UNIVERSITÀ DEGLI STUDI DI PADOVA

POLISH-JAPANESE ACADEMY OF INFORMATION TECHNOLOGY

Leveraging Persuasion and Intent for Analysis and Reasoning-based Detection of Disinformation with Large Language Models

PH.D. STUDENT

Arkadiusz Modzelewski

POLISH-JAPANESE ACADEMY OF INFORMATION TECHNOLOGY, DOCTORAL SCHOOL ICT & DESIGN

DISCIPLINE OF SCIENCE: INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITY OF PADUA, DEPARTMENT OF MATHEMATICS "TULLIO LEVI-CIVITA"

PH.D. COURSE IN: BRAIN, MIND AND COMPUTER SCIENCE (BMCS)

CURRICULUM: COMPUTER SCIENCE FOR SOCIETAL CHALLENGES AND INNOVATION

SERIES: XXXIX (COTUTELLE)

SUPERVISOR

Prof. Adam Wierzbicki

Polish-Japanese Academy of Information Technology

SUPERVISOR

Prof. Giovanni Da San Martino

University of Padua

BMCS COORDINATOR

Prof. Anna Spagnolli

University of Padua

ACADEMIC YEAR 2025/2026

Additional Resources

Following reproducibility standards emphasized in major Natural Language Processing conferences such as ACL, EMNLP, and EACL, the full codbase used to obtain the results in this dissertation has been made publicly available.

 ArkadiusDS/Understanding-and-Detecting-Disinformation

To one I deeply respect and admire,
for your patience, understanding,
and for the moments when your support made all the difference.

To my supervisors,
for the freedom you gave me,
and for guiding this work with insight and trust.

To my family, especially my parents,
for your unwavering belief in me,
and for nurturing my growth from childhood to this milestone.

Contents

Abstract

The rapid and widespread dissemination of online disinformation constitutes a major threat to democratic processes, public health, and societal stability. Although disinformation is commonly defined as intentionally misleading and persuasive communication, most Natural Language Processing (NLP) approaches still reduce its detection as a binary classification problem and overlook the mechanisms that make disinformation effective, namely, malicious intent and misleading persuasion techniques. This abstraction limits both the robustness and the explainability of existing systems, particularly in cross-domain, cross-genre, and emerging AI-generated content settings.

This dissertation addresses this gap by proposing a persuasion- and intent-augmented computational framework to improve large language model reasoning for disinformation detection. It advances the state of the art through four interrelated research studies. First, it introduces novel human-annotated datasets that move beyond binary disinformation labels by explicitly capturing manipulation techniques and malicious intent, including the first large-scale Polish corpus and the first English dataset annotated for malicious intent with stepwise annotations from each stage of the annotation process. These resources enable fine-grained empirical study of disinformation as a goal-driven communicative phenomenon and establish new benchmarks for multifaceted disinformation analysis in Polish and English.

Second, drawing on insights from psychology and communication studies, the dissertation proposes persuasion-augmented reasoning for large language models, demonstrating that explicitly analyzing persuasive strategies as intermediate reasoning steps significantly improves disinformation detection and generalizes across domains, genres, and temporal shifts. Third, it introduces intent-augmented reasoning, showing that explicit reasoning about malicious intent further enhances robustness and generalization, including across multiple languages.

Finally, the dissertation investigates AI-generated persuasive text, introducing a multilingual benchmark to compare the detection difficulty and linguistic characteristics of human-written and AI-generated persuasive content. Through extensive linguistic and empirical analyses, it shows that AI-generated persuasive texts exhibit systematic linguistic differences and pose new challenges for disinformation detection systems.

Overall, this work bridges theoretical definitions of disinformation with their computational treatment, delivering new datasets, reasoning frameworks, and empirical insights that support more transparent, generalizable, and future-proof disinformation detection systems in the era of generative AI.

Sommario

La rapida diffusione della disinformazione online può indebolire i processi democratici, la salute pubblica e la stabilità sociale. Sebbene la disinformazione sia comunemente definita come una comunicazione intenzionalmente fuorviante e persuasiva, la maggior parte degli approcci di elaborazione del linguaggio naturale (NLP) continua a ridurne il rilevamento a un problema di classificazione binaria e trascura i meccanismi che rendono efficace la disinformazione, ovvero gli intenti malevoli e le tecniche di persuasione ingannevoli. Questo limita sia la robustezza che l'interpretabilità dei sistemi esistenti, in particolare in contesti cross-domain, cross-genre e per contenuti emergenti generati dall'intelligenza artificiale.

Il presente lavoro propone un framework computazionale per integrare abilità di riconoscimento di tecniche di persuasione ed intento nei modelli linguistici di grandi dimensioni per migliorare le loro capacità di rilevare la disinformazione. Per tale scopo, quattro linee di ricerca, tra loro interconnesse, vengono presentate in questo lavoro. In primo luogo, si introducono nuovi dataset annotati che vanno oltre le etichette binarie di disinformazione, catturando esplicitamente le tecniche di manipolazione e gli intenti malevoli, tra questo il primo corpus di grandi dimensioni per la lingua polacca e il primo dataset in inglese annotato con intenti malevoli. Queste risorse consentono uno studio empirico dettagliato della disinformazione e stabiliscono nuovi parametri di riferimento per un'analisi su più livelli della disinformazione in polacco e inglese.

In secondo luogo, attingendo alle conoscenze della psicologia e degli studi sulla comunicazione, la tesi propone un ragionamento potenziato dalla persuasione per i modelli linguistici di grandi dimensioni, dimostrando che l'analisi esplicita delle strategie persuasive come fasi intermedie del ragionamento migliora significativamente il rilevamento della disinformazione e si generalizza attraverso domini, generi e cambiamenti temporali. In terzo luogo, viene introdotto una forma di ragionamento per i modelli linguistici arricchita dall'analisi dell'intento, con miglioramenti significativi in termini di capacità di generalizzazione, anche su più lingue.

Infine, la tesi confronta la difficoltà di rilevamento e le caratteristiche linguistiche dei contenuti persuasivi scritti dall'uomo e quelli generati dall'intelligenza artificiale. Attraverso approfondite analisi linguistiche ed empiriche, mostra che i testi persuasivi generati dall'intelligenza artificiale presentano differenze linguistiche sistematiche e pongono nuove sfide ai sistemi di rilevamento della disinformazione.

Nel complesso, questo lavoro colma il divario tra le definizioni teoriche della disinformazione e il loro trattamento computazionale, fornendo nuovi dataset, modelli di ragionamento e approfondimenti empirici per ottenere sistemi di rilevamento della disinformazione più trasparenti, generalizzabili e a prova di futuro nell'era dell'IA generativa.

Streszczenie

Szybkie rozpowszechnianie dezinformacji w przestrzeni internetowej stanowi poważne zagrożenie dla procesów demokratycznych, zdrowia publicznego oraz stabilności społecznej. Chociaż dezinformacja jest definiowana jako perswazyjna komunikacja celowo wprowadzająca w błąd, większość podejść redukuje jej automatyczne wykrywanie do problemu binarnej klasyfikacji, pomijając kluczowe mechanizmy jej skuteczności, takie jak manipulacyjne techniki perswazyjne oraz złośliwe intencje nadawcy. Ograniczenie to wpływa negatywnie na wyjaśnialność systemów, zwłaszcza w scenariuszach obejmujących różne kategorie tematyczne oraz gatunki tekstów.

Niniejsza rozprawa doktorska odpowiada na tę lukę badawczą poprzez zaproponowanie obliczeniowego podejścia do wykrywania dezinformacji z wykorzystaniem modeli językowych, wzbogaconego o rozumowanie oparte na perswazji i intencji. Praca rozwija stan wiedzy poprzez cztery powiązane ze sobą badania. Po pierwsze, wprowadza nowe zbiory danych, które wykraczają poza binarne etykiety dezinformacji, uwzględniając techniki manipulacji oraz złośliwe intencje, w tym pierwszy korpus w języku polskim oraz pierwszy anglojęzyczny zbiór anotowany pod kątem złośliwych intencji wraz z adnotacjami pochodzącymi z kolejnych etapów procesu anotacji. Zasoby te umożliwiają szczegółowe badanie dezinformacji jako zjawiska komunikacyjnego ukierunkowanego na realizację określonych celów oraz ustanawiają nowe punkty odniesienia dla wielowymiarowej analizy dezinformacji w języku polskim i angielskim.

Po drugie, czerpiąc z dorobku psychologii oraz badań nad komunikacją, rozprawa proponuje wzbogacone rozumowanie dla dużych modeli językowych, wykazując, że analiza perswazyji jako pośredniego kroku rozumowania istotnie poprawia skuteczność wykrywania dezinformacji oraz dobrze generalizuje się pomiędzy różnymi tematami, gatunkami tekstów i zmianami temporalnymi. Po trzecie, rozprawa przedstawia rozumowanie wzbogacone o analizę intencji, pokazując, że jawne wnioskowanie o złośliwych intencjach zwiększa zdolność modeli do wykrywania dezinformacji, również w kontekście wielojęzycznym.

Ponadto, rozprawa analizuje teksty perswazyjne generowane przez modele sztucznej inteligencji, wprowadzając wielojęzyczny zbiór danych umożliwiający porównanie trudności detekcji oraz charakterystyk językowych treści tworzonych przez ludzi i modele językowe. Na podstawie szeroko zakrojonych analiz lingwistycznych i empirycznych wykazano, że perswazja generowana przez AI charakteryzuje się systematycznymi różnicami językowymi i stanowi nowe wyzwanie dla istniejących systemów wykrywania dezinformacji.

Podsumowując, praca łączy teoretyczne definicje dezinformacji z ich automatyczną detekcją, dostarczając nowych zbiorów danych, ram rozumowania dużych modeli językowych oraz wniosków empirycznych, które wspierają rozwój przejrzystych, uogólnialnych i odpornych na przyszłe wyzwania systemów wykrywania dezinformacji w erze generatywnej sztucznej inteligencji.

1 Introduction

1.1 Motivation and Research Problem

The creation, dissemination, and consumption of online disinformation raise increasing concerns, driven by easy access to false content and limited public awareness of its misleading nature [1]. Disinformation spreads rapidly and poses growing risks to democratic institutions, public health, and social stability [2]. The High-Level Expert Group established by the European Commission defines disinformation as [3]:

> *"False, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit."*

This definition has been widely adopted in Natural Language Processing research as well as across the social sciences and humanities [4, 5, 6, 7].

Despite increasing recognition that disinformation is an intentionally misleading and persuasive form of communication, most NLP approaches explore disinformation detection as a binary classification task [8, 9, 10]. This paradigm abstracts away from the mechanisms that characterize disinformation, particularly the presence of malicious intent and the use of misleading persuasive techniques. As a result, current systems offer limited insight into why a piece of content is classified as disinformation, constraining their usefulness for analysts, policymakers, and end users who must interpret and act upon such decisions.

Research has shown that persuasion and manipulation are central components of disinformation [11, 12]. Another fundamental but underexplored dimension is malicious intent. Intent is part of the definition of disinformation [3], but in NLP research, the malicious intent behind disinformation is underexplored.

Existing theoretical frameworks seek to explain why disinformation is produced [4], yet there is little empirical and computational work, or annotated data, that directly captures these intentions.

Recent advances in large language models further enable reasoning-based detection approaches. Nevertheless, current methods largely underexploit the use of intermediate representations, such as explicit analyses of persuasion strategies or malicious intent, to guide model reasoning. The lack of frameworks for integrating this information may limit performance gains.

In this work, we argue that disinformation detection must move beyond binary judgments toward richer systems. We aim to develop methods that not only automatically classify disinformation but also enrich this classification with explanatory information that analyzes the persuasive strategies and intent underlying the content. By enriching disinformation detection outputs with explicit information about persuasive strategies and malicious intents, this dissertation aims to provide more informative predictions by offering users insight into how and why disinformation seeks to influence its audiences.

From a computational perspective, these conceptual gaps surrounding malicious intent and misleading persuasion are further compounded by the lack of high-quality, expert-annotated datasets that explicitly model these dimensions of disinformation. Available resources typically collapse diverse disinformation objectives into a single binary label, obscuring differences in underlying intent and persuasive strategy. As a result, malicious intent and the persuasiveness of disinformation, despite being central to its conceptual definition, remain largely absent from computational frameworks, constraining model generalization and explainability. To address this limitation, this dissertation introduces novel, expert-annotated datasets that explicitly capture these dimensions.

Moreover, prior work indicates that fine-tuned binary disinformation classifiers may generalize poorly across domains, with performance degrading when models are trained on one domain and evaluated on unseen topics [13]. This dissertation therefore systematically evaluates the generalization ability of automatic disinformation detection methods beyond their training domains and proposes approaches to improve robustness under both domain and temporal shifts. The high-quality dataset developed in this work serves as a foundation for the systematic evaluation of the proposed methods.

Finally, the growing prevalence of AI-generated persuasive text introduces an additional challenge for disinformation detection. Large language models are increasingly capable of producing persuasive content that may mirror the persuasiveness found in human-authored texts [14, 15, 16], potentially amplifying the scale and adaptability of disinformation. Understanding whether and how

AI-generated persuasion differs from human-written persuasion is therefore crucial for developing robust detection systems, as such differences may affect their generalizability. In the absence of empirical benchmarks and systematic comparative analyzes, current disinformation detection systems risk becoming increasingly fragile in the face of automated, persuasive content generation.

Consequently, the core research problem addressed in this dissertation is the design of an intent- and persuasion-augmented framework for automatic disinformation detection, grounded in expert knowledge, that supports intermediate reasoning and remains robust across domains. Furthermore, the dissertation advances the understanding of AI-generated persuasive texts by systematically examining whether such content is more challenging to detect automatically than human-written persuasion and analyzing its linguistic features.

1.2 Research Objectives

This dissertation attempts to achieve the following four research objectives:

- **Research Objective 1 [RO1]** - Design a human-annotated disinformation dataset and annotation framework that captures manipulation techniques and malicious intent, going beyond binary credibility labels, to benchmark language models for detecting disinformation, manipulation techniques, and malicious intent classification.
- **Research Objective 2 [RO2]** - Develop and evaluate a persuasion-augmented reasoning framework for large language models, enabling intermediate analysis and explanation of persuasive strategies to improve disinformation detection.
- **Research Objective 3 [RO3]** - Develop and evaluate an intent-augmented reasoning framework for large language models that uses knowledge and explanations of malicious intent to improve disinformation detection.
- **Research Objective 4 [RO4]** - Analyze and compare AI-generated and human-authored persuasive content based on linguistic characteristics and detection difficulty for automatic disinformation systems.

1.3 Contributions

This dissertation contributes to the field of Computer Science, particularly Natural Language Processing, by advancing the computational study of disinformation, persuasion techniques, and malicious intent.

The core of this dissertation is structured around four interrelated studies, each presented in a dedicated chapter. Collectively, these studies investigate disinformation through the lenses of persuasion and malicious intent, progressing from the creation of human-annotated datasets to the development of reasoning-based detection frameworks and the analysis of AI-generated persuasive content. Together, they form a coherent research trajectory that provides a multifaceted perspective on characterizing and detecting disinformation. The main contributions are summarized below.

Main contributions of the first study (presented in Chapter 4):

- **C1.1: Expert-Annotated Disinformation Dataset for a Lower-Resource Language (Polish).** The dissertation presents the first large-scale Polish disinformation corpus with annotations beyond binary credibility judgments. The dataset was developed as a part of the *InfoTester* project by multiple researchers and professional fact-checking experts. This process ensured high annotation quality and methodological rigor. The dataset captures manipulation techniques and malicious intent. It covers 10 thematic categories, including the War in Ukraine and LGBT+ issues. This contribution addresses the lack of high-quality disinformation resources for lower-resource languages. It enables fine-grained analyses of the intentional and manipulative dimensions of Polish disinformation.

- **C1.2: Benchmarking Language Models for Multifaceted Disinformation Analysis in Polish.** The dissertation establishes baseline language models for disinformation detection, manipulation technique classification, and malicious intent classification in Polish web articles. Using the proposed expert-annotated dataset, Polish language models are systematically benchmarked across these tasks, providing the first comprehensive evaluation of malicious intent and manipulation classification for Polish. All baseline models fine-tuned for disinformation, manipulation classification, and malicious intent are released on *HuggingFace*[1], enabling reproducibility and supporting future research on Polish disinformation.

As a result, Chapter 4 addresses **Research Objective 1 (RO1)** by presenting a high-quality annotated dataset that enables fine-grained disinformation analysis beyond binary classification. In addition, it goes beyond this objective by benchmarking language models for disinformation detection, manipulation technique classification, and malicious intent classification, and by releasing publicly available classification models on *HuggingFace*.

[1]Link to the collection of models: https://huggingface.co/collections/ArkadiusDS/mipd

Main contributions of the second study (presented in Chapter 5):

- **C2.1: English Disinformation Dataset with Stepwise Expert Annotations.** The dissertation presents a human-annotated English-language disinformation dataset developed using a multi-stage annotation protocol. The dataset was developed as part of the *InfoTester4Education* project by multiple researchers from different research institutions across four European countries, in collaboration with professional fact-checking and debunking experts. The annotation methodology and guidelines for the dataset were co-designed by the author of this dissertation, who also contributed to the annotation of the dataset. Crucially, the dataset preserves intermediate annotation decisions and enables not only disinformation detection research but also meta-analyses of the annotation workflow, including studies of annotation consistency and inter-annotator agreement.

- **C2.2: Persuasion-Augmented Reasoning for Disinformation Detection.** This study introduces a persuasion-augmented reasoning framework for large language models, in which analyzes of persuasive strategies are explicitly incorporated as intermediate reasoning steps. The proposed approach improves robustness and generalization in zero-shot, cross-domain, and cross-genre disinformation detection, demonstrating that structured reasoning about persuasion enhances performance in complex NLP disinformation detection scenarios.

As a result, Chapter 5 addresses **Research Objective 2 (RO2)** by developing and evaluating a persuasion-augmented reasoning framework for large language models, in which explicit analysis of persuasive strategies is incorporated as an intermediate reasoning step to improve disinformation detection. In addition, the chapter partially supports **Research Objective 1 (RO1)** by introducing an English-language disinformation dataset with stepwise expert annotations that preserve intermediate decisions and enable analysis of English disinformation.

Main contributions of the third study (presented in Chapter 6):

- **C3.1: English Disinformation Dataset Annotated for Malicious Intent with Stepwise Annotations.** This dissertation presents a human-annotated English-language disinformation dataset capturing distinct categories of malicious intent. The dataset was developed as a continuation of the *InfoTester4Education* project, in collaboration with multiple European research institutes and professional fact-checking and debunking experts. It supports research on intent-aware disinformation detection. By preserving

annotations from each stage of the annotation process, the dataset additionally enables analyses of annotation dynamics, including consistency and agreement across annotation steps.

- **C3.2: Intent-Augmented Reasoning for Disinformation Detection.** The dissertation introduces intent-augmented reasoning, also referred to as intent-based inoculation, as a novel approach to disinformation detection. This framework demonstrates that explicitly analyzing malicious intent significantly enhances large language model reasoning, particularly in challenging cross-domain, cross-genre, and multilingual settings.
- **C3.3: Baselines and Benchmarking Language Models for Malicious Intent Classification.** The dissertation establishes baseline methods and benchmarks twelve language models, spanning BERT-based models and generative large language models, on malicious intent classification tasks.

As a result, Chapter 6 primarily addresses **Research Objective 3 (RO3)** by developing and evaluating an intent-augmented reasoning framework for large language models, in which explicit analysis and explanation of malicious intent are incorporated to improve disinformation detection. In addition, the chapter contributes to **Research Objective 1 (RO1)** by presenting an English-language disinformation dataset annotated for malicious intent with stepwise expert annotations, and by establishing baselines and benchmarks for malicious intent classification across a diverse set of language models.

Main contributions of the fourth study (presented in Chapter 7):

- **C4.1: Systematic Comparison of Human-Written and AI-Generated Persuasive Texts.** The dissertation contributes to emerging research on generative AI by comparing human-authored and AI-generated persuasive texts. It introduces controlled text generation procedures to produce synthetic persuasive content, enabling rigorous evaluation of whether AI-generated persuasive content is more challenging to detect automatically than human-written persuasive text.
- **C4.2: Linguistic Analysis of AI-Generated Persuasive Texts.** The dissertation provides a detailed linguistic analysis identifying stylometric features that distinguish machine-generated persuasive texts from human-written persuasive texts. This analysis offers empirical insights into the linguistic properties of AI-generated persuasive texts.
- **C4.3: Dataset and Benchmarks for Detecting AI-Generated Persuasion.** The dissertation introduces a novel multilingual dataset and benchmarks

multiple large language models on the task of detecting AI-generated persuasive texts, establishing baselines and supporting future research on automated detection of persuasive AI content.

As a result, Chapter 7 addresses **Research Objective 4 (RO4)** by systematically analyzing and comparing human-authored and AI-generated persuasive content. The chapter investigates linguistic characteristics of AI-generated persuasion and evaluates the difficulty of automatically detecting such content, introducing a new dataset and benchmarks to support research on the detection of AI-generated persuasive texts.

Summary. Collectively, the contributions of this dissertation bridge the gap between theoretical definitions of disinformation, particularly its persuasive and intentional dimensions, and their computational treatment in Natural Language Processing. By introducing intent- and persuasion-augmented reasoning frameworks, releasing expert-annotated datasets, and conducting rigorous empirical evaluations, this work lays the foundations for more transparent, generalizable, and practically relevant disinformation detection systems. Furthermore, through a study of AI-generated versus human persuasion, the dissertation anticipates emerging threats posed by generative models and provides essential insights for developing future tools to detect persuasive AI-generated content that may be used in large-scale disinformation campaigns. These contributions support future research at the intersection of NLP, disinformation studies, and AI safety.

1.4 Publications Forming This Dissertation

Each core chapter of this dissertation is grounded in a research study that has been developed into a scholarly publication. These publications constitute the dissertation's core contributions and reflect the research outcomes achieved during the PhD studies. Two of the studies have been published in CORE A* conferences: EMNLP and ACL (Main Proceedings). A third paper has been accepted to the CORE A conference EACL (Main Proceedings), while the fourth study is currently under review for the ACL conference.

The following sections provide an overview of the four core chapters and cite the publications that serve as the basis for each chapter.

1.4.1 Manipulation and Malicious Intent in Polish Disinformation

The first core study, presented in Chapter 4, is based on a published paper:
Arkadiusz Modzelewski, Giovanni Da San Martino, Pavel Savov, Magdalena Anna Wilczyńska, and Adam Wierzbicki. 2024. *MIPD: Exploring Manipulation and Intention In a Novel Corpus of Polish Disinformation.* In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), Miami, Florida, USA. Association for Computational Linguistics.

In the first study, introduced in Chapter 4 and based on the study published as Modzelewski et al. [17], we present MIPD, a novel corpus of 15,356 Polish web articles designed to support disinformation research beyond binary veracity classification. The dataset explicitly captures the intentional and manipulative dimensions of disinformation.

The chapter introduces a multilayer annotation methodology carried out by professional fact-checkers, who label malicious intent types and manipulation techniques alongside disinformation. Using MIPD, we establish the first Polish baselines for disinformation detection, manipulation technique classification, and intent classification, and report experiments with Polish BERT-based models as well as zero-shot evaluations with GPT-3.5 and GPT-4.

1.4.2 Leveraging Persuasion for Disinformation Detection

The second core study, presented in Chapter 5, is based on a published paper:
Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. 2025. *PCoT: Persuasion-Augmented Chain of Thought for Detecting Fake News and Social Media Disinformation.* In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL Volume 1: Long Papers), Vienna, Austria. Association for Computational Linguistics.

In the second study, presented in Chapter 5 and based on the study published as Modzelewski et al. [18], we propose a novel reasoning framework for disinformation detection with large language models that explicitly leverages knowledge of persuasion. Inspired by findings from psychology and communication research, the study examines whether analyzing persuasive strategies as part of the reasoning process can improve ability of Large Language Models (LLMs) to recognize deceptive content in a zero-shot setting.

To this end, we introduce Persuasion-Augmented Chain of Thought (PCoT) , a two-stage reasoning approach in which LLMs first identify and explain persuasive strategies in a text and then use this analysis to inform the final disinformation judgment. Extensive experiments on news articles and social media

posts, including evaluations on two newly introduced datasets (EUDisinfo and MultiDis), show that PCoT consistently outperforms strong baselines, achieving an average improvement of 15%. These results demonstrate that persuasion-augmented reasoning substantially enhances generalization and robustness in disinformation detection.

1.4.3 Leveraging Malicious Intent for Disinformation Detection

The third core study, presented in Chapter 6, is based on an accepted paper:
Arkadiusz Modzelewski, Witold Sosnowski, Eleni Papadopulos, Elisa Sartori, Tiziano Labruna, Giovanni Da San Martino, and Adam Wierzbicki. 2026. *MALicious INTent Dataset and Inoculating LLMs for Enhanced Disinformation Detection.* In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Rabat, Morocco. Association for Computational Linguistics (accepted and soon to be published).

In the third study, presented in Chapter 6 and forthcoming as Modzelewski et al. [19], we introduce MALINT, the first human-annotated English corpus explicitly designed to capture both disinformation and the malicious intent of its creators. Developed in close collaboration with professional fact-checkers, MALINT follows a rigorous multi-stage annotation process and releases annotations from each step, enabling transparent and fine-grained analysis of intent in disinformation.

Building on this resource, the chapter presents the first systematic evaluation of malicious intent detection in English and introduces intent-based inoculation, an intent-augmented reasoning framework inspired by inoculation theory. In this approach, LLMs analyze hidden malicious intents as an intermediate reasoning step to strengthen disinformation detection. Extensive experiments across multiple datasets, models, and languages show consistent performance gains, demonstrating that intent-aware reasoning improves disinformation detection.

1.4.4 Persuaficial: Human vs. AI Persuasive Texts Comparison

The fourth core study, presented in Chapter 7, is based on a paper under review and available as a preprint:
Arkadiusz Modzelewski, Paweł Golik, Anna Kołos, and Giovanni Da San Martino. *Can AI-Generated Persuasion Be Detected? Persuaficial Benchmark and AI vs. Human Linguistic Differences*.

In the fourth study, presented in Chapter 7 and currently under review as Modzelewski et al. [20], we extend prior work on persuasion by systematically

analyzing persuasive language in human- and AI-generated texts. The study investigates whether persuasive content produced by large language models differs from human-written persuasion in its linguistic characteristics and whether AI-generated persuasive texts pose greater challenges for automatic detection.

To this end, we introduce Persuaficial, a multilingual benchmark designed to directly compare human-written and LLM-generated persuasive texts across six languages. Extensive experiments show that while some forms of AI-generated persuasion are relatively easy to detect, more subtly crafted persuasive content generated by LLMs remains significantly more difficult. Complementary linguistic analysis reveals consistent stylometric differences between human and machine-generated persuasion, offering insights that may influence the development of more robust persuasion detection systems.

1.5 Publications Not Included in the Dissertation

The publications listed below were written during the PhD period but are not included in this dissertation. They are presented for completeness and to document additional scientific contributions beyond the scope of the thesis.

Several of these works were published in leading international venues, including the CORE A* conference EMNLP. The list also includes a publication resulting from the co-organization of a shared task at the Slavic NLP workshop, co-located with ACL. In addition, several publications arose from international collaborations with researchers from European institutions.

List of published works:

1. Witold Sosnowski, **Arkadiusz Modzelewski**, Kinga Skorupska, Jahna Otterbacher, and Adam Wierzbicki. 2024. *EU DisinfoTest: a Benchmark for Evaluating Language Models' Ability to Detect Disinformation Narratives.* In Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA. Association for Computational Linguistics.

2. Witold Sosnowski, **Arkadiusz Modzelewski**, Kinga Skorupska, and Adam Wierzbicki. 2025. *DiNaM: Disinformation Narrative Mining with Large Language Models.* In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025), Suzhou, China. Association for Computational Linguistics.

3. Tiziano Labruna, **Arkadiusz Modzelewski**, Giorgio Satta, and Giovanni Da San Martino. 2026. *Detecting Winning Arguments with Large Language*

Models and Persuasion Strategies. In Findings of the Association for Computational Linguistics: EACL 2026, Rabat, Morocco. Association for Computational Linguistics (accepted and soon to be published).

4. **Arkadiusz Modzelewski**, Witold Sosnowski, Magdalena Wilczynska, and Adam Wierzbicki. 2023. *DSHacker at SemEval-2023 Task 3: Genres and Persuasion Techniques Detection with Multilingual Data Augmentation through Machine Translation and Text Generation.* In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1582–1591, Toronto, Canada. Association for Computational Linguistics.

5. Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michal Marcinczuk, **Arkadiusz Modzelewski**, Ivo Moravski, and Roman Yangarber. 2025. *SlavicNLP 2025 Shared Task: Detection and Classification of Persuasion Techniques in Parliamentary Debates and Social Media*. In Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 254–275, Vienna, Austria. Association for Computational Linguistics.

6. **Arkadiusz Modzelewski**, Witold Sosnowski, and Adam Wierzbicki 2023. *DSHacker at CheckThat!-2023: Check-Worthiness in Multigenre and Multilingual Content With GPT-3.5 Data Augmentation.* In Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023).

7. **Arkadiusz Modzelewski**, Paweł Golik, and Adam Wierzbicki 2024. *Bilingual propaganda detection in diplomats' tweets using language models and linguistic features.* Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024).

8. Paweł Golik, **Arkadiusz Modzelewski**, and Aleksander Jochym 2024. *DSHacker at CheckThat! 2024: LLMs and BERT for check-worthy claims detection with propaganda co-occurrence analysis.* Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)

List of publications under review:

1. Witold Sosnowski, **Arkadiusz Modzelewski**, Kinga Skorupska, and Adam Wierzbicki. 2025. *DiNO: Disinformation Narrative Observer.* Manuscript under review at the Conference of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026).

2. Jakub Piskorski, Dimitar Iliyanov Dimitrov, Marina Ernst, Jacek Haneczok, Michal Marcinczuk, **Arkadiusz Modzelewski**, and Roman Yangarber. 2026. *A Corpus of Persuasion Techniques in Slavic Languages.* Manuscript under review at the 2026 International Conference on Language Resources and Evaluation (LREC 2026).

3. Marina Ernst, **Arkadiusz Modzelewski**, Adam Wierzbicki, and Frank Hopfgartner. 2025. *Disinformation Detection with LLMs: Can Different Models Agree on Their Judgment?* Manuscript under review at *Annals of Operations Research*.

2 Background and Fundamental Concepts

2.1 Technical Foundations

2.1.1 Introduction Language Models

Language models constitute a foundational component of modern natural language processing, providing a formal framework for modeling, understanding, and generating human language.

Language Modeling. A language model defines a probability distribution over sequences of tokens. Given a sequence of tokens $w_1, w_2, \ldots, w_T$, the goal of language modeling is to estimate the joint probability:

$$P(w_1, w_2, \ldots, w_T). \tag{2.1}$$

By applying the chain rule of probability, this joint distribution can be factorized as:

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t \mid w_1, \ldots, w_{t-1}). \tag{2.2}$$

This formulation reduces language modeling to the problem of predicting the next token given its preceding context. It provides a unifying probabilistic framework underlying both classical statistical models and modern neural language models, and it serves as the basis for a wide range of downstream tasks, including text classification, information extraction, and text generation [21].

Modern Architectures of Language Models. While the probabilistic formulation of language modeling is architecture-agnostic, modern neural language

models are commonly instantiated using one of three architectural paradigms: *encoder-only*, *decoder-only*, and *encoder–decoder* architectures. These architectures are independent of the specific neural network realization [21]. In contemporary NLP, they are most commonly implemented using the Transformer architecture [22], which has become the dominant modeling framework for language models.

Decoder. The decoder is an architecture that takes as input a sequence of tokens (basic unit of text, e.g. subwords) and generates an output sequence one token at a time. Information flows strictly left-to-right, meaning each predicted word depends only on prior words. Decoders are generative models: given some input tokens, they can produce novel output tokens. This architecture underlies large language models like GPT, Claude, Llama, and Mistral [21].

Encoder. The encoder takes as input a sequence of tokens and outputs a vector representation for each token. Encoders are typically trained as masked language models, where certain words are masked and the model learns to predict them using surrounding context on both sides. Masked language models e.g., BERT, RoBERTA, and others in the BERT family are encoder models [21]. Because they are not designed to generate text, encoder models are not generative models. Instead, they are commonly fine-tuned for tasks such as classification, where text inputs are mapped to labels like sentiment, topic, or other categories [21].

Encoder–Decoder The encoder–decoder architecture takes a sequence of input tokens and produces a sequence of output tokens, but with a looser relationship between inputs and outputs than in decoder-only models [21]. The output tokens may differ substantially from the input tokens. This flexibility makes encoder–decoder models suitable for tasks that map between different kinds of representations, such as machine translation or speech recognition, where the input and output token sequences can be in different languages or of different lengths [21].

Language Models in This Work. Language model architectures differ primarily in how they represent and condition on context, which in turn determines their suitability for different classes of NLP tasks. Encoder-only models are particularly effective for discriminative tasks such as text classification and sequence labeling. Decoder-only models naturally support text generation and enable task adaptation through prompting without parameter updates. Encoder–decoder models are well suited for structured input–output transformations, including translation and speech recognition.

This dissertation focuses on encoder-based models for supervised text classification and on decoder-based large language models for prompt-based inference and reasoning, reflecting the complementary strengths of these architectures in addressing disinformation-related tasks.

2.1.2 Text Classification in Modern Natural Language Processing

Text Classification. Text classification is a fundamental task in natural language processing that aims to assign one or more labels to a textual input. Formally, given an input text sequence $x = (w_1, \ldots, w_T)$ composed of T tokens, the objective of text classification is to learn a function

$$f : \mathcal{X} \to \mathcal{Y}, \tag{2.3}$$

where $\mathcal{X}$ denotes the space of textual inputs and $\mathcal{Y}$ is a discrete label space. In the binary and multiclass settings, $\mathcal{Y} = \{1, \ldots, K\}$, where exactly one label is assigned to each input, whereas in the multilabel setting $\mathcal{Y} = \{0, 1\}^K$, where each component indicates the presence or absence of a particular label, allowing multiple labels to be assigned simultaneously to a single input.

Depending on the application, text classification can be performed at different levels of granularity, including document-level, sentence-level, or span-level classification. This task underlies a wide range of NLP applications such as sentiment analysis, topic categorization, disinformation detection, persuasion and intent detection.

Text Classification with Encoder-Based Models. The introduction of pretrained encoder-based language models marked a major shift in text classification. Models such as BERT and its variants produce contextualized token representations by attending to both left and right context, enabling rich semantic encoding of the input text. For classification tasks, a task-specific prediction head is typically added on top of the encoder, and the entire model is fine-tuned using labeled data. This approach has become the dominant paradigm for supervised text classification due to its strong empirical performance, sample efficiency, and robustness across domains and languages. Encoder-based models are particularly well suited for discriminative tasks, including multilabel classification and fine-grained categorization, which are central to applications such as disinformation, persuasion and intent detection.

Text Classification with Decoder-Based Models. More recently, large decoder-only language models have enabled an alternative approach to text classification

based on conditional text generation. Instead of learning a dedicated classification head, classification is formulated as a generation task in which the model is prompted with an input text and instructed to produce a label or a structured response. This paradigm supports zero-shot and few-shot classification without parameter updates, relying solely on inference-time prompting. Decoder-based models are especially flexible and can incorporate task instructions, label descriptions, or reasoning steps directly into the prompt. However, compared to fine-tuned encoder models, their performance and reliability can be sensitive to prompt design.

Text Classification in This Work. Modern text classification in this dissertation is approached through encoder-based and decoder-based language models. For supervised classification, encoder-based architectures are employed due to their strong discriminative capabilities in labeled settings. In particular, pretrained Transformer encoders including BERT, RoBERTa, and DistilBERT are used for English-language data, while Polish-language experiments rely on Polish pretrained models such as HerBERT and Polish RoBERTa.

In parallel, large decoder-only language models are explored as an alternative classification paradigm based on prompt-driven conditional generation. These models are used in combination with reasoning-enhancement techniques that aim to elicit structured intermediate reasoning during inference, improving classification reliability in complex settings. The basics and methods of reasoning with language model prompting are described in detail in the following section. Together, these approaches form the methodological foundation for the classification of disinformation, persuasive content, and malicious intent investigated in subsequent chapters.

2.1.3 Reasoning with Large Language Model Prompting

Reasoning with large language models has recently emerged as an important research direction in natural language processing, driven by the observation that sufficiently large pretrained models can exhibit non-trivial reasoning abilities when appropriately guided at inference time. Among several lines of research on LLM reasoning, one line relevant for this PhD dissertation focuses on *prompting*-based approaches. This line of research elicits enhanced reasoning without modifying model parameters through additional training, instead operating on frozen language models. A systematic overview and taxonomy of methods for reasoning with language model prompting is presented by Qiao et al. [23], which serves as the primary reference and the basis for this section.

Reasoning via Prompting. In the standard prompting setting, given a reasoning question Q, a prompt T, and a parameterized probabilistic language model p_{LM}, the objective is to maximize the likelihood of the correct answer A, i.e.,

$$p(A \mid T, Q) = \prod_{i=1}^{|A|} p_{LM}(a_i \mid T, Q, a_{<i}), \tag{2.4}$$

where a_i denotes the i-th token in the answer sequence, and $|A|$ denotes the length of the final answer [23]. In the few-shot prompting setting, the prompt T consists of K exemplar question–answer pairs (Q, A). Chain-of-Thought (CoT) prompting [24] further augments each exemplar with an explicit reasoning sequence C, such that

$$T = \{(Q_i, C_i, A_i)\}_{i=1}^{K}. \tag{2.5}$$

Accordingly, Equation 2.4 can be rewritten as follows:

$$p(A \mid T, Q) = p(A \mid T, Q, C)\, p(C \mid T, Q), \tag{2.6}$$

where $p(C \mid T, Q)$ and $p(A \mid T, Q, C)$ are defined as

$$p(C \mid T, Q) = \prod_{i=1}^{|C|} p_{\text{LM}}(c_i \mid T, Q, c_{<i})\,, \tag{2.7}$$

$$p(A \mid T, Q, C) = \prod_{j=1}^{|A|} p_{\text{LM}}(a_j \mid T, Q, C, a_{<j})\,, \tag{2.8}$$

where c_i denotes the i-th reasoning step, and $|C|$ is the total number of all reasoning steps [23].

Taxonomy of Reasoning Methods with LM Prompting. Following Qiao et al. [23], reasoning methods with large language model prompting can be broadly categorized into *strategy-enhanced reasoning* and *knowledge-enhanced reasoning*. Figure 2.1 presents the resulting taxonomy proposed by Qiao et al. [23].

Strategy-Enhanced Reasoning. Strategy-enhanced reasoning methods aim to improve the reasoning capability of large language models by modifying how prompts are structured, executed, or supported during inference, without introducing additional external knowledge. As shown in Figure 2.1, this class of methods can be broadly categorized into *prompt engineering*, *process optimization*, and *external engines* [23].

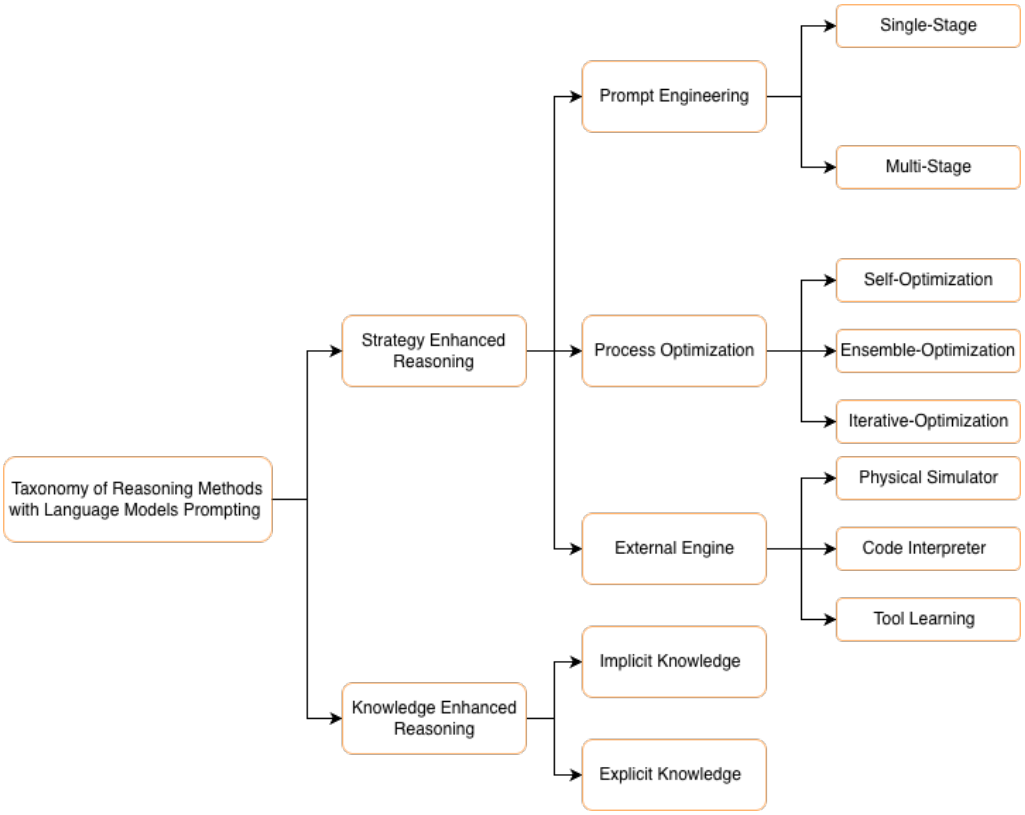

Figure 2.1: Taxonomy of Reasoning Methods with Language Models Prompting proposed by Qiao et al. [23].

For *prompt engineering* approaches, the primary objective is to enhance reasoning performance through improvements to the prompt T itself. Methods in this category can be broadly divided into *single-stage* and *multi-stage* approaches. Single-stage methods aim to elicit high-quality reasoning by constructing a single, well-designed prompt that conditions the model on (T, Q) in one generation step. In contrast, multi-stage methods decompose the reasoning process into a sequence of intermediate steps, where each reasoning step c_i is either incrementally appended to the existing context or generated using a dedicated prompt T_{c_i} tailored to that step [23].

Within strategy-enhanced reasoning, *process optimization* methods focus on improving the quality of reasoning by calibrating each reasoning step. A straightforward approach introduces an optimizer with parameters θ to calibrate the reasoning process C during answer generation, such approaches are referred to as *self-optimization* methods. Beyond single-process optimization, some methods exploit multiple reasoning trajectories and combine their outcomes to derive a

final answer, which was categorized as *ensemble-optimization* methods. Furthermore, the optimization process can be extended in an iterative manner by repeatedly generating reasoning triplets (Q, C, A) and using them to fine-tune the underlying language model p_{LM}. Approaches following this paradigm are denoted as *iterative optimization* methods [23].

Process optimization methods seek to enhance reasoning by refining the generation process itself, for example through *self optimization*, *ensemble optimization*, or *iterative optimization*, which aim to improve robustness, consistency, or accuracy by leveraging multiple reasoning trajectories or repeated refinement.

External-engine-based approaches augment the language model with auxiliary systems, such as simulators, code execution environments, or learned tools, enabling the model to offload specific reasoning subroutines while maintaining a prompt-centric interaction paradigm [23].

Knowledge-Enhanced Reasoning. *Knowledge-enhanced* reasoning methods are motivated by the observation that knowledge plays a critical role in an effective artificial intelligence reasoning systems [25]. As illustrated in Figure 2.1, knowledge-enhanced reasoning methods can be broadly categorized into *implicit knowledge* and *explicit knowledge* approaches. Implicit knowledge methods seek to elicit and structure latent knowledge already "included" within a language model's parameters, for example by prompting the model to generate intermediate knowledge representations or reasoning rationales that are subsequently used to guide downstream inference. In contrast, explicit knowledge methods integrate information from external sources directly into the reasoning process, typically through retrieval or grounding mechanisms, in order to mitigate issues such as factual hallucination or knowledge inconsistency. By incorporating externally provided evidence or structured information into prompts, these approaches enable language models to ground their reasoning in verifiable knowledge, particularly in settings that require multi-step or open-domain reasoning [23].

Reasoning with Language Models in This Work. This section introduced a background on reasoning with large language model prompting, focusing on approaches that enhance reasoning at inference time. Following the taxonomy of Qiao et al. [23], the methods developed in this dissertation, namely persuasion-augmented reasoning (persuasion-augmented chain of thought) and intent-augmented reasoning (intent-based inoculation), can be understood as multi-stage prompting approaches that combine both explicit and implicit knowledge. Specifically, they introduce external, structured knowledge about persuasion

strategies and techniques and malicious intent as explicit inputs, while simultaneously eliciting implicit reasoning in the form of model-generated analyses and explanations. By first prompting models to reason about intent and persuasion and then leveraging these intermediate analyses for final disinformation detection, these methods operationalize knowledge-enhanced, multi-stage reasoning.

2.2 Fundamental Definitions and Theoretical Foundations

2.2.1 Credibility and Disinformation

Credibility Credibility is a central concept in the evaluation of information quality and plays a foundational role in how information is assessed, trusted, and acted upon by receivers. In this dissertation, credibility is defined as: *a signal that may make a receiver believe that the information is true* [26]. This definition emphasizes that credibility is not an inherent property of a text or source, but rather an evaluative outcome formed by the receiver during information processing.

This perspective aligns with game-theoretic models of communication under asymmetric information, in which senders possess private information about their intentions, reliability, or truthfulness, while receivers must infer these hidden properties from observable signals [26]. In this setting, persuasion strategies, source characteristics, and factual consistency may function as signals that allow the receiver to estimate the credibility of the information. Credibility evaluation can thus be understood as an inference process under uncertainty, where receivers assess signals in order to reduce informational asymmetry.

A crucial distinction in credibility research is between *surface-level credibility* and *earned credibility*. Surface-level credibility refers to rapid evaluations made by message receiver [27], often within seconds or minutes, based on superficial cues such as website design, writing style, or perceived authority of the source. From a signal-based perspective, such evaluations rely on weak and noisy signals that provide only limited reduction of informational asymmetry between the sender and the receiver.

In contrast, earned credibility refers to credibility assessments grounded in systematic analysis, domain knowledge, and verification practices [27]. Earned credibility may be typically result of evaluation made by experts, such as journalists, fact-checkers, trained annotators, or subject-matter specialists, who assess information using evidence-based reasoning, cross-source verification, and contextual understanding. In this case, credibility is derived from stronger signals that substantially reduce informational asymmetry, as they are based on substantive properties such as factual accuracy and alignment with verified knowledge rather than on surface-level cues.

This distinction is central to the creation of novel disinformation datasets mentioned in this work. Rather than using surface-level credibility judgments, this work focuses on obtaining earned credibility evaluations for novel datasets, as performed by annotators, debunking and fact-checking professionals. Accordingly, the datasets presented in this dissertation are grounded in expert and trained annotators assessments of information credibility.

Disinformation is a complex and multidimensional phenomenon that has been conceptualized in diverse but related ways across disciplines [4]. In this dissertation, disinformation is defined as "false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit" [3]. This definition, proposed by the HLEG of the European Commission, is widely adopted in both academic and policy-oriented research [3].

Disinformation should therefore not be understood as a purely factual category, but rather as information that is disseminated with a harmful intention, and may be inaccurate or misleading due to the use of persuasive and manipulative strategies. As a result, assessing disinformation requires going beyond fact-checking individual claims and instead considering how information is constructed, framed, and disseminated within broader influence strategies.

Within this perspective, disinformation can be decomposed into two complementary dimensions: the *means* through which influence is exerted and the *goals* it is intended to achieve. The former concerns observable and relatively objective devices, such as persuasive and manipulative techniques, which can be identified directly in the text. The latter concerns the underlying goals that motivate the creation and dissemination of disinformative content, such as undermining trust, polarizing audiences, or advancing ideological or economic interests [4]. These goals are captured by the notion of malicious intent, which operates at a higher level of abstraction and is not easily observable in text.

In our research, disinformation classification is grounded in *earned credibility* evaluations performed by annotators and professional debunking or fact-checking experts. Within this framework, disinformation constitutes a specific subclass of non-credible content: while many forms of information (e.g., advertising or opinionated content) may be judged as non-credible, they do not necessarily qualify as disinformation. Disinformation is distinguished by the combination of degraded credibility signals arising from persuasive and manipulative language, the presence of malicious intent and the inclusion of factually false or misleading claims. These factors shape credibility judgments and serve as the empirical basis for the disinformation analysis conducted in this work.

In summary, this dissertation does not treat disinformation as a monolithic

category. Instead, dissemination of disinformation is approached as an intentional practice that integrates persuasive and manipulative means with strategic objectives. Persuasion and manipulation describe the observable tools of influence, while malicious intent captures the higher-level objectives guiding their use. The following sections elaborate on these two dimensions in detail, providing further theoretical foundation for the computational methods developed in this dissertation.

2.2.2 Persuasion

Persuasion is a fundamental communicative phenomenon in which language is used to influence readers' attitudes, judgments, or actions [28]. Persuasion encompasses a range of rhetorical and argumentative practices. Following this view, persuasive text is defined as text that employs specific linguistic and rhetorical choices in order to influence readers. Importantly, persuasion does not necessarily imply deception or harm [29]. It can be used for socially beneficial purposes, such as encouraging healthy behaviors, promoting civic engagement, or fostering awareness of societal issues. Consequently, persuasion should not be equated a priori with disinformation or malicious intent.

In line with this definition, this dissertation adopts a high-level categorization of persuasion techniques proposed by the Joint Research Centre, which is the European Commission's science and knowledge service that provides independent, evidence-based research and scientific advice to support European Union (EU) policies. Below mentioned high-level categorization of persuasion techniques that distinguishes between different mechanisms of influence [28]:

- **Attack on Reputation [AR]**: rather than engaging with the issue at hand, the argument shifts focus to the individual involved in order to cast doubt or even undermine their credibility. The object of the argumentation can also refer to a group of people, an organization, an object, or an activity,
- **Justification [J]**: the argument consists of two components: a claim and a supporting explanation or appeal, with the latter serving to justify and/or reinforce the claim,
- **Simplification [S]**: the argument reduces a complex issue to an overly simple explanation, often by oversimplifying its causes, effects, or the range of available options,
- **Distraction [D]**: the argument diverts attention from the central issue by shifting focus to an unrelated or less relevant point, thereby drawing the reader away from the main line of reasoning,

- **Call [C]**: the text does not present an argument, but instead urges the reader to act or to adopt a particular way of thinking or acting,
- **Manipulative Wording [MW]**: the text does not constitute an argument in itself, but it uses specific language that includes non-neutral, confusing, exaggerated, or loaded words and expressions in order to emotionally influence the reader.

The high-level categorization of persuasion techniques adopted in this dissertation has been widely used in Natural Language Processing research [30, 31]. It has also served as the basis for several Shared Tasks organized within the International Workshop on Semantic Evaluation across multiple years [32, 33], and was further applied in a Shared Task at the Slavic NLP Workshop [34].

Within this broader framework, manipulation is treated as a specific and more restrictive subset of persuasion. Manipulation differs from persuasion primarily in intent [29]. While persuasion aims to influence beliefs, opinions, or attitudes, manipulation seeks to induce recipients to make choices or adopt behaviors that primarily serve the manipulator's goals [29]. Crucially, manipulation is inherently associated with malicious intent: the communicator deliberately exploits cognitive biases, emotional responses, or informational asymmetries to guide the audience toward a predetermined outcome.

An important consequence of this distinction is that persuasion techniques cannot be treated as inherently harmful signals. Techniques that belong to e.g., *Justification* or *Call* (call to action) may legitimately appear in journalistic opinion pieces, public health campaigns, or educational materials. In contrast, manipulation techniques are indicative of an underlying intent to deceive or control.

In summary, this dissertation follows the definition of persuasion proposed by Piskorski et al. [28] and treats persuasion as a general communicative strategy aimed at influencing readers, which may be used with or without malicious intent. Manipulation is understood as a distinct, intent-driven persuasion, characterized by exploitative goals and deceptive practices. Maintaining this distinction allows for a more nuanced analysis of influence in text and provides a principled foundation for separating benign persuasive communication from harmful disinformation and manipulation.

2.2.3 Malicious Intent

Malicious intent constitutes a core analytical dimension in contemporary disinformation research, distinguishing disinformation from other forms of problematic information such as misinformation, satire or parody [3]. In line with the

definition proposed by the High-Level Expert Group of the European Commission, disinformation is not merely characterized by falsity or inaccuracy, but by the presence of an *intention* to cause public harm or to generate profit [3]. This intentional dimension implies purposive action: disinformation is produced and disseminated to achieve specific goals that extend beyond the mere transmission of incorrect information.

Building on this premise, malicious intent can be understood as the underlying goal structure that motivates the creation and dissemination of disinformative content. As argued by Hameleers [4], a comparative review of definitions of disinformation proposed by different researchers reveals that the intentional dimension is one of the most consistently recognized elements across the literature. While definitions vary in terms of scope, actors, and techniques, they converge on the understanding that disinformation involves purposive deception rather than accidental error. On this basis, Hameleers [4] concludes that disinformation can be defined as intentionally created or disseminated deceptive content, designed to mislead recipients in pursuit of political, ideological, or economic objectives.

In this dissertation, malicious intent is operationalized as a higher-level abstraction that generalizes across and groups recurring disinformation narratives. Disinformation narratives are observable, recurring patterns found across several disinformation articles [7]. Malicious intent, by contrast, encapsulates the broader strategic objective that guides the selection and deployment of such narratives.

In our understanding, this distinction is analytically important for three reasons. First, it enables the systematic study of disinformation beyond isolated false claims or specific disinformation narratives, allowing researchers to identify patterns of coordinated or recurrent behavior across topics, platforms, and time. Second, intent-based analysis provides explanatory depth: rather than asking only what is false, it asks why a particular form of falsity is produced and to what end. As prior research emphasizes, uncovering intent is essential for understanding how disinformation seeks to influence public beliefs, erode trust, or mobilize specific audiences Hameleers [4]. Third, analyzing the intentions of malign actors and their targeted disinformation campaigns can offer a starting point for legal and policy interventions addressing the causes of deliberately false information Hameleers [4].

As a result, malicious intent is not only a theoretically meaningful construct but also practically applicable in real-world settings, such as fact-checking, and, potentially, legal and policy contexts. Precisely because malicious intent can be used in practice, the intent taxonomy proposed in this work was developed in

close cooperation with domain experts and practitioners from fact-checking and debunking organizations accredited by the International Fact-Checking Network (IFCN)[1]. These experts bring extensive experience in analyzing disinformation campaigns across domains and platforms, enabling the translation of theoretical intent constructs into a set of empirically grounded, annotatable categories. This expert-driven approach ensures that the proposed taxonomy reflects not only conceptual rigor but also real-world applicability in contemporary disinformation ecosystems.

Building on this practically grounded approach, malicious intent is treated in this work as a multi-label construct, acknowledging that a single disinformative article may simultaneously pursue multiple objectives and spread multiple disinformation narratives. For example, an article may seek to undermine trust in public institutions while also promoting anti-scientific views or reinforcing social antagonisms. Such overlap reflects the complex and strategic nature of disinformation campaigns, which often exploit multiple vulnerabilities within the information environment.

To conclude, this work conceptualizes malicious intent as a central and defining component of disinformation, capturing its purposeful and strategic nature in its creation and dissemination. By abstracting beyond individual narratives, the notion of intent allows disinformation to be examined in terms of broader goals and impact. Informed by both existing theory and professional fact-checking practice, this perspective acknowledges the complexity of disinformation campaigns and provides a coherent foundation for the intent taxonomy and analytical framework developed in the subsequent chapters.

[1]The IFCN accredits fact-checking and debunking organizations that adhere to its code of principles. See https://www.poynter.org/ifcn/

3 Literature Review

3.1 Disinformation Detection with Language Models

Research on disinformation detection has expanded rapidly in recent years, reflecting its growing impact on digital communication and societal trust [35, 36, 37, 38]. Although substantial work has focused on identifying fake news in online content, it is essential to distinguish between *fake news* and *disinformation*. Fake news is commonly defined as *"fabricated information that mimics news media content in form but not in organizational process or intent"* [39]. This term, however, is insufficient to represent the broader and more complex phenomenon of disinformation, even though fake news constitutes a subset of it [40]. Consequently, research in these two areas often overlaps, and many studies address them jointly.

Several recent surveys have provided comprehensive overviews of methods for detecting fake news and disinformation, examining diverse models, feature sets, definitions, and problem formulations [41, 42, 43, 44]. A central observation across this body of work is that most approaches treat disinformation detection as a supervised classification problem. The predominant setup involves binary classification, typically *real* vs. *fake* or *disinformation* vs. *reliable information*. Numerous systems rely on fine-tuned transformer-based models, especially BERT and its variants, demonstrating strong performance in such settings [8, 9, 45, 46, 10]. Prior work also emphasizes that fake news is generally crafted with the intent to mislead [47], motivating continued research into improving the linguistic and contextual understanding of detection models.

Beyond standard fine-tuning, recent findings indicate that pretraining or auxiliary training on related tasks can enhance performance in disinformation de-

tection. For example, pretraining on fine-grained sentiment analysis has been shown to provide beneficial inductive biases for distinguishing misleading from reliable content [48]. Traditional machine learning and deep learning approaches have additionally incorporated lexical, semantic, and engagement-based features [49, 50, 51], reflecting the multifaceted nature of disinformation signals. Given the high-stakes applications of these models, explainability has become a critical research priority. Hybrid approaches combining deep neural architectures with feature-level explanations improve transparency and user trust [52, 53, 54, 55]. The increasing prevalence of both human-written and LLMs generated disinformation has further stimulated research interest [13, 56].

A persistent challenge in the field is the scarcity of high-quality annotated datasets. According to Capuano et al. [41], most human-labeled datasets focus on political news in English, such as the political dataset provided by Wang [57]. The global impact of the COVID-19 pandemic, along with the rapid spread of related disinformation, has prompted the creation of several datasets specifically targeting COVID-19 and vaccine misinformation [58]. Other expert-labeled medical information datasets include the work of Nabożny et al. [59]. Early datasets in the field, such as LIAR [57], contain short statements annotated for veracity, while FakeNewsNet [60] enriches news data with social and temporal context.

The scarcity of high-quality annotated datasets especially in the non-English language prompted growing interest in zero-shot and few-shot learning methods. Such techniques leverage the adaptability of large pretrained transformers, enabling effective performance even without task-specific supervision [61, 62, 63, 64]. Recent studies have demonstrated that zero-shot LLMs, such as GPT-4, can outperform fully supervised models like BERT in detecting various forms of disinformation [65, 66, 67]. These findings highlight the potential of large-scale language models to generalize across domains and content types without expensive annotation, positioning them as a promising direction for future research in disinformation detection.

Disinformation is intentionally misleading [47]. In response, the Vietnamese dataset RMDM [68] introduces four distinct labels: *real*, *misinformation*, *disinformation*, and *malinformation*, to distinguish between unintentional errors and deliberate attempts to cause harm. Different studies have shown that disinformation often uses persuasion and manipulation to mislead audiences [11, 12, 69, 70]. First attempts to use persuasion as intermediate labels in healthcare misinformation detection within a few-shot scenario have shown promising potential [71].

3.2 Computational Approaches to Persuasion Detection

Research on persuasion detection intersects deeply with the broader study of manipulation, propaganda as these phenomena share conceptual foundations and often rely on overlapping rhetorical strategies. Recent efforts have introduced multilingual datasets annotated for persuasion techniques, improving system evaluation and generalizability across languages and contexts [72]. Social media has emerged as a particularly important domain in this regard, with studies examining political persuasion, the role of opinion leaders, and exposure to dissenting views [73, 74]. LLMs-based approaches have also been developed to simulate multi-agent conversations, assess persuasiveness, and identify techniques such as trust-building and logical reasoning to counter disinformation [75, 76, 77].

Although persuasion and manipulation have distinct characteristics, they share significant theoretical and methodological commonalities, and they also overlap with propaganda [72, 78]. Several datasets have been created to support the study of manipulation and fallacies as forms of persuasive reasoning. Jin et al. [79] introduced LOGIC, comprising 13 classes of fallacious arguments drawn from online educational sources, although it includes no non-fallacious examples. Habernal et al. [80] presented *Argotario*, a game designed to teach argumentation fallacies while collecting data annotated with five fallacy types, including non-fallacious cases, later used for classification experiments with neural and feature-based models [81]. Alhindi et al. [82] released the CLIMATE dataset containing expert-annotated fallacies across 679 text segments from climate change articles, covering 10 categories including a *No Fallacy* class [83].

The study of persuasion is also tightly connected to propaganda detection. Early work focused on document-level judgments, such as the four-way distant-supervision classification task introduced by Rashkin et al. [84] (trusted, satire, hoax, propaganda). Barrón-Cedeno et al. [85] expanded this line of research by proposing a binary corpus comparing propaganda to non-propaganda and analyzing stylistic and readability features, ultimately showing that models sometimes learn to infer source characteristics rather than propaganda cues. As the field progressed, research shifted from coarse document-level labels to the identification of specific persuasive or propagandistic techniques. Habernal et al. [80, 81] annotated arguments with five fallacies, while Da San Martino et al. [86] created a corpus with 18 fine-grained propaganda techniques and explored tasks such as sentence-level detection supported by a multigranular gated neural network. This body of work evolved further through the Prta system [87], followed by models addressing transformer limitations [88] or aiming for greater

interpretability [89]. Work by Da San Martino et al. [86, 90] was extended in a multilingual setting by Piskorski et al. [72], who defined 23 techniques grouped into six coarse-grained categories. Recent work has increasingly leveraged large language models to identify rhetorical or credibility cues in text, using weakly supervised methods to annotate persuasion strategies and enhance veracity classification [91, 92].

Beyond textual news, research has expanded to multimodal and multilingual persuasion detection, including techniques in memes [93], code-switched content [94], and analysis of propaganda in coordinated communities in social media [95]. Additional studies explored how propaganda connects to metaphor [96], fake news dynamics [97], and COVID-19–related online content [98, 99]. A comprehensive overview of computational propaganda research is provided in Da San Martino et al. [100].

This growing interest has led to multiple shared tasks centered on persuasion and propaganda detection. SemEval-2020 Task 11 [90] targeted span and technique identification for 14 propaganda strategies in news articles, while NLP4IF-2019 [101] addressed fine-grained detection of 18 techniques. SemEval-2021 Task 6 extended these efforts to multimodal memes [102], and WANLP'2022 organized a shared task on detecting 20 techniques in Arabic tweets [103]. More recently, the Slavic NLP Workshop introduced tasks on persuasion detection and multilabel persuasion classification at the paragraph level in parliamentary debates and social media across Slavic languages, including Polish and Russian [104].

Parallel research has focused on the persuasive capabilities of generative artificial intelligence. This is being addressed by many fields of science, including computer science, social sciences, and complexity science [105]. Recent progress in large language models has drawn attention to their potential for persuasion and related applications [14, 16]. Early studies by Wang et al. [106] explored personalized persuasive dialogue systems designed to promote socially beneficial outcomes. Subsequent studies have investigated how individuals respond to persuasive machine-generated text and how they perceive its effectiveness [107, 108, 15]. In addition, Schoenegger et al. [109] explored whether LLMs can be more persuasive than humans, while Pauli et al. [110] analyzed the extent to which LLMs are capable of generating persuasive language across different domains.

3.3 The Intentional Dimension of Disinformation in NLP

Intent (or *intention*) discovery has been approached from multiple perspectives within NLP and beyond. Prior work ranges from purely textual analyses [111] to multimodal settings where text is complemented with images [112] or videos [113]. Research has examined how intentions relate to human behavior [114] and how they guide individuals toward achieving goals [115]. Other studies focus more explicitly on identifying and categorizing intentions, particularly in the context of disinformation.

A central contribution to conceptualizing disinformation intent is provided by Hameleers [4], who propose a framework connecting actors, intentions, and techniques used to create and disseminate misleading content. They outline four categories of malicious intent: delegitimization, mobilization, ideological motivations, and financial gain. While this framework is useful for understanding intent categories, it does not include annotated datasets or empirical evaluations with language models.

Several works investigate user or agent intent within broader misinformation or fact-checking pipelines. Gupta et al. [116] examine user intent behind fact-checking queries and develop a chatbot to counter fake news, but do not explore whether intent features can enhance disinformation detection. In contrast, Zhou et al. [117] focus on the intent of fake news spreaders, algorithmically labeling intents and comparing them with a subset of manual annotations. Their findings indicate that incorporating user propagation intent into Heterogeneous Graph Neural Networks yields slight performance improvements over prior models. They also observe that most users spreading fake news do so unintentionally, suggesting that concentrating on the intent of disinformation agents, rather than spreaders, may be more insightful.

Complementary to this, Wang et al. [118] annotate public real/fake news datasets with agent-level intent classes such as *Public*, *Emotion*, *Individual*, *Popularize*, *Clout*, *Conflict*, *Smear*, *Bias*, and *Connect*. They demonstrate that adding intent features improves misinformation detection in a T5-based architecture.

Beyond spreader-focused intent, other research examines intent in news creation. Notably, Wang et al. [119] introduce a formal definition and analytical framework for news creation intent, reflecting the recognition that intention is fundamental to news understanding [120, 121]. Zhou et al. [122] contribute the first assessment of intentional versus unintentional fake news dissemination via an *influence graph*. Guo et al. [123] extend this by categorizing spreading intent into five classes, though these studies remain centered on news and fake news rather than disinformation more broadly.

Despite substantial progress in disinformation research [124, 125, 126, 127, 13], the intent behind the creation of disinformation content remains insufficiently explored. Existing work primarily considers spreader intent or high-level conceptual frameworks, leaving a gap in empirical studies directly addressing disinformation agents' malicious intentions.

3.4 Conclusions and Knowledge Gaps

Despite substantial progress across disinformation, persuasion, and intent detection, the literature reveals several critical gaps that motivate the contributions of this dissertation.

The Role of Persuasion in Disinformation Detection. The second cluster of research gaps concerns the role of persuasion, which is a broader concept than manipulation, as an intermediate signal for disinformation detection. Although many studies acknowledge the persuasive nature of disinformation, prior work does not provide a structured, model-agnostic framework for systematically incorporating persuasion knowledge into disinformation detection pipelines. Existing studies show only localized and domain-specific improvements, for example, in the detection of healthcare misinformation, but the field still lacks a method that can generalize across different Large Language Models and datasets. So far, no study has demonstrated a clear, measurable benefit of applying persuasion knowledge to disinformation detection.

The Role of Malicious Intent in Disinformation Detection. A third significant gap in the literature concerns the integration of malicious intent as a core signal for disinformation detection. Although intent is a defining component of disinformation in widely accepted policy and academic definitions, research in natural language processing has paid remarkably little attention to this dimension. Existing work provides almost no empirical investigation of intent in disinformation, and the field lacks high-quality human-annotated resources that capture malicious intent in disinformative content. Furthermore, current approaches offer no model-agnostic frameworks for incorporating intent signals into inference pipelines, and no study to date has systematically examined whether malicious intent can enhance the reasoning processes of large language models.

Disinformation Research in Low-Resource Languages and Novel Taxonomies. A significant gap in current research is the overwhelming focus on English-

language disinformation, leaving limited resources and systematic studies for lower-resource languages. Existing approaches often treat disinformation detection as a binary classification task, failing to account for the nuanced dimensions of malicious intent and manipulation strategies. Additionally, the field lacks comprehensive, systematic taxonomies that categorize manipulation techniques and malicious intent, which are critical for understanding the mechanisms underlying disinformation. There is also a shortage of studies benchmarking natural language processing models on non-English datasets, particularly for languages such as Polish, which limits the generalizability of findings across diverse linguistic contexts. More broadly, current research tends to conceptualize disinformation detection in a simplified manner, overlooking frameworks that integrate both manipulation techniques and intent as essential analytical components.

Understanding and Detecting LLM-Generated Persuasion. A further gap concerns the growing role of LLMs as generators of persuasive content, an area where scientific understanding remains limited. While recent studies examine the persuasive capabilities of large language models, prior research has not addressed whether LLM-generated persuasive texts are easier or harder to detect than human-written persuasion. Moreover, the field lacks a systematic, comprehensive linguistic comparison of human- and machine-produced persuasive messages. Existing work typically evaluates persuasiveness at a behavioral level (e.g., whether people are convinced), but does not analyze the linguistic features that differentiate LLM-generated persuasion from its human counterpart. As a result, there is no established benchmark for studying LLM-specific persuasion, nor any empirical assessment of how well current detection models perform on this emerging content category. This gap is particularly consequential given the rapid proliferation of AI-generated persuasive text in political communication, advertising, social media, and disinformation campaigns. A better understanding of LLM-generated and human-written persuasive texts is particularly important as it may offer insights that may guide the development of more interpretable and robust persuasion and disinformation detection tools.

4 Benchmarking Models on Polish Disinformation Detection

This chapter presents contributions C1.1 and C1.2 described in Section 1.3. The chapter is based on the following published paper:
Arkadiusz Modzelewski, Giovanni Da San Martino, Pavel Savov, Magdalena Anna Wilczyńska, and Adam Wierzbicki. *MIPD: Exploring Manipulation and Intention In a Novel Corpus of Polish Disinformation*. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, Florida, USA. Association for Computational Linguistics (EMNLP 2024).

Mitigating the spread of disinformation on the web has become an important social challenge. Numerous significant events, including the COVID-19 pandemic and the Russo-Ukrainian conflict, highlight disinformation's negative impact on individuals and society [128, 129].

The High-Level Group of Experts set up by the European Commission defines disinformation as *"false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit"* [3]. There are two significant aspects in this definition: intention types ("the why") and misleading manipulations ("the how"). However, to our knowledge, no study in the literature examines intention types and manipulation in disinformation together, possibly due to a lack of quality annotated data. Therefore, we share with the research community the **M**anipulation and **I**ntention in **P**olish corpus of **D**isinformative web articles: the **MIPD** dataset. The MIPD dataset sheds light on the authors' intention and manipulation techniques in disinformation. Our high-quality open corpus, annotated by five professional fact-checkers and debunkers, will provide a multifaceted understanding of disinformation. Initially, we focus on Polish, the 5th most spoken language in the European Union [130].

We chose this language because it is the largest of the V4 countries (Slovak Republic, Czech Republic, Poland, and Hungary), which have been particularly vulnerable to disinformation in recent years due to the Russo-Ukrainian conflict [131, 132].

4.1 Construction and Annotation of the MIPD Dataset

MIPD is a novel dataset that includes 15,356 web articles in Polish. In addition to the article content, the data we publish contains four annotations: *(i)* whether an article is disinformative or credible; *(ii)* the intention types; *(iii)* the manipulation techniques used in the article; *(iv)* the thematic category of the article. Additionally, we publish the sources from which we derived our articles. What distinguishes this corpus is its focus on uncovering the hidden malicious intent and manipulation techniques within disinformative content. Each article was annotated by experts using a multifaceted methodology. This resource provides a foundation for benchmarking Polish Language Models on the task of disinformation detection.

4.1.1 Annotation Process

We can break down our annotation process into four stages:

1. **Methodology creation stage** - professional annotators and researchers collaborated to develop a methodology for annotating articles.
2. **Initial annotation stage** - the most experienced specialist and the leader of the annotators' group trained other less experienced participants. In this initial step, the annotators tested the methodologies on a small sample of articles.
3. **Article annotation stage** - each text was annotated independently by at least two annotators. We included articles in our dataset if the annotations from experts were the same. If not, the article passed to the fourth annotation stage.
4. **Annotation consensus** - if the evaluations of at least two experts did not match, the annotators met and discussed their evaluations, seeking consensus. The lack of consensus resulted in adding a *hard-to-say* label to an article. Article with *hard-to-say* label was excluded from our dataset. The discussion and development of consensus have always occurred during face-to-face meetings.

Additionally, our annotators divided the labeling phases into subject areas. They labeled articles topic by topic. Each time, they underwent additional training provided by the most experienced person in a specific thematic area before the annotation process. The training ensured in-depth understanding and accurate identification of disinformation.

In order to ensure high-quality annotations, our annotation guidelines and methodology were created by fact-checking and debunking experts. We employed five Polish native-speaker experts with at least three years of fact-checking and debunking experience (on a one-year competitive salary). All debunking experts working on the project were previously employed in debunking organizations with the accreditation of the International Fact-Checking Network[1]. The same experts used the methodology to annotate the articles in MIPD. The methodology described here is also an educational tool for students who wish to learn how to detect disinformation.

The methodology is divided into five main steps:

1. Determining the article's thematic category.
2. Evaluating the credibility of the article's source and author (if known).
3. Determining the article's main class: *credible*, *disinformation*, *misinformation* or *hard-to-say*.
4. For disinformation, evaluating of manipulation.
5. For disinformation, evaluating of intention types and narratives.

Experts could return to previous steps and typically re-evaluate the main class after a detailed investigation of an article suspected to contain disinformation.

4.1.2 Data Sources

We selected articles from more than 400 sources, each being freely available and not requiring any subscription. Our articles partially come from general and common news sources operating within the public sector, i.e., official sources managed by the government and its institutions. We also incorporate articles from alternative and independent opinion-oriented media, websites, and blogs sharing scientific insights. Additionally, we collected articles from websites containing conspiracy theories and Russian propaganda. The list of sources is not exhaustive. We aimed to collect the least biased dataset possible. Therefore, we focused on including the broadest spectrum of views and beliefs. We publish our dataset and the sources from which we obtained the articles.

4.1.3 Thematic Category

Given a web article, we start with an initial content analysis and determine the topic. Categorizing web articles into thematic domains enables future research on distinct features and patterns within different disinformation topics. Our

[1]The International Fact-Checking Network gives accreditation to debunking organizations that sign its code of principles. See https://www.poynter.org/ifcn/

assessment allows us to classify the articles into one of 10 detailed thematic categories. We base our taxonomy of thematic categories on a prior analysis of the work of fact-checking and debunking organizations, such as Snopes[2], "Counteracting Disinformation" Foundation[3], Demagog[4] Association, and Debunk EU[5]. We consider the following categories (if created, acronym in parentheses): *COVID-19 (COVID)*, *Migrations (MIG)*, *LGBT+*, *Climate Crisis (CLIM)*, *5G*, *War in Ukraine (WUKR)*, *Pseudomedicine (PSMED)*, *Women's rights (WOMR)*, *Paranormal Activities (PA)*, *News or Other (NEWS)*. The topics in our dataset significantly overlap with the most significant disinformation topics published in the recent EU DisinfoLab report [133].

4.1.4 Evaluation of Source Credibility

For each article, the experts evaluate the credibility of the article's source (publishing portal or organization) and author (if known). Source and author credibility did not determine the overall evaluation of the article, but the experts maintain a list of sources with their credibility evaluation. The experts used this list to search for the next articles for evaluation. Sources were evaluated in three classes: *reliable*, *unreliable* or *mixed*. Articles from unreliable sources could be evaluated as credible, while articles from other sources could be evaluated as disinformation.

4.1.5 Main Credibility Evaluation

Given a web article, annotators identify from its content whether it contains *disinformation*, *misinformation*, or *credible information*. Annotators could also use a fourth category *hard-to-say*.

In our annotation methodology, we adopt a disinformation definition provided by the European Commission's HLEG group (see Section 4). Disinformation is intentionally misleading or false. Unlike disinformation, misinformation is *misleading information shared by people who do not recognize it as such* [3].

We exclude articles with *misinformation* and *hard-to-say* labels from the primary published dataset. In this study, we wanted to focus on a binary classification: disinformation versus credible articles.

[2] Snopes

[3] Counteracting Disinformation

[4] Demagog

[5] Debunk EU

4.1.6 Manipulation Techniques

Debunking experts identify the usage of manipulation techniques in disinformative articles. The annotation of manipulation techniques is a multiclass multilabel problem. The following presents our taxonomy and short descriptions of manipulation techniques adopted in our annotation methodology:

- **Cherry Picking [CHP]** Presenting information utilizing only data that supports a given hypothesis or argument, while ignoring the broader context [134].
- **Quote Mining [QM]** Using a short fragment of someone's longer speech in a way that significantly distorts its original tone [135].
- **Anecdote [AN]**. The use of evidence in the form of personal experience or an isolated case, possibly rumor or hearsay, most often to discredit statistics [136].
- **Whataboutism [WH]**. Responding to a substantive argument not by addressing the heart of the matter, but by raising a new point that is unrelated to the topic at hand. [137].
- **Strawman [ST]**. It involves distorting someone else's argument in a way that makes it easier to refute it. It is often done by attributing a stance to opponents, who do not share it. [138].
- **Leading Questions [LQ]**. Flooding the recipient with a series of consecutive suggestive questions or putting them together leads the recipient to a predetermined thesis [139].
- **Appeal to Emotion [AE]**. The use of words and phrases that are to arouse in the recipient extreme emotion and attitude to the presented matter [140].
- **False Cause [FC]**. The individual employing this technique assumes a cause-and-effect relationship solely based on the observed correlation [141].
- **Exaggeration [EG]**. The author overstates a phenomenon, making it appear larger, better, or worse, or oversimplifies a phenomenon making it seem less significant or smaller than it truly is [142, 86].
- **Reference Error [RE]**. In this technique, the author refers to fake experts, propaganda statements made by politicians, anonymous entries published on social media, or false quotes from famous people to authenticate the presented thesis [143]. It may present false choices and false analogies.
- **Misleading Clickbait [MC]**. A technique involves giving a title to the text that misrepresents or contradicts the content discussed within the article. Title created with a purpose to attract attention [144].

Manipulation and persuasion techniques have a lot in common. Detection of the latter has already been examined in previous studies, such as in work done by Da San Martino et al. [86]. Manipulation can be seen as distinct from

persuasion in that it is concerned not with changing individuals' beliefs but with inducing them into choices that the manipulator desires [29]. Therefore, we can assume that a manipulation technique is always used with malicious intent, which is also explored in our methodology and the MIPD dataset. On the other hand, persuasion techniques can be used without malicious intent (for example, persuading individuals to stop smoking or make other better health choices).

Our list of manipulation techniques includes techniques not considered in previous studies, e.g., in Da San Martino et al. [86], such as *Cherry-Picking* and *Quote Mining*. More about taxonomy of manipulation techniques in MIPD dataset available in Appendix A.1.

4.1.7 Malicious Intention Type

Debunking experts explore the intention types and narratives of creators of disinformative articles. Classifying the creator's intention in disinformative articles allows us to understand their characteristics and detect patterns in disinformation content. In our methodology, each intention corresponds to several narratives. An intention is a generalization of a narrative that we can define as a repeating pattern found in several disinformative articles [133]. Intention encapsulates the broader goal of the author, which guides specific narratives used to achieve that goal.

Figure 4.1 provides a breakdown of our taxonomy and brief explanations of the intention types. The annotation of intention is a multiclass multilabel problem.

4.1.8 Impartiality and Bias Prevention

To avoid bias in the dataset, our methodology requires each article to be annotated independently by two experts. Due to the complexity and time cost of the evaluation (the evaluation of a disinformative article took 30 minutes on average), we could only assure two evaluations per article. Instead, in case of disparity in an article's annotations, the two evaluating experts attempted to reach a consensus. We removed all articles that did not reach consensus from our dataset. All article annotations in the dataset are the result of a consensus between two expert annotators. During the consensus building, annotators discussed their interpretation of the methodology. Therefore, double verification helped to avoid biases and human errors while also serving as a standardization of the methodology's application.

Negating Scientific Facts [NSF]: Authors deny established scientific facts, such as challenging the existence and severity of COVID-19, promoting alternative treatments, questioning the safety of 5G, and denying the reality of climate change and human impact on the environment. The objective is to create skepticism and erode public trust in scientific consensus.

Undermining the Credibility of Public Institutions [UCPI]: Authors try to erode trust in public institutions by engaging in activities, i.e., discrediting pandemic control measures, reproaching human rights violations, negating defense capabilities, and undermining strategies addressing migration and climate crises. These actions weaken the trust and confidence in the reliability and authority of government bodies and public organizations.

Challenging an International Organization [CIO]: Involves a deliberate effort to erode confidence in international organizations like the EU, WHO, UN, and NATO by disseminating content that blames them for regional conflicts, accuses them of aggression against specific countries, undermines defense capabilities, and discredits international climate agreements.

Promoting Social Stereotypes/Antagonisms [PSSA]: Authors promote social stereotypes and antagonisms through tactics such as enhancing homophobia, transphobia, xenophobia (linked to economic, security, and health aspects), religious conflicts, and anti-semitism.

Weakening International Alliances [WIA]: Authors disseminate false or misleading information to undermine the strength and unity of partnerships between countries. The goal is to create doubt among allied nations, undermining the trust and cooperation necessary for their mutual security and strategic interests.

Changing Electoral Beliefs [CEB]: Authors influence public opinion, especially during elections. Authors with this intention capitalize on exploiting public sentiments surrounding sensitive issues such as LGBT rights and migrations to sway voters, polarize opinions, and potentially impact political decisions during elections.

Undermining International Position of a Country [UIPC]: Authors spread claims aimed at deteriorating a nation's global standing by accusing it of meddling in the political processes of other countries. Authors may erode trust and confidence in the state's governance and humanitarian standards. It seeks to damage the state's reputation on the international stage through unfounded allegations.

Causing Panic [CP]: Authors spread false information to incite fear and unrest among the public. This strategy exploits readers' emotions to destabilize societal trust and order.

Raising Morale of a Conflict's Side [RMCS]: Authors intend to boost the spirit and confidence of a particular group involved in a conflict. It aims to positively influence supporter perceptions and commitment towards their side's objectives and actions.

Figure 4.1: Malicious intention types with a brief description. We give an acronym for each intention in brackets.

4.1.9 Dataset Quality

We evaluate inter-rater reliability using a consensus measure. Consensus estimates of inter-rater reliability assume that annotators can agree on their evaluations. It is most suited for nominal evaluations where different scale levels represent qualitatively different ideas [145]. In our case, the main annotation class includes categories: *credible*, *disinformation*, *misinformation*, and *hard-to-say*. The difference between credible information and disinformation is complex to describe. This complexity is evident in the subsequent steps of the methodology, which aim to illustrate different aspects of disinformation. Similarly, evaluating manipulation techniques and intention types requires using qualitatively different concepts for each rating level.

Statistic	PA	CLIM	COVID	5G	LGBT+	MIG	NEWS	PSMED	WUKR	WOMR	All
AVG_w	724	736	804	756	633	716	662	978	782	708	767
AVG_{ch}	5,062	5,280	5,764	5,471	4,552	5,091	4,672	7,085	5,517	5,046	5,485
$\#DOC$	1,046	1,011	6,049	1,048	1,036	1,030	1,033	1,013	1,026	1,064	15,356

Table 4.1: Data statistics per thematic category: average article length in number of words (AVG_w), average article length in number of characters (AVG_{ch}), number of articles (#DOC). Acronyms in columns provide information about topic (see subsection 4.1.3)

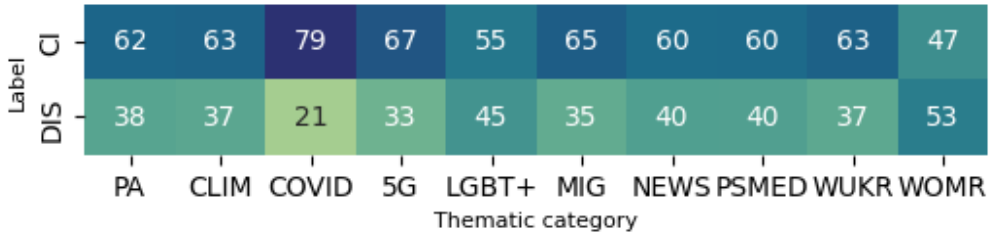

Figure 4.2: Percentage of disinformative (DIS) and credible (CI) articles per thematic category.

In total, 15,510 articles in our dataset had two independent annotations. After the double independent annotations, experts attempted to establish consensus. Our experts did not reach a consensus in 49 cases (we removed these 49 articles from the dataset). The percentage of articles that reached consensus is 99.69%. However, during the consensus-building process, annotators could agree that the annotation should be *hard-to-say*. Experts placed 105 articles in that category. We have removed these articles from the dataset, considering them as articles that did not reach a consensus. Therefore, the final percentage of articles that reached consensus as credible, disinformation, or misinformation is 99%.

4.2 Multi-Dimensional Data Analysis

In Table 4.1, we present some statistics per different thematic categories, such as the number of articles, average number of words, and average number of characters in the article. Figure 4.2 shows the percentages of articles with credible information and disinformation per thematic category. We publish 10,359 credible articles and 4,997 articles with disinformation. Details about the number of articles with specific intention type are presented in Table 4.2 and manipulation techniques in Table 4.3.

We present in Figure 4.4 and Figure 4.3 the percentages of articles with specific manipulation techniques and intention types per thematic category. These

NSF	UCPI	CIO	PSSA	WIA	CEB	UIPC	CP	RMCS
2879	1522	915	1887	296	294	113	96	122

Table 4.2: Number of articles per intention type.

CHP	QM	AN	WH	ST	LQ	AE	FC	EG	RE	MC
1526	125	442	426	434	127	909	915	2153	1108	177

Table 4.3: Number of articles per manipulation technique.

Figures show that neither manipulations nor intents are specific to the topic of the articles. One may consider the *Raising Morale of a Conflict's Side* intention type specific to the *War in Ukraine* topic. Nevertheless, the RMCS could likely be applicable when analyzing articles about other conflicts. In addition, we observe that a single manipulation technique can be used in articles with different intention types. Furthermore, an article designed with a particular intention may contain various manipulation techniques.

4.3 Experiments

Our experiments aim to test the data quality further and provide a baseline upon which future works can build.

4.3.1 Models and GPU

Small Language Models. We fine-tuned two pre-trained Polish BERT-based Language Models: HerBERT [146], and Polish RoBERTa [147]. We chose these two models because they are the ones that perform best on the KLEJ[6] Benchmark. KLEJ Benchmark is a comprehensive benchmark for Polish Language Understanding [148]. We used the HerBERT-base (HerBERT-B) and Polish-RoBERTa-v2-base (PL-RoBERTa-B) versions as well as the larger models: HerBERT-large (HerBERT-L) and Polish-RoBERTa-v2-large (PL-RoBERTa-L) versions. As of 30.11.2025, these models are available on HuggingFace[7] under the following names:

- `sdadas/polish-roberta-large-v2`
- `sdadas/polish-roberta-base-v2`

[6] KLEJ Benchmark leaderboard accessed on 15th April 2024 https://klejbenchmark.com/leaderboard/

[7] https://huggingface.co/models

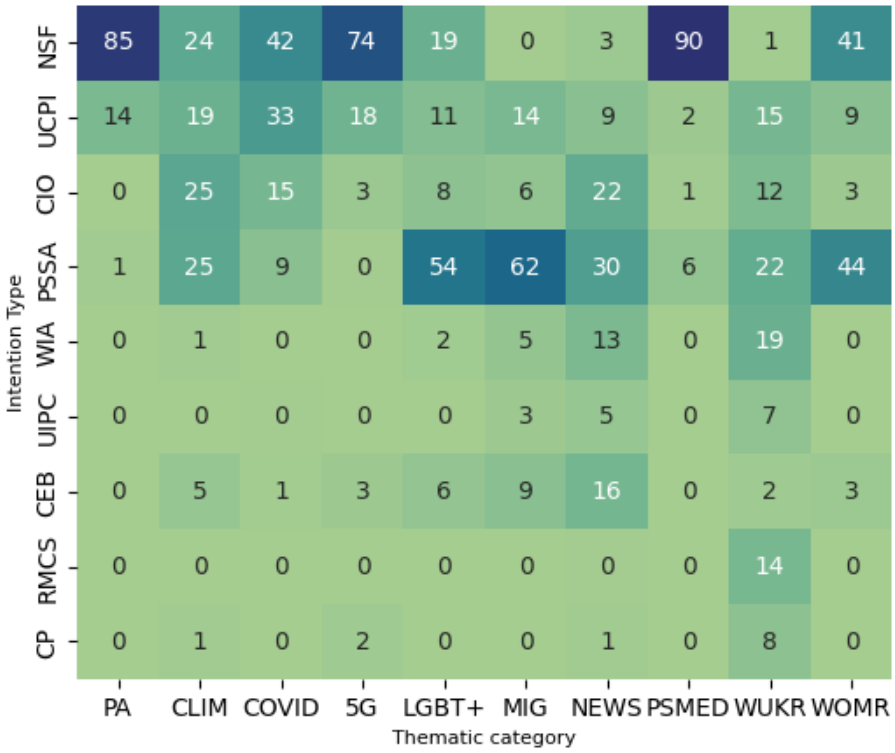

Figure 4.3: Percentage of different intention types per thematic categories among articles with malicious intention

- `allegro/herbert-base-cased`
- `allegro/herbert-large-cased`

For our computations to find the optimal hyperparameters and final fine-tuning of the models, we used the NVIDIA L40 GPU.

Large Language Models. In addition, we decided to explore the efficacy of generative models in disinformation classification. Specifically, we experimented with two OpenAI generative models that are accessible via their APIs: GPT-3.5 and GPT-4[8].

[8]Details on the models used: We utilized a snapshot of GPT-4 from June 13th, 2023, named `gpt-4-0613`, and `gpt-3.5-turbo-instruct`, which has capabilities similar to GPT-3 era models. The last access to these models was on 28th May 2024.

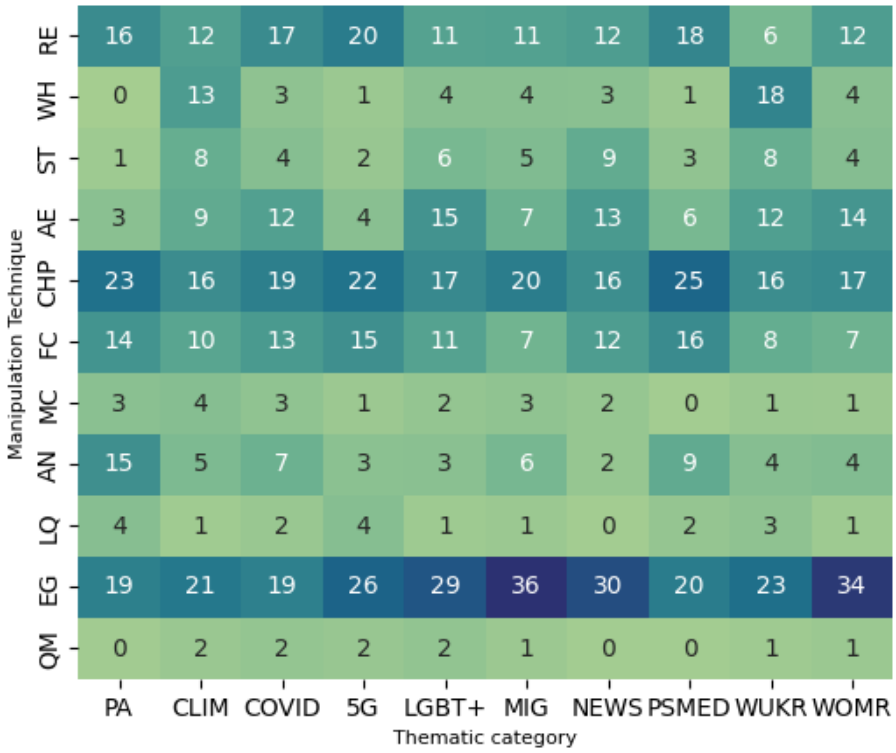

Figure 4.4: Percentage of different manipulation techniques per thematic category among articles with manipulation.

4.3.2 Experimental Setup

Small Language Models. We began our experiments by dividing the data into train and validation in the proportions of 70%/30%. From the validation set, we randomly selected about 30% of the data as a test dataset. At the end of data preparation, we got datasets segmented into train/validation/test sets comprising 10,749 articles for training, 3,086 for validation, and an additional 1,521 for testing purposes. Next, we utilized our data and models to identify the optimal hyperparameters for training the model for disinformation binary classification and two multilabel multiclass tasks: manipulation and intention type classification. We accomplished this by performing a hyperparameter search for various learning rate values (ranging from 1e-6 to 1e-4) and weight decay (ranging from 0.005 to 0.2). Additionally, we implemented a linear warmup for the first 6% of the training steps. Batch size was not tuned for optimal value. We assumed 16 for train and evaluation batch size. We performed hyperaprameters tuning for all versions of chosen models. To check all optimal values for learning rate and weight decay see Table 4.4. After fine-tuning models for text classification, we

named the resulting models **PolBERT** when the base model was HerBERT, and **PolBERTa** when the base model was Polish RoBERTa. Our fine-tuned models are publicly available on *HuggingFace* under the MIPD collection[9]. Finally, we used these trained models with optimal hyperparameters to predict the classes of the provided test dataset.

Model	**Disinformation**		**Manipulation**		**Intention**	
	lr	wd	lr	wd	lr	wd
HerBERT-B	3e-5	0.1	1e-5	0.03	1e-5	0.2
PL-RoBERTa-B	2e-5	0.2	1e-5	0.1	3e-5	0.03
HerBERT-L	1e-5	0.03	1e-5	0.02	1e-5	0.03
PL-RoBERTa-L	1e-5	0.02	2e-5	0.01	2e-5	0.1

Table 4.4: Optimal hyperparameters (lr = learning rate, wd = weight decay) for all MIPD tasks: disinformation detection, manipulation multilabel classification and intent multilabel clasisfication.

Large Language Models. Our objective was to assess the ability of GPT-3.5 and GPT-4 models to classify articles as containing disinformation using a zero-shot classification approach. We employed two zero-shot strategies for each model: one without defining disinformation and the other including the definition. The definition we utilized was proposed by the HLEG established by the European Commission (see introduction of the Chapter 4).

First, we randomly drew a sample of 10% of the articles from our entire dataset. Then, we used a prompt to classify articles with generative models (our prompts are available in Appendix A.2). We repeated these steps for two approaches: (i) zero-shot classification with a disinformation definition included in the prompt; (ii) and zero-shot classification without a definition. Finally, we calculated various evaluation metrics, including F_1 score over positive (disinformative) class.

4.4 Results

We computed results on test data that was unavailable during the fine-tuning process. The final result is an average of metric scores produced by models trained with five seeds. Tables 4.5, 4.7, and 4.8 show the final results with their corresponding standard deviations.

[9] https://huggingface.co/collections/ArkadiusDS/mipd

Model	*Acc.*	F_{w}	F_1
HerBERT-B	0.94 ± 0.004	0.94 ± 0.004	0.91 ± 0.007
HerBERT-L	0.95 ± 0.003	0.95 ± 0.003	0.93 ± 0.004
PL-RoBERTa-B	0.94 ± 0.005	0.94 ± 0.005	0.91 ± 0.008
PL-RoBERTa-L	**0.96** ± 0.001	**0.96** ± 0.001	**0.93** ± 0.002

Table 4.5: Results for disinformation detection task. Table shows accuracy (*Acc.*), weighted F_1 score (F_{w}), and F_1 score on test data for pre-trained Polish BERT-based models. The results show the average metrics and their standard deviations, calculated from five different seeds.

4.4.1 Polish Disinformation Detection

Small Language Models. Table 4.5 presents the results of four fine-tuned Polish BERT-based models on a disinformation detection task. This task was a binary classification to distinguish between disinformative and credible articles. Since the dataset is imbalanced, we adopted a weighted F_1 score as the primary evaluation metric. Notably, all models demonstrate high effectiveness. Evaluation metrics indicate minor variations across models. As for other evaluation metrics, the PL-RoBERTa-L model stands out with the highest weighted F_1 score.

Large Language Models. Although our findings with LLMs are preliminary and warrant further in-depth analysis, we present them to demonstrate the potential of generative models in classifying disinformation.

Table 4.6 presents the result of these calculations. Our investigation reveals that Polish BERT-based models fine-tuned on the MIPD dataset significantly outperform chosen generative models: GPT-4 and GPT-3.5. The GPT models, when applied in a zero-shot approach without a definition of disinformation, achieved weighted F_1 scores of 0.84 for GPT-4 and 0.61 for GPT-3.5, respectively. In the zero-shot approach with the given definition of disinformation, both the GPT-4 and GPT-3.5 models improved their results. Nevertheless, these results are inferior to any Polish BERT-based models. Our findings highlight the effectiveness of HerBERT and Polish RoBERTa in handling Polish disinformation, likely due to their specialized training and fine-tuning utilizing Polish datasets. In contrast, the results of GPT models suggest that generative models may require domain-specific fine-tuning to reach the performance of language-specific BERT variants in the disinformation classification task.

4.4.2 Manipulation Techniques Detection

Table 4.7 provides the performance of fine-tuned selected models, detailing results across individual manipulation techniques. Moreover, we show the mod-

Model	Prompt Type	*Acc.*	F_w	F_1
GPT-4	Without Definition	0.85	0.84	0.73
	With Definition	0.86	0.86	0.77
GPT-3.5	Without Definition	0.60	0.61	0.51
	With Definition	0.70	0.70	0.56

Table 4.6: Results of the disinformation detection task for GPT-4 and GPT-3.5, showing accuracy (*Acc.*), F_1 score, and weighted F_1 score (F_w). The results present Zero-Shot Classification with and without a definition of disinformation.

Model	CHP	QM	AN	WH	ST	LQ	AE	FC	EG	RE	MC	F_w
HerBERT-B	0.45	0.00	0.14	0.19	0.27	**0.02**	0.40	0.31	0.64	0.43	0.00	0.42
	±0.01	±0.00	±0.05	±0.03	±0.03	±0.05	±0.02	±0.01	±0.01	±0.01	±0.00	±0.006
HerBERT-L	**0.48**	0.00	0.36	**0.30**	**0.30**	0.00	0.44	0.37	0.66	**0.50**	0.08	**0.47**
	±0.01	±0.00	±0.04	±0.03	±0.02	±0.00	±0.01	±0.03	±0.01	±0.00	±0.01	±0.008
PL-RoBERTa-B	0.44	0.00	0.08	0.20	0.25	0.00	0.38	0.33	0.64	0.38	0.00	0.41
	±0.02	±0.00	±0.08	±0.03	±0.03	±0.00	±0.01	±0.02	±0.01	±0.01	±0.00	±0.011
PL-RoBERTa-L	0.46	0.00	**0.39**	0.26	0.28	0.00	**0.45**	**0.38**	**0.67**	0.48	**0.15**	**0.47**
	±0.01	±0.00	±0.02	±0.05	±0.02	±0.00	±0.00	±0.02	±0.01	±0.02	±0.06	± 0.003

Table 4.7: Results for manipulation techniques classification. Table shows F_1 scores for pre-trained Polish BERT-based models in each manipulation type. Moreover, we present a weighted F_1 score (F_w) for the overall task. The results show the average metrics and their standard deviations, calculated from five different seeds. All evaluation metrics were computed for test data.

els' overall effectiveness in the task using a weighted F_1 score. The HerBERT-L model and PL-RoBERTa-L performed best in this multilabel multiclass task. PL-RoBERTa-L achieved the highest F_1 score for five manipulation techniques. Importantly, *Quote Mining*, *Leading Questions*, and *Misleading Clickbait* were particularly challenging. Specifically, none of the models could detect the *Quote Mining* technique. The decreased performance observed in classifying these three techniques is likely due to their relatively rare occurrence in our dataset.

4.4.3 Malicious Intention Types Detection

In the task of intention classification, PL-RoBERTa-L exhibits the best results, reaching a weighted F_1 score of 0.71. A closer examination of the performance across distinct intention categories presented in Table 4.8 reveals that PL-RoBERTa-L outperforms other models in 8/9 categories of intention types.

Model	UCPI	CEB	UIPC	CIO	WIA	PSSA	NSF	CP	RMCS	F_w
HerBERT-B	0.56	0.19	0.31	0.52	0.46	0.69	0.81	0.22	0.42	0.65
	±0.01	±0.03	±0.07	±0.03	±0.03	±0.01	±0.01	±0.12	±0.04	±0.006
HerBERT-L	0.62	0.27	**0.38**	0.60	0.46	0.71	0.84	0.24	0.51	0.69
	±0.01	±0.05	±0.04	±0.01	±0.03	±0.01	±0.01	±0.08	±0.05	±0.006
PL-RoBERTa-B	0.56	0.17	0.38	0.55	0.48	0.67	0.81	0.25	0.46	0.65
	±0.02	±0.01	±0.04	±0.02	±0.02	±0.00	±0.01	±0.06	±0.04	±0.009
PL-RoBERTa-L	**0.62**	**0.30**	0.37	**0.63**	**0.49**	**0.74**	**0.86**	**0.27**	**0.56**	**0.71**
	±0.01	±0.02	±0.05	±0.01	±0.01	±0.01	±0.00	±0.07	±0.04	±0.005

Table 4.8: Results for malicious intention type classification. Table shows F_1 scores for pre-trained Polish BERT-based models in each intention type. Moreover, we present a weighted F_1 score (F_w) for the overall task. The results show the average metrics and their standard deviations, calculated from five different seeds. All evaluation metrics were computed for test data.

4.5 Discussion

This chapter introduced MIPD, a multifaceted corpus of Polish disinformation. MIPD enables deeper analysis of misleading content and captures dimensions such as manipulation techniques and malicious intent types, rather than reducing disinformation to a binary classification. This design reflects the theoretical view that disinformation is deliberately misleading information.

A key aspect of MIPD is its focus on annotation quality and methodological rigor. Annotation guidelines were created with professional fact-checkers and debunking experts. Experienced specialists from accredited organizations annotated the dataset to ensure data quality and consistency. This expert-driven process enhances annotation reliability. By releasing the dataset and methodology, this work provides a reusable framework adaptable to other languages.

The experimental results obtained using MIPD provide several insights into current disinformation detection approaches. Fine-tuned Polish BERT-based models achieve strong performance across all tasks, including binary disinformation detection and the more challenging multilabel classification of manipulation techniques and intention types. In contrast, experiments with large language models in a zero-shot setting, namely GPT-4 and GPT-3.5, show substantially lower performance on Polish data. This gap suggests that, despite their general reasoning capabilities, supervised and fine-tuned approaches remain crucial for disinformation detection in non-English languages.

While MIPD focuses on Polish, the underlying annotation framework and conceptual distinctions are not language-specific. Extending this approach to additional languages and cultural contexts would allow for further development

of disinformation research in Natural Language Processing. As such, MIPD serves not only as a standalone resource but also as a foundation for broader research on English disinformation in subsequent chapters of this dissertation.

5 Persuasion-Augmented Reasoning for Disinformation Detection

This chapter presents contributions C2.1 and C2.2 described in Section 1.3. The chapter is based on the following published paper:
Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. *PCoT: Persuasion-Augmented Chain of Thought for Detecting Fake News and Social Media Disinformation*. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vienna, Austria. Association for Computational Linguistics (ACL 2025).

The growing accessibility of digital media, coupled with reduced funds for traditional fact-checking efforts and the rise of alternatives like Birdwatch on Platform X (formerly Twitter), underscores the urgent need for complementary disinformation detection systems [149, 150]. Traditional supervised detection methods, which rely on human-annotated data, face challenges in generalization and the scarcity of labeled data. This reinforces the need for zero-shot detection systems.

A critical aspect of disinformation is its coexistence with manipulation and persuasion to mislead audiences [17, 12]. Psychological studies show that teaching individuals to recognize persuasive fallacies improves their ability to distinguish between real and fake news [151]. Building on this, we explored whether infusing knowledge of persuasion into generative LLMs enhances disinformation detection.

As a result, we present **Persuasion-Augmented Chain of Thought (PCoT)**, a novel zero-shot method leveraging persuasion signals to improve disinformation detection that more effectively addresses generalization and annotated data

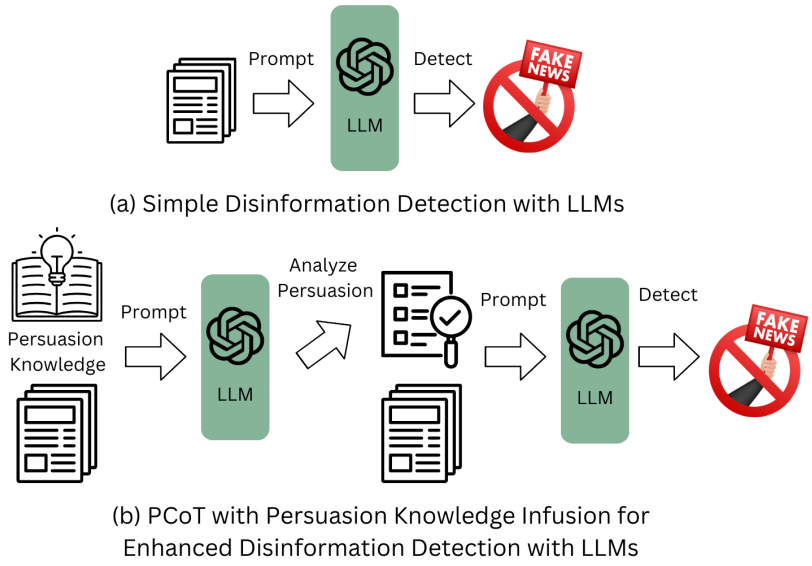

Figure 5.1: The comparison between detecting disinformation with LLMs in a simple zero shot setting and detecting with PCoT and infused knowledge about persuasion.

scarcity challenges compared to supervised models. PCoT operates through a two-stage process where the LLM first identifies and analyzes persuasion within a given text, using infused knowledge. This analysis is then utilized in subsequent reasoning to determine the presence of disinformation. By augmenting models's decision-making process with persusion knowledge, PCoT achieves significant gains in detection performance across multiple datasets.

We conducted experiments on five datasets covering fake news and social media disinformation to evaluate our method rigorously. We evaluated PCoT on two novel datasets, MultiDis and EUDisinfo, which contain up-to-date articles from 2024 onwards, ensuring they were not part of the pretraining data for any tested LLMs. The **Multi**topic **Dis**information is a high-quality dataset developed with fact-checking experts with prior experience in debunking organizations accredited by the International Fact-Checking Network. For a comprehensive evaluation, we also used three publicly available datasets containing texts before the knowledge cutoff of all tested models.

For evaluation, we selected three top-performing methods on zero-shot disinformation detection from Lucas et al. [13] and adapted them to incorporate our PCoT approach. Using five different LLMs, we demonstrate that PCoT delivers significant performance improvements over chosen competitive methods.

5.1 Datasets Used for Experiments

To ensure robust performance of our method across diverse data conditions and inspired by the work of Lucas et al. [13], we designed our evaluation to address potential dataset overlap with LLMs pretraining. We tested our method on two dataset types: (i) prior-cutoff datasets, which may contain pretraining content, and (ii) two novel datasets of articles published after the models' knowledge cutoff. This setup enables a rigorous evaluation of our PCoT method on potential pretraining content and entirely new information. Moreover, we evaluated our method on social media posts versus longer articles, such as news.

5.1.1 Existing Datasets Used for Experiments

The following datasets, published before January 1, 2024, may overlap with the models' training data and are regarded as **Prior-Cutoff Datasets**.

- **CoAID** – A dataset for COVID-19 misinformation detection, comprising 4k+ news articles and 1k+ social posts, all annotated with ground-truth labels [152].
- **ISOT Fake News** – A dataset of 44k+ fake and truthful articles from reputable and unreliable sources, identified via Politifact[1] [153, 154].
- **ECTF** – An extended version of CTF [155] for detecting fake news on Platform X about COVID-19, with additional data to improve early-stage detection [156].

5.1.2 MultiDis Dataset

The **Multi**topic **Dis**information Dataset comprises nearly 2,000 English articles on European and global disinformation. It has been created by researchers from multiple European universities to support disinformation detection research. It is one of the two datasets that in our study is categorized as **Post-Cutoff Dataset**.

Data Sources and Collection. We selected diverse sources to ensure access to both reliable and unreliable content, categorizing each as *Reliable*, *Unreliable*, or *Mixed*. A team of experts evaluated sources through consensus, thoroughly analyzing the source's regularly published content and cross-checking with established tools (e.g., Media Bias/Fact Check). The assigned categories were not revealed to annotators to prevent biases toward sources.

The MultiDis dataset includes a variety of sources: global news agencies, regional publications, thematic platforms, fact-checking organizations, and in-

[1] PolitiFact is a nonprofit fact-checking project by the Poynter Institute.

dependent media. All used 44 distinct sources are freely accessible. To ensure transparency, we make these sources publicly available.

Annotation Methodology and Guidelines. The annotation process involved four key stages:

1. **Methodology and Data Preparation** – Researchers, fact-checking and debunking experts developed a robust methodology and guidelines before collecting a database of articles.
2. **In-Depth Training** – A three-day hybrid training led by the most experienced fact-checking expert aimed to deliver in-depth on-site training to all European teams while ensuring accessibility for remote annotators. Each team was assigned two supervisors, usually a disinformation researcher. The training concluded with an initial annotation round, reviewed and discussed by a fact-checking expert. These preliminary annotations were excluded from the final dataset to maintain high quality.
3. **Article Annotation** – Independent annotation by a less experienced annotator and a supervisor.
4. **Final Evaluation** – The supervisor reviewed both annotations and resolved disagreements through discussion when necessary. A senior fact-checking expert contributed when needed. If consensus was unattainable, the article was labeled *Hard-to-say*.

Appendix A.3 shows details about annotation guidelines used to create MultiDis.

Thematic Category. Before detailed analysis, articles were manually assigned to one of eight thematic categories. The selection of these topics was informed by the EU DisinfoLab report[2] [133]. The categories are: (i) *Anti-Europeanism and Anti-Atlanticism*; (ii) *Anti-migration and Xenophobia*; (iii) *Climate Change and the Energy Crisis*; (iv) *Health*; (v) *Institutional and Media Distrust*; (vi) *Gender Issues*; (vii) *Ukraine War and Refugees*; (viii) *LGBT+*. Table 5.1 shows the distribution of articles by thematic category in the MultiDis dataset. Annotators, during the first credibility evaluation, could label articles as *Inconsistent with the topic*, excluding them from further analysis to ensure high-quality topic assignments.

Credibility Annotation. Annotators assessed each article using a debunking technique, auxiliary complemented by fact-checking, as defined by the NATO Strategic Communications Centre of Excellence [157].

[2]The EU DisinfoLab's report, grounded in expert research from 20 countries across Europe, guarantees high quality and credibility.

Category	*#DOC*	*#PERC*
Anti-Europeanism & Anti-Atlanticism	219	11.4%
Anti-migration and Xenophobia	117	6.1%
Climate Change and the Energy Crisis	324	16.9%
Health	285	14.8%
Institutional and Media Distrust	317	16.5%
Gender Issues	97	5.0%
Ukraine War and Refugees	361	18.8%
LGBT+	202	10.5%

Table 5.1: Number of articles (#*DOC*) per thematic category and their (#*PERC*) percentage in MulitDis dataset.

Given an article, annotators analyze its content to determine whether it belongs to one of four categories. The main categories in our guidelines are: *Credible Information*, *Disinformation*, with the latter following the European Commission's High-Level Expert Group: *Disinformation is false, inaccurate or misleading information designed, presented, and promoted to intentionally cause public harm or for profit* [3]. This definition has also been adopted in other disinformation studies [17, 6]. Two additional labels, *Hard-to-say* and *Inconsistent with the Topic*, were respectively assigned to articles where annotators did not reach a consensus or where the content did not match the assigned topic. Articles labeled with these or published before January 2024 were excluded from experiments.

Annotation and Data Quality Control. Our guidelines require each article to be annotated independently by two experts to minimize bias. Given the time demands of the annotation process, only two independent evaluations per article were guaranteed. Supervisors provided the third final annotation by reviewing the two previous annotations. However, supervisors were instructed to resolve uncertainties through discussions, and the lead fact-checking expert provided clarification when needed. These discussions helped ensure consistency among annotators and reduced human errors and bias. We achieved full agreement in the first two rounds for 86.78% of articles, with the remainder undergoing a more detailed third analysis.

Note: We publicly release the complete dataset, including annotations from all three rounds.

5.1.3 EUDisinfo Dataset

We introduce the EUDisinfo dataset, collected with usage of the EUvsDisinfo database[3], which comprises 18,464 disinformation cases[4]. EUvsDisinfo is an EU initiative dedicated to identifying, analyzing, and countering pro-Kremlin disinformation. Each entry concisely summarizes a disinformation case, along with links to the original misleading content and credible sources debunking the claims. Since EUvsDisinfo provides predefined evaluation for each disinformation case as either *credible* or *disinformation*, we did not conduct additional annotation. The EUvsDisinfo database comprises articles published in multiple languages, some previously analyzed in Leite et al. [158]. However, as all articles in that study date before 2024, this dataset was unsuitable for our research. To address this limitation, we independently curated a collection of approximately 400 English articles published in 2024 or later.

To collect English news article content, we leveraged the *Trafilatura* tool [159], which efficiently scrapes web content while preserving article structure. Additionally, we employed *Selenium* [160] to navigate and extract HTML pages and *Beautiful Soup 4* [161] to parse article content.

5.1.4 MultiDis and EUDisinfo Analysis and Statistics

Table 5.2 presents the percentage of articles per main credibility category in our two datasets. Moreover, Table 5.3 reports the number of articles within the three credibility categories in MultiDis dataset: *Credible Information*, *Disinformation*, and *Hard-to-say*. Additionally, Table 5.3 includes articles labeled as *Inconsistent with the Topic*. Articles in this category were excluded from further analysis. Similar statistics for the EUDisinfo dataset are presented in Table 5.4, which includes only two categories: *Credible Information* and *Disinformation*.

Category	MultiDis	EUDisinfo
Credible Information	65.3%	67.1%
Disinformation	32.8%	32.9%

Table 5.2: Percentage of articles per main credibility category in MultiDis and EUDisinfo datasets.

[3] https://EUvsDisinfo.eu/disinformation-cases/

[4] Database size recorded as of February 11, 2025.

Category	*#DOC*	*#PERC*
Credible Information	1256	65.3%
Disinformation	630	32.8%
Hard-to-say	18	0.95%
Inconsistent with the Topic	18	0.95%

Table 5.3: Number of articles (#*DOC*) per each credibility evaluation category and their (#PERC) percentage in MultiDis dataset.

Category	#DOC	#PERC
Credible Information	241	67.1%
Disinformation	118	32.9%

Table 5.4: Number of articles (#DOC) per main credibility evaluation category and their (#PERC) percentage in the EUDisinfo test dataset.

5.2 Proposed PCoT Method

In this section, we introduce the Persuasion-Augmented Chain of Thought (PCoT) method, which leverages persuasion to enhance zero-shot disinformation detection using generative LLMs.

Empirical studies have shown that persuasion is an integral part of disinformation [12]. Insights from psychological research highlight the potential of leveraging persuasion knowledge to more effectively discern between fake and credible news [151]. Inspired by this, we propose the Persuasion-Augmented Chain of Thought. The PCoT method employs a two-stage reasoning process that improves LLM's disinformation detection by persuasion knowledge infusion.

In the first stage, an LLM is prompted to perform multi-faceted reasoning by analyzing persuasion strategies (see section 2.2.2) within the text. The second stage performs the disinformation detection task, enriched by the previously generated analysis of persuasion strategies. Figure 5.1 presents a simplified comparison between traditional zero-shot disinformation detection using LLMs and our PCoT method. Final prompt templates for each stage of our PCoT method are available in Appendix A.4.

5.2.1 Persuasion Detection Step

In the first stage LLM performs multifaceted reasoning by tackling the multi-class, multi-label task of detecting persuasion strategies, along with contextual question answering by explaining persuasion usage within each text. The per-

suasion detection task can be formally represented as follows: The model M takes as input the text T, the impersonation I_P, the infused knowledge K_P and guidelines G_P. Here, I_P establishes the context and overrides alignment tuning, while K_P encapsulates knowledge about a predefined set of high-level persuasion strategies P, and guidelines G_P that determine the task and specify the structure of the expected response. This combined input is represented as $X = (T, I_P, K_P, G_P)$, where the set of persuasion strategies is given by $P = \{p_1, p_2, \ldots, p_k\}$. For each text, the model generates an output in a structured textual format that can be decoded into a *JSON*-like dictionary. This output contains, for each persuasion strategy $p_i \in P$, two components: a binary label y_{p_i} (*'Yes'* or *'No'*) indicating the presence of p_i in the text, and an explanation E_{p_i} justifying the prediction. The output can be formally expressed as:

$$A_T = \{p_i : (y_{p_i}, E_{p_i}) \mid p_i \in P\}. \tag{5.1}$$

The model M generates the output A_T by leveraging the combined input X, capturing both the text and infused persuasion knowledge:

$$A_T \sim M(T, I_P, K_P, G_P). \tag{5.2}$$

This stage leverages the capabilities of generative LLMs to integrate knowledge about persuasion into the reasoning process. The rationale for our approach is based on the observations that explanations can enhance the robusteness of the final prediction [162], and that previous works have shown that incorporating explanations can improve zero-shot classification performance [163].

5.2.2 Disinformation Detection Step

In the final stage of the PCoT method, LLM performs zero-shot binary classification on each input text. Formally, the model M evaluates the input text T to detect disinformation. It processes the combined input $X = (T, I_D, A_T, G_D)$, where I_D defines the impersonation that establishes the context, A_T provides the persuasion analysis from the first stage of PCoT, and G_D defines the task and specifies the structure of the expected response. The model then generates the output, Y_T indicating whether T contains disinformation (*'Yes'* or *'No'*).

$$Y_T \sim M(T, I_D, A_T, G_D). \tag{5.3}$$

We explored the zero-shot setting as many studies have shown that zero-shot prompting of LLMs like GPT-4 can outperform supervised models like BERT in detecting disinformation [65, 66, 67]. In addition, Lucas et al. [13] demonstrated

that fine-tuning BERT on different datasets and testing on unseen data leads to worse performance than zero-shot with LLMs. We confirm these findings on our data in one of the following sections, specifically section 5.5.1.

5.3 PCoT Design, Experiments and Evaluation

5.3.1 Experimental Setup for PCoT

We created five test sets by randomly selecting texts from each dataset. To evaluate our PCoT method's ability to detect disinformation in data unseen by LLMs, we used two novel test sets, MultiDis and EUDisinfo. These test sets contain only articles published from 2024 onward, ensuring that the content was not part of any LLM training data. Each test set contained 400-500 articles or posts. Table 5.5 presents the class distribution of *Disinformation* and *Credible Information* across the five test datasets. In addition, Table 5.6 shows the same distribution across different content categories and time-based splits, indicating that social media posts and prior cutoff texts contain a higher proportion of disinformation.

Dataset	Disinformation	Credible Information
CoAID	21%	79%
ECTF	41%	59%
EUDisinfo	33%	67%
ISOT Fake News	55%	45%
MultiDis	26%	74%

Table 5.5: Class distribution across evaluation datasets. The proportions reflect the nature of each dataset and its composition regarding disinformation and credible content.

Category	Disinformation	Credible Information
All Texts	35%	65%
Articles	33%	67%
Social Media Posts	41%	59%
Prior Cutoff	39%	61%
Post Cutoff	29%	71%

Table 5.6: Class distribution by text type and time period. Social media and pre-cutoff texts show a higher share of disinformation compared to articles and post-cutoff samples.

We conducted all experiments on five different LLMs: *GPT 4o Mini*, *Llama 3.1 8B*, *Claude 3 Haiku*, *Llama 3.3 70B*, and *Gemini 1.5 Flash*. To ensure the most de-

terministic results possible, we set the hyperparameter temperature to 0 in each model. We aimed to include widely recognized, state-of-the-art models from the largest available while ensuring they remain affordable. We also selected two open-weight models to demonstrate that our method can be applied without access to closed models through APIs. Additionally, we chose the smaller Llama 3.1 with 8B parameters to ensure that our method could be applied to models that do not require costly infrastructure. Table 5.7 lists the Large Language Models used in our experiments, detailing their knowledge cutoff dates, access methods, licenses, and sizes. The knowledge cutoff dates confirm that our datasets, *MultiDis* and *EUDisinfo*, which contain articles from 2024 onward, were not part of the models' pretraining.

API Model Name	Cutoff	Access Details	License	Size
`gpt-4o-mini`	10.2023	OpenAI API 02.2025	COMM	N/A
`gemini-1.5-flash`	11.2023	Google API 02.2025	COMM	N/A
`claude-3-haiku-20240307`	08.2023	Anthropic API 02.2025	COMM	N/A
`meta-llama/Llama-3.3-70B-Instruct-Turbo`	12.2023	DeepInfra API 02.2025	ML3COM	70B
`meta-llama/Meta-Llama-3.1-8B-Instruct`	12.2023	DeepInfra API 02.2025	ML3COM	8B

Table 5.7: Large Language Models used in our experiments. The column *Cutoff* denotes the training-data knowledge cutoff date for each model. Model sizes for commercial models were not disclosed, so we list "N/A". In the *License* column, "COMM" stands for Commercial, and "ML3COM" stands for Meta Llama 3 Community.

PCoT was evaluated using the F_1 score. To assess the significance of its difference from competitive methods, we used McNemar's test, which suits binary tasks comparing two methods on the same dataset [164, 165]. This statistical test has been widely applied in NLP [166, 167].

5.3.2 Persuasion Detection Step

To enhance first stage of the PCoT method we designed prompts that explicitly infuse knowledge about persuasion. This knowledge is grounded in the taxonomy proposed by Piskorski et al. [28, 72], which organizes persuasion techniques into six high-level strategies: *Attack on reputation*, *Justification*, *Simplification*, *Distraction*, *Call*, and *Manipulative wording* (see section 2.2.2 for definitions). The taxonomy was developed by researchers at the Joint Research Centre (JRC) of the European Commission and is accompanied by publicly available, expert-annotated datasets. These resources have been extensively used in prior work, including multiple shared tasks on persuasion detection at the International Workshop on Semantic Evaluation [32, 33]. Leveraging this well-established taxonomy enables a rigorous evaluation of the first stage of PCoT using reliable

ground truth data. Moreover, to our knowledge, it remains the only high-quality persuasion taxonomy that has been systematically applied to longer-form news articles, which constitute a central component of our experimental setup.

Method	F_1 Micro
DMT	↑9% 0.722 ±0.035
DTAT	↑4% 0.689 ±0.042
Base MT	0.664±0.030

Table 5.8: Average F_1 micro ($\pm std$, over five LLMs) for three methods evaluated in the first stage of the PCoT method. Percentage changes are computed relative to the *Base MT* method. The *DMT* variant is selected as the final best-performing method for this stage.

To develop the most effective prompt for detecting persuasive strategies, we conducted extensive experiments on the SemEval 2023 dataset [32], using 536 English news articles with ground truth on persuasion strategies and five LLMs. We used F_1 micro as the evaluation metric for this stage, following its use in a closely related task at SemEval 2023 [32].

We tested various prompts, including:

- Detailed Multitask (DMT) - a single prompt for detecting all strategies and their explanations. Prompt with infused knowledge about persuasion strategies and their definitions (see section 2.2.2,), and the specific techniques with definitions (see Appendix A.5) that fall under each strategy. These techniques are categorized according to the taxonomy proposed by Piskorski et al. [28, 72].
- Detailed One Task At a Time (DTAT) - individual prompts for binary detection and explanations per strategy, infusing the same knowledge as DMT but divided into six parts as there are six persuasion strategies.
- Base Multitask - our baseline single prompt for detecting all strategies. It does not incorporate persuasion knowledge but simply lists strategy names and prompts identification of those present in the text. This served as our starting point.

As shown in Table 5.8, the DMT method achieved the highest F_1 micro score, outperforming our baseline prompt by 9%. In addition, we decided to evaluate futher approaches:

- Multitask (MT) - In this approach, we used a single prompt that included the names and definitions of persuasion strategies, as outlined by Piskorski et al. [72] and showed in section 2.2.2. This zero-shot method guided the LLM in classifying persuasion strategies across multiple labels and categories. Furthermore, we instructed the model to provide explanations for each classification decision.

- One Task at a Time (TAT) - In this approach, we used a separate prompt for each persuasion strategy, treating each as a binary classification task. This approach resulted in six distinct prompts, each focusing on a specific persuasion strategy from Piskorski et al. [72]. Each prompt included only the name and definition of a single strategy, as listed in section 2.2.2. Additionally, we asked the model to explain each classification decision related to the corresponding persuasion strategy.
- One Task at a Time with Broad Knowledge (TATB) - This approach is similar to the TAT method but with a broader scope. Instead of providing knowledge about a single persuasion strategy per prompt, we used six distinct prompts, each containing knowledge about all the persuasion strategies. However, the LLM was still tasked with detecting and analyzing only one specific strategy within each prompt, treating it as a binary classification task.

Table 5.9 presents the average results per each persuasion strategy for presented approaches. It includes the overall average performance in detecting persuasion strategies and the results for each persuasion strategy. The *Detailed Multitask* method outperformed the others in the average performance over all persuasion strategies detecting persuasion. As a result, DMT was used in the first stage of our final PCoT method. Our experiments revealed important finding that: *Using a single prompt to identify all persuasion strategies was more effective than separate prompts for each strategy's binary classification.*

Approach	AR	J	S	D	C	MW	Average
MT	0.6407	0.6616	0.6198	0.7537	0.6366	0.7813	0.6823
	± 0.130	± 0.031	± 0.022	± 0.090	± 0.046	± 0.093	± 0.069
DMT	0.7368	0.6710	0.6290	0.7082	0.6326	0.8440	0.7036
	± 0.103	± 0.044	± 0.028	± 0.104	± 0.046	± 0.058	± 0.081
TAT	0.6522	0.6489	0.5940	0.5153	0.6541	0.7276	0.6320
	± 0.143	± 0.041	± 0.026	± 0.188	± 0.011	± 0.217	± 0.065
DTAT	0.6963	0.6896	0.5985	0.4810	0.6407	0.8045	0.6518
	± 0.086	± 0.013	± 0.028	± 0.147	± 0.023	± 0.100	± 0.109
TATB	0.6455	0.6437	0.5851	0.5269	0.6299	0.6858	0.6195
	± 0.133	± 0.045	± 0.047	± 0.248	± 0.058	± 0.225	± 0.056

Table 5.9: The table presents F_1 scores for each persuasion strategy (shortcuts presented in section 2.2.2) and approach. Standard deviations, calculated from the results across five different LLMs, are provided below their corresponding scores. The final approach used in the PCoT method is the best-performing *DMT*.

The persuasion detection step provides an analysis that includes binary labels and explanations. These explanations further improves reasoning of models in terms of disinformation detection. To assess the impact of these explanations, we

also evaluated PCoT without them. Testing PCoT without explanations showed that including LLM-generated insights improved performance by average of 1.6 percentage points. Table 5.10 presents a comparative analysis of PCoT with and without persuasion strategy explanations across various models.

model	Explanation	F_1 Score
GPT 4o mini	Yes	0.841
	No	0.830
Gemini 1.5 Flash	Yes	0.817
	No	0.798
Claude 3 Haiku	Yes	0.789
	No	0.771
Llama 3.3 70B	Yes	0.844
	No	0.842
Llama 3.1 8B	Yes	0.785
	No	0.756
Average	Yes	0.815
	No	0.799

Table 5.10: Results for PCoT with usage of explanation for each persuasion strategy and without explanation.

The impact of explanations varies, with the most significant improvement observed in the smallest open-weight model, Llama 3.1 8B, while Llama 3.3 70B shows minimal change. We observe a consistent average improvement when using explanations. Since inference is conducted with a temperature of 0, making the results more stable and reproducible, this further reinforces the importance of explanations. Notably, the benefits are most pronounced for smaller models, underscoring the value of explanations in enhancing their disinformation detection performance. As a result of our analysis and performance gains, we decided to include explanations as they further enhance reasoning of models in disinformation detection.

This step of persuasion strategies detection and explanation establishes the foundation for the second stage of PCoT by analyzing the persuasion signals present in the input text.

5.3.3 Disinformation Detection Step

For disinformation detection stage, we selected three top-performing competitive methods based on an extensive evaluation by Lucas et al. [13], specifically those that excelled on human-annotated datasets [152, 60]. We outline the three methods below:

- *VaN* - A vanilla prompt serving as a fundamental baseline, offering concise instructions to LLMs [13].
- *Z-CoT* - Extends *VaN* with a prompt encouraging step-by-step reasoning, inspired by Kojima et al. [168]'s findings on zero-shot reasoning.
- *DeF-SpeC* - Emphasizes contextual, deductive, and abductive reasoning [13], addressing LLM limitations in inductive and multi-step reasoning [66].

The chosen competitive methods served as baselines, allowing us to evaluate the effectiveness of PCoT. We then adapted these methods to our PCoT approach by modifying prompts to incorporate persuasion analysis from the first stage. This approach enabled us to determine whether the PCoT method is sensitive to prompt variations or exhibits consistent behavior. For a rigorous evaluation, we conducted experiments on five datasets covering various themes and genres, such as news and social media posts. The diverse selection of datasets allows us to assess PCoT's generalizability.

To demonstrate the need for two-stage PCoT, we tested a more straightforward single-step approach, where LLMs analyzed persuasion and detected disinformation simultaneously. As shown in Table 5.11 single-step PCoT outperformed the baseline by 8%, while the two-stage method provided an additional significant 7% improvement.

Method	F_1 Score
PCoT	↑15% 0.815 ±0.027
PCoT Single Step	↑8% 0.765 ±0.072
Base	0.711±0.055

Table 5.11: Average F_1 ($\pm std$, over five LLMs) for *PCoT* (two-stage) and *PCoT Single Step*, which uses one prompt for simultaneous persuasion analysis and disinformation detection. Percentage changes are computed relative to the *Base* method.

5.4 Results and Discussion

5.4.1 General Overview

The results of our experiments, presented in Table 5.12, compare the performance of our PCoT method with baseline approaches. PCoT significantly improves performance, achieving an average F_1 score of 0.815, about a 15% improvement over the baselines. To evaluate the statistical significance of PCoT, we conducted McNemar's test comparing each prompting method to its PCoT-adjusted counterpart across various language models. The results, presented in Table 5.13, show that PCoT consistently improves performance at the 0.01 significance level across

	Overall		Articles		Posts		Prior Cutoff		Post Cutoff	
	Base	PCoT	Base	PCoT	Base	PCoT	Base	PCoT	Base	PCoT
GPT 4o Mini										
VaN	0.759	0.845 ↑11%	0.788	0.885 ↑12%	0.700	0.762 ↑9%	0.742	0.830 ↑12%	0.790	0.874 ↑11%
Z-CoT	0.765	0.846 ↑11%	0.801	0.884 ↑10%	0.696	0.767 ↑10%	0.747	0.835 ↑12%	0.801	0.869 ↑8%
DeF-SpeC	0.772	0.834 ↑8%	0.816	0.867 ↑6%	0.690	0.766 ↑11%	0.742	0.813 ↑10%	0.832	0.875 ↑5%
Gemini 1.5 Flash										
VaN	0.681	0.810 ↑19%	0.673	0.843 ↑25%	0.695	0.748 ↑8%	0.683	0.778 ↑14%	0.679	0.875 ↑29%
Z-CoT	0.689	0.808 ↑17%	0.681	0.838 ↑23%	0.703	0.752 ↑7%	0.670	0.777 ↑16%	0.687	0.872 ↑27%
DeF-SpeC	0.744	0.834 ↑12%	0.764	0.876 ↑15%	0.708	0.754 ↑6%	0.721	0.810 ↑12%	0.790	0.884 ↑12%
Claude 3 Haiku										
VaN	0.710	0.797 ↑12%	0.714	0.820 ↑15%	0.702	0.747 ↑6%	0.728	0.797 ↑9%	0.677	0.796 ↑18%
Z-CoT	0.588	0.774 ↑32%	0.601	0.800 ↑33%	0.550	0.716 ↑30%	0.565	0.767 ↑36%	0.626	0.786 ↑26%
DeF-SpeC	0.780	0.795 ↑2%	0.806	0.810 ↑0%	0.727	0.763 ↑5%	0.809	0.812 ↑0%	0.727	0.766 ↑5%
Llama 3.3 70B										
VaN	0.740	0.845 ↑14%	0.747	0.881 ↑18%	0.727	0.768 ↑6%	0.733	0.839 ↑14%	0.752	0.856 ↑14%
Z-CoT	0.722	0.843 ↑17%	0.725	0.878 ↑21%	0.718	0.770 ↑7%	0.707	0.837 ↑18%	0.750	0.855 ↑14%
DeF-SpeC	0.732	0.832 ↑14%	0.740	0.863 ↑17%	0.717	0.768 ↑7%	0.719	0.806 ↑12%	0.755	0.880 ↑17%
Llama 3.1 8B										
VaN	0.627	0.792 ↑26%	0.565	0.802 ↑42%	0.736	0.773 ↑5%	0.649	0.788 ↑21%	0.585	0.801 ↑37%
Z-CoT	0.660	0.791 ↑20%	0.623	0.804 ↑29%	0.725	0.764 ↑5%	0.670	0.789 ↑18%	0.638	0.795 ↑25%
DeF-SpeC	0.697	0.773 ↑11%	0.688	0.784 ↑14%	0.712	0.752 ↑6%	0.683	0.767 ↑12%	0.724	0.785 ↑8%
Average	0.711	0.815 ↑15%	0.715	0.842 ↑18%	0.700	0.758 ↑8%	0.705	0.803 ↑14%	0.721	0.838 ↑16%

Table 5.12: Results with F_1 scores for five LLMs. The *Base* columns shows the competitive method results, while the *PCoT* columns presents results for prompts adapted to the PCoT method. McNemar's test confirmed that, across all models and methods, PCoT achieves significantly better results on *Overall* data at the 0.01 significance level.

all models and methods in overall evaluation. However, certain cases, such as experiments on posts for Llama 3.1 8B and experiments on articles for Claude 3 Haiku for DeF-Spec method, exhibit non-significant differences. McNemar's test confirmed that, almost across all models and methods, PCoT consistently achieves significantly better results on overall data at the 0.01 significance level.

PCoT significantly improves disinformation detection across various scenarios, including news articles, social media posts, and novel post-cutoff datasets. It achieves the most substantial improvement in articles, with a 18% increase. Additionally, PCoT shows a 16% improvement for post-cutoff datasets, leading to our next key finding: *Infusing persuasion knowledge into prompts improves generative LLMs' disinformation detection, especially for long texts and data not seen during pretraining*.

Better performance on unseen data confirms superior effectiveness on longer articles, as these datasets consist exclusively of such texts. We attribute PCoT's improved effectiveness on articles to the greater prevalence of persuasive strategies in longer texts, which complicate disinformation detection even for humans [169], underscoring the need for persuasion knowledge. Furthermore, PCoT

Method	Data	Gemini 1.5 Flash	Claude 3 Haiku	GPT 4o mini	Llama 3.3 70B	Llama 3.1 8B
VaN	overall	0.01	0.01	0.01	0.01	0.01
VaN	articles	0.01	0.01	0.01	0.01	0.01
VaN	posts	0.01	0.01	0.01	0.01	Non-Significant
VaN	prior	0.01	0.01	0.01	0.01	0.01
VaN	post	0.01	0.01	0.01	0.01	0.01
Z-CoT	overall	0.01	0.01	0.01	0.01	0.01
Z-CoT	articles	0.01	0.01	0.01	0.01	0.01
Z-CoT	posts	0.01	0.01	0.01	0.01	Non-Significant
Z-CoT	prior	0.01	0.01	0.01	0.01	0.01
Z-CoT	post	0.01	0.01	0.01	0.01	0.01
DeF-Spec	overall	0.01	0.01	0.01	0.01	0.01
DeF-Spec	articles	0.01	Non-Significant	0.01	0.01	0.01
DeF-Spec	posts	0.01	0.01	0.01	0.01	0.01
DeF-Spec	prior	0.01	Non-Significant	0.01	0.01	0.01
DeF-Spec	post	0.01	0.01	0.01	0.01	0.05

Table 5.13: Results of McNemar's test, comparing each prompting method (*VaN*, *Z-CoT*, and *DeF-Spec*) against its PCoT-adjusted counterpart across various language models. The values represent significance levels for different evaluation metrics, with *Non-Significant* indicating no statistically significant difference at the 0.05 threshold.

deliver the largest average improvement, about 18%, for the smallest model.

5.4.2 Impact of Persuasion

As Figure 5.2 shows, at least one persuasion strategy was found in 92% of disinformation and in 72% of credible texts. These results suggest that persuasion is more commonly used in disinformation than in credible information, though a significant proportion of credible content also contains persuasion. The strongest correlation is observed between disinformation and the prediction of four specific strategies: *Attack on reputation*, *Simplification*, *Distraction*, and *Manipulative wording*. In contrast, the remaining two strategies, namely *Justification* and *Call*, occur with similar frequencies in both disinformation and credible information. The comparable presence of *Call* and *Justification* in both disinformation and credible content may be explained by the broad applicability of the persuasion techniques they encompass. For instance, *Call* techniques like *Slogans* such as "*Make America Great Again!*" are highly persuasive but not inherently misleading, making them familiar across various types of content. Similarly, *Conversation Killers* like "*That's just your opinion*" appear in discussions to shut down debate rather than mislead. Likewise, *Justification* includes techniques often found in credible information. For instance, *Appeal to Authority* is a standard persuasion technique in legitimate discourse, where expert opinions are cited to

Figure 5.2: Average percentage of persuasion strategies predicted across 5 models for disinformation (*DIS*) and reliable information (*REL*). *ALL* represents the percentage of instances with at least one detected persuasion strategy. Other abbreviations are explained in section 2.2.2.

support claims. Similarly, *Appeal to Popularity*, justifying an argument based on widespread acceptance can be found in factual contexts.

In addition to Figure 5.2, we provide a heatmap in Figure 5.3 showing the distribution of predicted persuasion strategies within the final-stage predictions of the PCoT method. Figure 5.3 shows that the LLM-predicted distribution of persuasion strategies for predicted disinformation and reliable information closely matches the results in Figure 5.2.

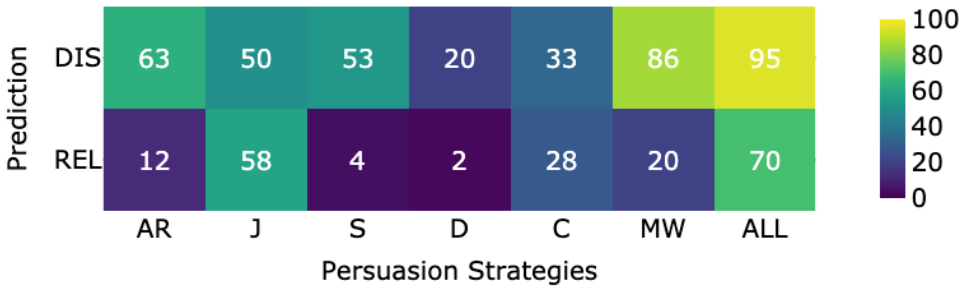

Figure 5.3: Averaged percentage of persuasion strategies predicted across 5 models in predicted disinformation (*DIS*) and predicted reliable information (*REL*). *ALL* represents the percentage of instances with at least one detected persuasion strategy. Other abbreviations are explained in section 2.2.2.

The results presented in Tables 5.14 and 5.15 provide further key insights into the relationship between persuasion strategies and disinformation across different models and prompting methods. Table 5.14 presents the Matthews

correlation coefficient (MCC) between various persuasion strategies and ground truth disinformation labels. The results from Table 5.14 reinforce previous find-

	persuasion	AR	J	S	D	C	MW
GPT 4o mini	0.228	0.528	-0.160	0.611	0.230	0.008	0.507
Gemini 1.5 Flash	0.173	0.476	-0.219	0.511	0.203	-0.000	0.627
Claude 3 Haiku	0.220	0.378	-0.054	0.354	0.201	-0.029	0.628
Llama 3.3 70B	0.328	0.546	0.152	0.536	0.347	0.118	0.591
Llama 3.1 8B	0.178	0.484	-0.054	0.301	0.303	0.064	0.474

Table 5.14: The Matthews correlation coefficient between persuasion strategies and ground truth disinformation label. Table presents coefficients for each persuasion strategy. In addition, *persuasion* column shows correlation with predicted at least one persuasion strategy. Abbreviations of persuasion strategies are explained in section 2.2.2.

ings, showing that across all models, *Attack on Reputation*, *Simplification*, *Distraction*, and *Manipulative Wording* exhibit positive correlations with disinformation, indicating that these strategies are strong signals of misleading content. In contrast, *Justification* and *Call* show in general a negligible correlation, suggesting that these strategies may be equally characteristic of credible and disinformation content. Table 5.15 extends this analysis by evaluating the correlation between persuasion strategies and final disinformation predictions under different PCoT-adapted methods (*VaN*, *Z-CoT*, and *DeF-SpeC*). The results demonstrate consistent patterns across all configurations, suggesting that PCoT's effectiveness is not highly prompt-sensitive and remains stable across different prompting approaches. It is important to note that we could not assess the impact of individual persuasive strategies in complete isolation, as all strategies were detected simultaneously. However, this analysis still provides valuable insight into which persuasive strategies are more characteristic of disinformation versus credible information.

In addition, we evaluated how the PCoT method enhances disinformation detection across two subsets of the datasets used: one in which at least one persuasion strategy was predicted and another in which none was detected. As shown in Table 5.16, PCoT improves detection by an average of 12% in the persuasion-present subset and about 7% in the persuasion-absent subset. While Table 5.16 provides average results for all prompting methods enhanced with PCoT reasoning, the results presented in Tables 5.17, 5.18, and 5.19 underscore the effectiveness of the proposed PCoT approach across various models and all individual prompting methods. The improvement is evident in detecting disinformation in texts with predicted persuasive strategies (*Persuasion* subset) and those without (*No Persuasion* subset). PCoT consistently outperforms the baseline

	persuasion	AR	J	S	D	C	MW
VaN with PCoT							
GPT 4o mini	0.307	0.601	-0.187	0.720	0.279	0.027	0.592
Gemini 1.5 Flash	0.228	0.495	-0.222	0.638	0.268	0.036	0.670
Claude 3 Haiku	0.353	0.491	-0.003	0.466	0.273	0.011	0.788
Llama 3.3 70B	0.422	0.648	0.176	0.622	0.385	0.158	0.700
Llama 3.1 8B	0.151	0.479	-0.155	0.362	0.333	0.027	0.481
Z-CoT with PCoT							
GPT 4o mini	0.308	0.597	-0.183	0.720	0.273	0.023	0.585
Gemini 1.5 Flash	0.227	0.495	-0.212	0.640	0.267	0.034	0.669
Claude 3 Haiku	0.334	0.504	0.018	0.419	0.257	0.012	0.766
Llama 3.3 70B	0.419	0.642	0.184	0.625	0.385	0.154	0.693
Llama 3.1 8B	0.166	0.484	-0.134	0.356	0.334	0.026	0.504
DeF-SpeC with PCoT							
GPT 4o mini	0.276	0.558	-0.203	0.720	0.277	0.025	0.557
Gemini 1.5 Flash	0.250	0.522	-0.225	0.638	0.260	0.032	0.709
Claude 3 Haiku	0.346	0.478	-0.003	0.443	0.255	0.010	0.782
Llama 3.3 70B	0.395	0.613	0.159	0.655	0.408	0.145	0.667
Llama 3.1 8B	0.163	0.455	-0.161	0.373	0.340	0.046	0.477

Table 5.15: The Matthews correlation coefficient between persuasion strategies and the final disinformation prediction. Table shows results for each base prompting method adopted to PCoT usage. Table presents coefficients for each persuasion strategy. In addition, *persuasion* column shows correlation with predicted at least one persuasion strategy. Abbreviations of persuasion strategies are explained in section 2.2.2.

prompting methods (*VaN*, *Z-CoT*, *DeF-SpeC*) in the *Persuasion* subset, where at least one persuasive strategy is identified. While PCoT also shows improvements in the *No Persuasion* subset, the gains are lower, highlighting the challenge of detecting misleading content without persuasive cues. Our findings highlight that: *Detecting disinformation is particularly challenging in texts where no persuasion strategy has been predicted*. Persuasive strategies may introduce emotionally charged language, making deception more apparent when these strategies are analyzed carefully. In contrast, when persuasion is absent, false statements alone may evade detection [6]. In this scenario, fact-checking techniques become more crucial, and semantic analysis of the language alone may be insufficient.

5.5 Further PCoT Evaluation and Ablation Study

We present additional experiments: a comparison between BERT-based model and LLMs with and without persuasion-augmented reasoning (section 5.5.1), a

Model	Persuasion		No Persuasion	
	PCoT	Base	PCoT	Base
GPT 4o Mini	0.872↑ ± 0,006	0.824 ± 0,008	0.342↑ ± 0,025	0.305 ± 0,009
Gemini 1.5 Flash	0.844↑ ± 0,014	0.738 ± 0,036	0.444↑ ± 0,013	0.430 ± 0,007
Claude 3 Haiku	0.831↑ ± 0,014	0.756 ± 0,101	0.177↓ ± 0,043	0.295 ± 0,084
Llama 3.3 70B	0.871↑ ± 0,007	0.781 ± 0,007	0.409↑ ± 0,010	0.343 ± 0,006
Llama 3.1 8B	0.812↑ ± 0,008	0.679 ± 0,050	0.536↑ ± 0,014	0.494 ± 0,059
Average	0.847↑	0.753	0.392↑	0.368

Table 5.16: Results comparison across two subsets: *Persuasion*, containing texts with at least one predicted persuasion strategy, and *No Persuasion* texts with no predicted persuasion. The table reports the average F_1 score and standard deviation for each model across three different prompting methods.

Model	Persuasion		No Persuasion	
	PCoT	Base	PCoT	Base
GPT-4o-mini	0.876	0.815	0.315	0.303
Gemini 1.5 Flash	0.837	0.713	0.438	0.424
Claude 3 Haiku	0.840	0.787	0.128	0.304
Llama 3.3 70B	0.876	0.789	0.407	0.346
Llama 3.1 8B	0.816	0.631	0.551	0.561

Table 5.17: Performance comparison based on F_1 scores across two subsets: *Persuasion*, containing texts with at least one predicted persuasion strategy, and *No Persuasion*, containing texts with no predicted persuasion strategies. The table reports the F_1 score for *VaN* prompting method as *Base* and for our adaptation to *PCoT*.

comparison with other prompting methods (section 5.5.2), a comparison of PCoT against cutting-edge reasoning models (section 5.5.3), and an ablation study to assess the impact of the definitions of the persuasion strategies to the overall performance (section 5.5.4).

5.5.1 Comparing BERT and LLMs on Unseen Data

Experiments in this section aim to validate the findings of Lucas et al. [13], which suggest that LLMs generalize more effectively and outperform BERT models in disinformation detection on unseen datasets. Furthermore, confirming these results strengthens the significance of our persuasion-augmented reasoning approach in advancing zero-shot classification. We also compare the BERT performance on unseen data to LLMs with baseline methods and with

Model	Persuasion		No Persuasion	
	PCoT	Base	PCoT	Base
GPT-4o-mini	0.876	0.827	0.348	0.297
Gemini 1.5 Flash	0.836	0.723	0.434	0.429
Claude 3 Haiku	0.815	0.644	0.196	0.206
Llama 3.3 70B	0.875	0.775	0.404	0.331
Llama 3.1 8B	0.818	0.676	0.535	0.473

Table 5.18: Performance comparison based on F_1 scores across two subsets: *Persuasion*, containing texts with at least one predicted persuasion strategy, and *No Persuasion*, containing texts with no predicted persuasion strategies. The table reports the F_1 score for *Z-CoT* prompting method as *Base* and for our adaptation to *PCoT*.

Model	Persuasion		No Persuasion	
	PCoT	Base	PCoT	Base
GPT-4o-mini	0.865	0.829	0.364	0.315
Gemini 1.5 Flash	0.861	0.779	0.459	0.437
Claude 3 Haiku	0.837	0.838	0.208	0.374
Llama 3.3 70B	0.863	0.780	0.415	0.351
Llama 3.1 8B	0.803	0.730	0.523	0.448

Table 5.19: Performance comparison based on F_1 scores across two subsets: *Persuasion*, containing texts with at least one predicted persuasion strategy, and *No Persuasion*, containing texts with no predicted persuasion strategies. The table reports the F_1 score for *DeF-SpeC* prompting method as *Base* and for our adaptation to *PCoT*.

persuasion-augmented reasoning with PCoT.

5.5.1.1 Experimental Setup

Datasets. We first selected three datasets: (i) CoAID, (ii) ISOT Fake News, and (iii) ECTF, to construct our training and validation sets. The validation set contained around 6,000 texts, while the training set included approximately 40,000. For testing, we used the same subsets as in the primary PCoT evaluation experiments, enabling a direct comparison with zero-shot classification results from baseline methods and our PCoT approach. Furthermore, no articles from EUDisinfo or MultiDis were included in the training or validation sets, ensuring they remained entirely unseen by BERT.

Model and Optimization. We fine-tuned widely used pre-trained BERT model. The Hugging Face model name is as follows: `google-bert/bert-large-uncased`[5].

[5]Hugging Face link to the BERT model and its details: google-bert/bert-large-uncased

This model was also used by Lucas et al. [13]. For our computations, including hyperparameter optimization and final fine-tuning, we utilized an NVIDIA L40 GPU. Since these experiments were not the primary focus of our study, our hyperparameter exploration was limited in scope. However, we systematically varied two key hyperparameters: learning rate and weight decay. Specifically, we experimented with learning rates ranging from 5e-6 to 5e-5 and weight decay values between 0.005 and 0.03. The final selected values were a learning rate 1e-5 and a weight decay of 0.03. Other training hyperparameters were kept constant, including a batch size of 16 for both training and evaluation, three training epochs, and a warm-up phase covering approximately 8% of the total training steps.

5.5.1.2 Results and Discussion.

Tables 5.20 and 5.21 present the results of our experiments comparing the baseline method and the PCoT method across various models with result on BERT model. These tables present performance of each model in detecting disinformation on unseen data, so not available during pretraining and fine-tuning of any of models. BERT performs worse than all other models, with an F_1 score of 0.485.

model	F_1 Score
BERT	0.485
GPT 4o mini	0.808
Gemini 1.5 Flash	0.719
Claude 3 Haiku	0.677
Llama 3.3 70B	0.752
Llama 3.1 8B	0.649

Table 5.20: Comparison between BERT performance and LLMs used with baseline methods (result averaged over 3 base methods) on post-cutoff datasets.

5.5.2 Prompting Methods Comparison

Comparison We compare PCoT with other recent prompting methods, including CoT [24] in a zero-shot version (Z-CoT) [13], Chain-of-Verification (CoVe) [170] and Rephrase and Respond (RaR) [171]. As shown in Table 5.22, PCoT consistently outperforms these methods.

model	F_1 Score
BERT	0.485
GPT 4o mini	0.873
Gemini 1.5 Flash	0.877
Claude 3 Haiku	0.783
Llama 3.3 70B	0.864
Llama 3.1 8B	0.794

Table 5.21: Comparison between BERT performance and LLMs used with PCoT method (result averaged over 3 PCoT runs) on post-cutoff datasets.

Model	**Z-CoT**	**RaR**	**CoVe**	**PCoT**
GPT 4o Mini	0.765	0.698	0.790	**0.846**
Gemini 1.5 Flash	0.689	0.573	0.736	**0.808**
Claude 3 Haiku	0.588	0.768	0.441	**0.774**
Llama 3.3 70B	0.722	0.657	0.835	**0.843**
Llama 3.1 8B	0.660	0.566	0.764	**0.791**

Table 5.22: Overall F_1 scores of different prompting methods on five datasets.

5.5.3 Evaluation Against Reasoning Models

To further evaluate our approach, we compared PCoT-enhanced models to OpenAI's advanced reasoning models, *o1-mini* and *o3-mini*. Specifically, we selected the best-performing (*GPT-4o Mini*) and worst-performing (*Llama 3.1 8B*) models from our zero-shot disinformation detection experiments using PCoT (see Table 5.12) and compared them against the reasoning models.

As shown in Table 5.23, even the weakest model, when used with PCoT, outperforms both *o1-mini* and *o3-mini* in zero-shot disinformation detection. This highlights PCoT's ability to boost reasoning performance, even in smaller models.

Model	**Overall**
GPT 4o Mini + PCoT	**0.846**
Llama 3.1 8B + PCoT	0.791
o3-mini	0.770
o1-mini	0.634

Table 5.23: Overall F_1 scores for PCoT-enhanced models vs. OpenAI reasoning models on five datasets.

5.5.4 PCoT Base Version and Ablation Results

To better understand the contribution of explicit persuasion knowledge in PCoT, we conducted an ablation study using a simplified base version, which provides in the prompt a general definition of persuasion avoiding to mention persuasion strategies.

Remarkably, even without detailed knowledge, this simplified version yields notable performance gains over baseline prompting methods across five datasets (see Table 5.24). Although the original variant of PCoT remains stronger, these findings underscore the role of persuasion-augmented reasoning in zero-shot disinformation detection.

Model	PCoT BV	Base
GPT 4o Mini	$0.814 \uparrow \pm 0{,}007$	$0.765 \pm 0{,}007$
Gemini 1.5 Flash	$0.790 \uparrow \pm 0{,}014$	$0.705 \pm 0{,}034$
Claude 3 Haiku	$0.736 \uparrow \pm 0{,}013$	$0.693 \pm 0{,}097$
Llama 3.3 70B	$0.831 \uparrow \pm 0{,}007$	$0.731 \pm 0{,}009$
Llama 3.1 8B	$0.785 \uparrow \pm 0{,}011$	$0.661 \pm 0{,}035$

Table 5.24: Comparison of average F_1 scores and standard deviations between Base prompts and PCoT without persuasion strategy augmentation. Results are shown for VaN, Z-CoT, and DeF-SpeC (as *Base*), and their adaptations for PCoT's base version (*PCoT BV*).

5.6 Discussion

This chapter investigated whether explicitly incorporating persuasion knowledge into the reasoning process of large language models can improve zero-shot disinformation detection. The proposed Persuasion-Augmented Chain of Thought (PCoT) framework presents that modeling persuasion as an intermediate reasoning step leads to consistent and substantial performance gains across models, datasets, domains, and temporal settings. Rather than treating disinformation detection as a simple binary classification, PCoT operationalizes a more process-oriented view, encouraging models to first reason about how a text attempts to influence the reader.

A key insight emerging from the experiments is that persuasion-augmented reasoning is particularly beneficial for longer-form content, such as news articles. These texts typically employ multiple rhetorical and persuasive strategies that unfold over extended discourse, making them more challenging for both humans and models to assess. The larger improvements observed in articles compared to shorter social media posts suggest that persuasion signals become increasingly

informative as textual complexity increases.

Persuasion strategies such as *Attack on reputation*, *Simplification*, *Distraction*, and *Manipulative wording* were found to be most strongly associated with disinformation. Importantly, persuasion was not exclusive to disinformation, highlighting that persuasion itself is not inherently malicious. Instead, its diagnostic value lies in how certain strategies are combined, emphasized, or exploited within deceptive contexts. This distinction underscores the importance of modeling persuasion in a nuanced manner rather than treating it as a binary indicator.

Another important aspect of the discussion concerns generalization. By evaluating PCoT on datasets published after the knowledge cutoff of the tested models, this work provides strong empirical evidence that persuasion-augmented reasoning remains effective even when factual memorization of LLMs is unlikely to play a role. This suggests that PCoT captures structural properties of disinformation that are more stable over time than topical or entity-specific cues. In this sense, persuasion functions as a component that supports robustness to distributional shifts.

More broadly, the results contribute to an emerging line of research that moves beyond purely label-driven detection toward more explainable approaches. By explicitly exposing intermediate reasoning about persuasion, PCoT provides a foundation for more explainable disinformation detection systems, potentially supporting downstream applications such as analyst-assisted fact-checking or educational tools to improve media literacy.

At the same time, these findings point to several open challenges. The current formulation relies on a fixed persuasion taxonomy and English-language data, and future work should explore how persuasion-augmented reasoning transfers across languages. Moreover, while persuasion-augmented reasoning improves detection performance, texts lacking clear persuasive signals remain difficult to classify, suggesting that persuasion should be viewed as a complementary signal rather than a complete solution.

Overall, this chapter demonstrates that integrating persuasion knowledge into LLM reasoning is a promising direction for advancing disinformation detection. By bridging insights from communication theory and natural language processing, PCoT leads to more robust, generalizable, and explainable detection systems.

6 Intent-Augmented Reasoning for Disinformation Detection

This chapter presents contributions C3.1 - C3.3 presented in Section 1.3.
The chapter is based on the following accepted paper:
Arkadiusz Modzelewski, Witold Sosnowski, Eleni Papadopulos, Elisa Sartori, Tiziano Labruna, Giovanni Da San Martino, and Adam Wierzbicki. 2026. *MALicious INTent Dataset and Inoculating LLMs for Enhanced Disinformation Detection.* In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Rabat, Morocco. Association for Computational Linguistics (accepted and soon to be published).

Researchers in communication theory emphasize that, since disinformation involves a deliberate attempt to deceive and persuade, the intentionality inherent in disinformation definition is critical [4]. Uncovering intentions can help future research detect more effectively goal-driven attempts to influence public beliefs [4]. Although English resources support disinformation research [57, 60, 153], none address the varying types of intent behind malicious agents. To fill this gap, we introduce **MALINT**, the first English corpus that annotates disinformation and the most common **MAL**icious **INT**ention types of disinformation agents. The MALINT dataset is a high-quality resource developed in collaboration with fact-checking experts from organizations accredited by the International Fact-Checking Network. We use MALINT to pursue two core objectives. The first one is (i) *Intent Classification*, and the second objective is (i) *Intent-Augmented Disinformation Detection*.

Intent Classification. We present the first investigation into how well different language models can detect malicious intent in English texts. We evaluate small

and large language models on binary and multilabel classification tasks.

Intent-Augmented Disinformation Detection. Inoculation theory in psychology suggests that exposing individuals to weakened forms of disinformation can build resistance to deception [172, 173]. Building on this idea, we explore whether weakening disinformation via integration of malicious intent knowledge can enhance the LLMs' zero-shot disinformation detection. To evaluate it, we propose intent-based inoculation (IBI) and conduct experiments on five established disinformation datasets that include only disinformation labels, as well as on MALINT. In analysis, we use three data splits:

- a genre-based (articles vs. posts),
- a temporal split separating texts published before and after the LLMs' knowledge cutoff dates,
- a language split employed to assess the usefulness of intent-based reasoning in a multilingual context, covering even low-resource languages such as Estonian and Polish.

We demonstrate that IBI outperforms competitive methods by an average of 9% in English and achieves even larger gains in other languages.

6.1 MALINT Dataset

MALINT is a novel corpus of online news articles designed to advance research on disinformation and the malicious intents behind it. During annotation, annotators first assess each article's credibility. Articles deemed disinformative are then annotated for the underlying malicious intent of the disinformation agents. This approach is grounded in the recognition that disinformation is deliberately crafted to serve specific malicious objectives [4]. Credible content, by its nature, is free of such intent. This perspective draws on the growing consensus in disinformation research that malicious intent is a defining feature of disinformative content [117, 174, 4, 175].

6.1.1 Data Sources and Collection

To build a representative dataset, we collected articles from about 50 online sources spanning mainstream media, outlets promoting alternative or incidental narratives. Sources were reviewed by fact-checking experts and classified by consensus into one of three categories: *Reliable*, *Unreliable*, or *Mixed/Biased*. Classification was based on systematic content review and cross-checking with

fact-checking tools such as Media Bias/Fact Check[1]. Articles were collected from all sources and subsequently provided for disinformation and malicious intent annotation. To prevent annotation bias, source categories were hidden from annotators. We release the full list of sources.

6.1.2 Annotation Methodology and Guidelines

A rigorous, multi-stage annotation approach was used to ensure high-quality and consistent annotations of the dataset.

Guidelines Creation. Our project began with the development of detailed guidelines by a team of experienced disinformation researchers and fact-checking experts, each with 3+ years of expertise in IFCN-accredited organizations. The guidelines specified annotation categories and established rules for ambiguous or complex cases, ensuring a consistent and robust framework for the project (Appendix A.3 presents annotation methodology and guidelines).

Annotator Training and Calibration. To ensure consistent application of guidelines, all annotators participated in a training that combined remote and on-site sessions. The training featured hands-on exercises and calibration annotation rounds, allowing annotators to converge in their understanding of guidelines and receive targeted feedback from the lead fact-checking trainer. Annotations from the calibration phase were used solely for training purposes and were excluded from the final dataset.

Annotation and Review Workflow. After training, annotation followed a structured workflow to ensure quality and reliability:

1. **Independent Annotation**: Each article was independently labeled by a primary annotator and by a supervisor with expertise in disinformation. Discrepancies were resolved through discussion to reach a consensus. The supervisor also conducted a third annotation, with unresolved cases passed to the next stage if necessary.
2. **Resolving Ambiguities**: If the primary annotator and supervisor could not reach consensus, the article could be reviewed with a senior fact-checking expert. If it remained ambiguous after this step, it was labeled as *Hard-to-say*.

[1] An initiative where domain experts perform a careful manual analysis based on clear guidelines [176] Link: https://mediabiasfactcheck.com/

Credibility Annotation. Each article was reviewed using a methodology that incorporated the debunking technique, complemented by fact-checking principles, as outlined by the NATO Strategic Communications Centre of Excellence [157]. Annotators assigned one of three labels. Two main annotations are: *Credible Information* and *Disinformation*. We used the definition of disinformation proposed by the European Commission's High-Level Expert Group [3], widely applied in recent research [4, 17, 6]. Moreover, we introduced an additional annotation label: *Hard-to-say*, for cases where annotators could not agree on veracity. Articles falling into the latter class were excluded from our experiments.

Malicious Intent Annotation. Given that disinformation is deliberately disseminated, we defined five intent categories: *Undermining the Credibility of Public Institutions*, *Changing Political Views*, *Undermining International Organizations and Alliances*, *Promoting Social Stereotypes/Antagonisms*, and *Promoting Anti-scientific Views*. Since annotators could assign any number of these categories to a single article (including none or all) this task constitutes a multilabel annotation problem.

Our categories and intent definition are based on the study proposed by Modzelewski et al. [17] and refined in collaboration with fact-checking experts to better reflect the current disinformation landscape. Figure 6.1 shows malicious intent definition and detailed descriptions for each category.

6.2 Annotation and Data Quality Control

To ensure annotation reliability and reduce bias, each article was independently reviewed by two annotators (a primary annotator and their supervisor). The supervisor performed a third pass, considering independent annotations. In the event of disagreement, supervisors were encouraged to consult with the initial annotators and, as needed, with a senior fact-checking expert.

In the first stage, two annotators achieved an agreement of approximately 85.31% on the credibility task. They reached 65.19% agreement on the more complex multilabel intent task. These figures reflect pre-consensus agreement and indicate how challenging it was to prepare the final consensus annotation rather than measuring the final label quality [72].

In the second stage, the supervisor performed a third annotation. Disagreements were resolved through consensus, and expert input was utilized when necessary. This process raised the final annotation agreement to over 95% for both tasks. This improved the reliability and quality of the dataset. If consensus could not be reached for credibility analysis, the article was assigned a

MALICIOUS INTENT

Malicious intent is a generalization of a disinformation narrative that we can define as a repeating pattern found in several disinformative articles. Malicious intent encapsulates the broader goal of the author, which guides the specific narratives used to achieve that goal [17]. In our study, we categorize the following types of malicious intent:

Undermining the Credibility of Public Institutions [UCPI]

The goal of many disinformation creators is to destroy trust in public institutions. This can be done by undermining official communications, insinuating bad intentions or falsely exposing corruption (e.g., accusing governments of population control with vaccines). The idea is to make citizens disbelieve in the effectiveness of their own state, undermine the sense of its existence or actively fight against it. This is ultimately meant to lead to resentment of the system, thus undermining the very essence of Western democracies. As a result, it becomes easier to spread false information, and the public's resistance to outside influence decreases.

Changing Political Views [CPV]

CPV is aimed at strengthening one side of a political dispute and arousing resentment against the others. It may involve the simultaneous promotion of politicians from extremist movements, which are treated as an alternative to the major parties. It is often based on the portrayal of mainstream politicians as corrupt and evil to the bone (e.g., portraying them as traitors to the nation, dependent on the outside influence of global elites).

Undermining International Organizations and Alliances [UIOA]

UIOA is often part of disinformation activities carried out by external forces. These are aimed at breaking up alliances of democratic countries to facilitate propaganda efforts by authoritarian states (e.g., portraying NATO as an aggressor that will drag peaceful states into war). Numerous extreme political movements also have an interest in shattering trust in international institutions. This is part of a populist influence on society and a way to gain power. International institutions are then most often portrayed as entities that take away the sovereignty of member states (e.g., presenting the EU as an authoritarian organization that imposes its will on others).

Promoting Social Stereotypes/Antagonisms [PSSA]

Deepening social divisions is a frequent goal of disinformation efforts. Divided society is less resistant to manipulation, and mutual distrust also promotes a collapse of confidence in the institution of the state and democracy. This causes internal problems to absorb most of the attention, giving room for external centers of influence to operate. This can take the form of reinforcing xenophobia (e.g., stirring up resentment against Ukrainian refugees and portraying them as dangerous). Aversion to specific social groups can also be exploited (e.g., portraying homosexuals as pedophiles).

Promoting Anti-scientific Views [PASV]

Science is a frequent enemy of disinformation authors. Science enhances critical thinking and is an important part of the strength of Western democracies. Presenting it as an enemy aids in undermining the system under which Western countries operate. Reinforcing anti-scientific attitudes also enables short-term financial gain (e.g., selling pseudo-medical remedies for various diseases). The fight against science can be based on a direct attack on scientists (e.g., the claim that vaccines are designed to depopulate humanity), but is also a significant element of conspiracy theories (e.g., medicine is not used to cure people, but to make money).

Figure 6.1: Definition and categories of malicious intent.

Hard-to-say label and excluded from further steps.
Note: We publish annotations from each stage.

6.3 MALINT Analysis and Statistics

Table 6.1 presents the key statistics for the **MALINT** dataset, comprising news articles. The MALINT corpus consists of 1,600 news articles, with an average length of 963 words (approximately 6,045 characters) per article. This reflects a

collection of relatively long-form texts, suitable for analyzing complex disinformation narratives and underlying malicious intent.

Statistic	Value
Total No. of Articles	1,600
Avg. Article Length (words)	963
Avg. Article Length (characters)	6,045

Table 6.1: Overview of the **MALINT** corpus.

Credibility and Malicious Intent Labels. The dataset includes two primary credibility labels: *Credible*, comprising 63.5% instances, and *Disinformation*, which accounts for the remaining 584 articles (36.5%). Table 6.2 details the distribution of the five malicious intent categories in dataset.

Statistic	UCPI	UIOA	PASV	PSSA	CPV
Count	321	234	154	222	197
%	20.06	14.63	9.63	13.88	12.31

Table 6.2: Malicious intent types distribution in MALINT.

Malicious Intents Multiplicity. Approximately 12.1% of articles are tagged with a single intent, while around 24% contain two or more intent labels. Among these, the most common pattern is the presence of exactly two intents, observed in 15.5% of all articles. The most frequent intent pair is *UIOA* and *UCPI*, co-occurring in 127 articles.

6.4 Intent Classification

To evaluate the ability of LMs to detect malicious intent, we use the MALINT dataset to assess performance across two classification tasks. These tasks are intended to capture different dimensions of intent recognition and provide a broad view of model accuracy when faced with malicious content. We evaluate LMs on the following tasks:

- **Binary Detection Per Class** - Models are evaluated on their ability to detect the presence of a specific malicious intent. Each intent is treated as an independent binary classification problem. This setup allows us to analyze how well models can isolate and identify individual intent categories.

- **Multilabel Detection** - Evaluating a model's ability to identify multiple co-occurring intents as a multilabel classification task, where models must assign all relevant intent labels to each input.

As shown in Table 6.2, our tasks are challenging due to class imbalance across intent categories.

6.4.1 Experimental Setup

For our experiments, the MALINT dataset was split into 770 training, 330 validation, and 500 test instances. Binary classification was evaluated using F_1 over the positive class, while the multilabel task used weighted F_1 to address class imbalance. All metrics were computed on test sets.

Setup for SLMs. We fine-tuned a range of pre-trained Small Language Models (SLMs), selected to represent different architectures and computational requirements. We have chosen *BERT* [177], *RoBERTa* [178], *DeBERTa V3* [179, 180] and *DistilBERT* [181]. DistilBERT was included to assess the model suitable for environments with limited computational resources.

Each model was fine-tuned for the two tasks across 42 model-task combinations, testing multiple hyperparameter settings, totaling around 2,000 experiments. For all models and tasks, we used the Hugging Face Transformers library to load both the tokenizer and the model. The input content was tokenized with truncation and padding enabled, using a maximum sequence length of 256 tokens. The training procedure involved feeding data into a training loop using the `Trainer` API. Hyperparameter tuning was performed over the following grid:

- Learning rate: {1e-5, 2e-5, 3e-5, 4e-5, 5e-5}
- Warmup ratio: {0.06, 0.1}
- Weight decay: {0.01, 0.03, 0.05, 0.1}

The best-performing hyperparameter configuration for each model and each binary detection task is reported in Table 6.3. All models were evaluated using F_1 on the positive class. Table 6.4 presents the corresponding best configurations for the multilabel malicious-intent classification task, where models were evaluated using macro-weighted F_1 to account for class imbalance and the multilabel setting. All experiments were run on an NVIDIA L40 GPU.

Setup for LLMs. We evaluated five cutting-edge LLMs via different APIs: *GPT 4o Mini*, *GPT 4.1 Mini*, *Gemini 2.0 Flash*, *Gemma 3 27b it*, *Llama 3.3 70B*. We aimed to include widely recognized, state-of-the-art models from the largest available while ensuring they remain affordable. We also selected two open-weight models to demonstrate that intent-based reasoning can be applied without access to

Model	Identifier	Learning Rate	Weight Decay	Warmup Ratio
Undermining the Credibility of Public Institutions **(UCPI)**				
BERT-base	google-bert/bert-base-uncased	2e-5	0.03	0.06
BERT-large	google-bert/bert-large-uncased	1e-5	0.05	0.06
RoBERTa-large	FacebookAI/roberta-large	2e-5	0.05	0.06
RoBERTa-base	FacebookAI/roberta-base	1e-5	0.05	0.1
DeBERTa-v3-large	microsoft/deberta-v3-large	1e-5	0.01	0.06
DeBERTa-v3-base	microsoft/deberta-v3-base	2e-5	0.05	0.06
DistilBERT-base	distilbert/distilbert-base-uncased	1e-5	0.03	0.06
Changing Political Views **(CPV)**				
BERT-base	google-bert/bert-base-uncased	2e-5	0.01	0.1
BERT-large	google-bert/bert-large-uncased	2e-5	0.01	0.06
RoBERTa-large	FacebookAI/roberta-large	2e-5	0.03	0.1
RoBERTa-base	FacebookAI/roberta-base	2e-5	0.1	0.06
DeBERTa-v3-large	microsoft/deberta-v3-large	1e-5	0.05	0.1
DeBERTa-v3-base	microsoft/deberta-v3-base	1e-5	0.03	0.06
DistilBERT-base	distilbert/distilbert-base-uncased	2e-5	0.01	0.06
Undermining International Organizations and Alliances **(UIOA)**				
BERT-base	google-bert/bert-base-uncased	5e-5	0.03	0.1
BERT-large	google-bert/bert-large-uncased	1e-5	0.05	0.1
RoBERTa-large	FacebookAI/roberta-large	1e-5	0.05	0.06
RoBERTa-base	FacebookAI/roberta-base	2e-5	0.05	0.06
DeBERTa-v3-large	microsoft/deberta-v3-large	1e-5	0.03	0.06
DeBERTa-v3-base	microsoft/deberta-v3-base	3e-5	0.05	0.06
DistilBERT-base	distilbert/distilbert-base-uncased	1e-5	0.1	0.06
Promoting Social Stereotypes/Antagonisms **(PSSA)**				
BERT-base	google-bert/bert-base-uncased	2e-5	0.01	0.06
BERT-large	google-bert/bert-large-uncased	2e-5	0.03	0.06
RoBERTa-large	FacebookAI/roberta-large	1e-5	0.01	0.06
RoBERTa-base	FacebookAI/roberta-base	2e-5	0.01	0.06
DeBERTa-v3-large	microsoft/deberta-v3-large	1e-5	0.05	0.1
DeBERTa-v3-base	microsoft/deberta-v3-base	1e-5	0.03	0.1
DistilBERT-base	distilbert/distilbert-base-uncased	1e-5	0.1	0.1
Promoting Anti-scientific Views **(PASV)**				
BERT-base	google-bert/bert-base-uncased	1e-5	0.05	0.1
BERT-large	google-bert/bert-large-uncased	2e-5	0.01	0.06
RoBERTa-large	FacebookAI/roberta-large	1e-5	0.03	0.1
RoBERTa-base	FacebookAI/roberta-base	1e-5	0.01	0.1
DeBERTa-v3-large	microsoft/deberta-v3-large	1e-5	0.03	0.06
DeBERTa-v3-base	microsoft/deberta-v3-base	1e-5	0.05	0.06
DistilBERT-base	distilbert/distilbert-base-uncased	1e-5	0.03	0.1

Table 6.3: Optimal hyperparameters for binary classification of all malicious intent categories. Identifiers of all models given as of 21.07.2025.

closed models through APIs. To ensure as deterministic results as possible, we prompted all models with the temperature parameter set to zero. All evaluations were conducted in a zero-shot setting, as many documents were too long for few-shot prompting within the LLM context limits. The full set of experiments

Model	Identifier	Learning Rate	Weight Decay	Warmup Ratio
BERT-base	google-bert/bert-base-uncased	2e-5	0.1	0.06
BERT-large	google-bert/bert-large-uncased	2e-5	0.01	0.1
RoBERTa-large	FacebookAI/roberta-large	1e-5	0.03	0.06
RoBERTa-base	FacebookAI/roberta-base	1e-5	0.05	0.1
DeBERTa-v3-large	microsoft/deberta-v3-large	1e-5	0.03	0.1
DeBERTa-v3-base	microsoft/deberta-v3-base	4e-5	0.05	0.1
DistilBERT-base	distilbert/distilbert-base-uncased	2e-5	0.03	0.1

Table 6.4: Optimal hyperparameters for each model in multilabel malicious intent classification. Identifiers of all models as of 21.07.2025.

involved approximately 15,000 API calls.

Table 6.5 lists the Large Language Models used in our experiments, detailing their knowledge cutoff dates, access methods, licenses, and sizes. To enable evaluation on both prior and post cutoff content, we used a knowledge cutoff date of September 2024 to split the EUDisinfo and MALINT datasets accordingly. All other datasets contain only texts published prior to this date. Prompts used for all tasks are provided in Appendix A.6.

API Model Name	Cutoff	Access Details	License	Size
`gpt-4o-mini`	10.2023	OpenAI API 07.2025	COMM	N/A
`gemini-2.0-flash`	06.2024	Google API 07.2025	COMM	N/A
`gpt-4.1-mini-2025-04-14`	06.2024	OpenAI API 07.2025	COMM	N/A
`meta-llama/Llama-3.3-70B-Instruct`	12.2023	DeepInfra API 07.2025	ML3COM	70B
`google/gemma-3-27b-it`	08.2024	DeepInfra API 07.2025	GTU	27B

Table 6.5: Large Language Models used in our experiments. The column *Cutoff* denotes the training-data knowledge cutoff date for each model. Model sizes for commercial models were not disclosed, so we list "N/A". In the *License* column, "COMM" stands for Commercial, and "ML3COM" stands for Meta Llama 3 Community, and "GTU" stands for Gemma Terms of Use.

Baselines. We implemented two baselines for all tasks: a random classifier and logistic regression. For binary classification, the logistic regression model was trained on bag-of-words features represented as sparse token count vectors, with English stop words removed. For the multilabel task, logistic regression was applied using a one-vs-rest strategy [182].

6.4.2 Evaluation Results

Binary Detection Per Each Class. As shown in Table 6.6, DeBERTa V3 Large and RoBERTa models consistently achieved the highest F_1 scores across most

intent categories among SLMs. Among LLMs, GPT 4.1 Mini performed best for *UCPI* and *PSSA*, while Llama 3.3 70B for *PASV* and *CPV* categories. LLMs achieved superior results compared to fine-tuned SLMs across three intent categories. The logistic regression baseline with bag-of-words outperformed random predictions but remained below most LMs, serving as a simple baseline.

Model	UCPI	UIOA	PASV	PSSA	CPV
Small Language Models					
BERT Base	0.562	0.484	0.500	**0.614**	0.293
BERT Large	0.528	0.437	0.543	0.529	0.306
DeBERTa V3 Base	0.675	0.505	0.580	0.523	0.400
DeBERTa V3 Large	**0.696**	**0.649**	**0.683**	0.547	0.460
RoBERTa Base	0.693	0.547	0.674	0.515	**0.486**
RoBERTa Large	0.682	0.630	0.680	0.505	0.444
DistilBERT Base	0.599	0.547	0.564	0.450	0.400
Large Language Models					
GPT 4o Mini	0.543	0.547	0.632	0.458	0.324
GPT 4.1 Mini	**0.702**	0.469	0.717	**0.479**	0.371
Gemini 2.0 Flash	0.639	**0.604**	0.722	0.452	0.444
Llama 3.3 70B	0.569	0.427	**0.738**	0.415	**0.496**
Gemma 3 27B it	0.682	0.395	0.667	0.424	0.407
Baselines					
Random	0.279	0.205	0.122	0.179	0.162
LR with BoW	0.581	0.477	0.595	0.424	0.376

Table 6.6: LMs' F_1 scores on binary intent classification across five categories, compared to random and BoW logistic regression baselines.

Multilabel Detection. We show results for this task in Table 6.7. DeBERTa V3 and RoBERTa continued to perform best among the SLMs, achieving the highest weighted F_1 scores. The best-performing LLM, LlaMA 3.3 70B, lagged noticeably behind the top SLMs. Surprisingly, the logistic regression baseline (using a one-vs-rest strategy) outperformed most LLMs. However, fine-tuned SLMs demonstrated superior ability to capture the complexity of co-occurring intent labels, underscoring the effectiveness of supervision.

6.5 Intent-Augmented Disinformation Detection

Inoculation theory, introduced by McGuire [183], uses a biological metaphor. It suggests that, just as people can be protected against viruses through vaccines, they can also be "vaccinated" to resist persuasive messages [183]. An inoculation

Model	Micro F_1	Weighted F_1
Small Language Models		
BERT Base	0.421	0.414
BERT Large	0.578	0.521
DeBERTa V3 Base	0.812	0.804
DeBERTa V3 Large	**0.817**	0.815
RoBERTa Base	0.813	**0.821**
RoBERTa Large	0.775	0.808
DistilBERT Base	0.759	0.769
Large Language Models		
GPT 4o Mini	0.446	0.457
GPT 4.1 Mini	0.489	0.498
Gemini 2.0 Flash	0.410	0.404
Llama 3.3 70B	**0.542**	**0.570**
Gemma 3 27B it	0.440	0.485
Baselines		
Random	0.192	0.201
LR with BoW (OvR)	0.503	0.491

Table 6.7: Performance of LMs on multilabel intent classification compared to baselines: a random classifier and a one-vs-rest logistic regression using BoW approach.

message has two parts: a *threat* and *refutational preemption*. The threat alerts individuals that a persuasive attack is coming [184]. Refutational preemption (or prebunking) involves providing people with arguments or tools to resist persuasive attacks, helping them better recognize and respond to such attempts [185]. Building on this theory and its applicability to improve disinformation detection [172], we pose the following research question: *Does inoculating LLMs against malicious intent improve their disinformation detection performance in a zero-shot setting?*

To answer this question, we designed an intent-based inoculation (an IBI, we call it also intent-augmented reasoning) experiment in which the threat is an information that the text might hide malicious intent. Refutational preemption in the IBI consists of the LLM-generated analysis of intent. This analysis is generated by utilizing knowledge about types of malicious intent from our taxonomy.

6.5.1 Existing Datasets Used in Experiments

We rigorously evaluate intent-augmented reasoning on the MALINT dataset and 5 additional datasets covering diverse topics, text genres and languages:

- **ISOT Fake News**: Thousands of real/fake news articles from reputable

sources and sites flagged by PolitiFact[2] [153, 154].
- **CoAID**: COVID-19 misinformation dataset with news and social media posts [152].
- **EUDisinfo**: The latest English disinformation dataset collected from the EUvsDisinfo database[3] [18].
- **ECTF**: COVID-19 fake post detection dataset from Platform X (Twitter) [156].
- **EUvsDisinfo**: Multilingual EUvsDisinfo's texts with pro-Kremlin propaganda [158].

We used five datasets (including MALINT) to assess the usefulness of intent-based reasoning across genres and temporal splits in English. To evaluate its cross-lingual generalizability, we also used the EUvsDisinfo texts, splitting it into six language-specific datasets: German, French, Polish, Estonian, Russian, and Spanish.

6.5.2 Intent-based Inoculation Design

As a first step in the IBI framework, the model M generates a structured intent analysis of the input text T. This is a multilabel task over a predefined taxonomy of malicious intents $I = \{i_1, i_2, \ldots, i_m\}$, accompanied by natural language explanations. To facilitate this, the model receives the text T, external knowledge K_I describing common types of malicious intent, and G_A that provides task guidance and desired output structure. We define the intent analysis prompt as:

$$X_T = (T, K_I, G_A) \tag{6.1}$$

The model M produces a structured output:

$$A_I(T) = \{i_j : (r_{i_j}, R_{i_j}) \mid i_j \in I\}, \tag{6.2}$$

where each $r_{i_j} \in \{\texttt{Yes}, \texttt{No}\}$ is a binary label indicating whether intent i_j is present in the text, and R_{i_j} is the accompanying rationale.

Formally, we define:

$$A_I(T) \sim M(X_T) = M(T, K_I, G_A). \tag{6.3}$$

To test if intent-inoculated LLMs improve disinformation detection, we design an inoculation prompt with a threat and a refutational preemption:

[2]PolitiFact is a nonprofit fact-checking organization.

[3]The EUvsDisinfo comprises 19,455 disinformation cases (number as of October 2, 2025). Link: https://EUvsDisinfo.eu/disinformation-cases/

- The **threat** θ is a textual warning that the input text may contain hidden malicious intent.
- The **refutational preemption** is constructed from the previously generated analysis $A_I(T)$.

These elements are combined with the original text T, and detection-specific task guidelines G_I. The full IBI input is then:

$$Z_T = (T, \theta, A_I(T), G_I). \tag{6.4}$$

The model M uses this input to binary detection, indicating whether T is considered disinformative:

$$\hat{y}_T \sim M(Z_T) = M(T, \theta, A_I(T), G_I). \tag{6.5}$$

This design of experiments allows us to answer our research question. By adding a threat and LLM-generated refutational preemption, IBI applies inoculation theory to enhance disinformation detection.

6.5.3 Experimental Setup

We created five test sets by randomly sampling about 400-500 texts from each of the datasets. Moreover, we sampled approximately 3,000 texts from EUvsDisinfo, creating test sets of about 500 per language across six languages. Table 6.8 reports class proportions across all test sets.

Dataset	Disinformation	Credible
MALINT	30%	70%
ISOT Fake News	55%	45%
CoAID	21%	79%
EUDisinfo	33%	67%
ECTF	41%	59%
EUvsDisinfo	49%	51%

Table 6.8: Class distribution across test datasets.

Our experiments were conducted on the same five LLMs that we used for experiments in section 6.4. In these experiments, we again set the temperature hyperparameter to zero for all models. We focused on zero-shot settings with LLMs to evaluate the effectiveness of the IBI when texts lack explicit information about malicious intent. Additionally, prior studies show that LLMs can outperform supervised models such as BERT in disinformation detection [65, 66]. Lucas

et al. [13] also found that fine-tuned BERT models perform worse on unseen data compared to zero-shot with LLMs.

To thoroughly evaluate IBI, we used two data splits on English data: (a) a genre-based split, separating long-form news articles from social media posts (from Platform X, formerly Twitter), and (b) a temporal split, comparing texts published before and after the LLMs' knowledge cutoffs. The temporal split is possible because EUDisinfo and MALINT include post-cutoff articles not seen during model training. Following Lucas et al. [13], this setup enables a rigorous evaluation of IBI across two genres and, crucially, on unseen data, providing a more realistic test of its generalization. IBI was further evaluated on six languages to highlight its cross-lingual generalization.

For the disinformation detection stage, we selected three strong baseline methods identified by Lucas et al. [13] as top performers, particularly on human-annotated datasets such as CoAID [152] and FakeNewsNet [60]. The selected methods are as follows:

- *VaN* – A basic prompt that provides minimal, direct instructions to the LLM, serving as a foundational baseline [13].
- *Z-CoT* – Builds on *VaN* by encouraging step-by-step reasoning, following the zero-shot chain-of-thought prompting strategy introduced by Kojima et al. [168].
- *DeF-SpeC* – A more advanced prompt designed to elicit contextual, deductive, and abductive reasoning [13], addressing limitations in LLMs' ability to perform multi-step or inductive inference [66].

To adapt above methods to our setting, we modified the original prompts to incorporate malicious intent analysis, as introduced in the first stage of our pipeline. This allowed us to examine whether IBI is robust across different prompting strategies or sensitive to prompt. All prompts templates are available in Appendix A.6.3.

We use F_1 for the positive class as evaluation metric. To assess significance between IBI and baselines, we used McNemar's test, a standard method for comparing two models on the same binary task [165, 164], widely used in NLP [167, 166].

6.5.4 Results

Results for MALINT and Other Datasets. Table 6.9 compares baseline prompting strategies (Base) with their intent-based inoculation (IBI) counterparts on the MALINT dataset. Across all models, IBI consistently improves disinformation detection, with average gains ranging from around 2% for GPT-4o Mini up to 8% for Gemini 2.0 Flash.

	GPT 4o Mini		*GPT 4.1 Mini*		*Gemini 2.0 Flash*		*Gemma 3 27b it*		*Llama 3.3 70B*	
	Base	IBI	Base	IBI	Base	IBI	Base	IBI	Base	IBI
VaN	0.815	0.856 ↑5%	0.856	0.825	0.873 ↑6%	0.789	0.855 ↑8%	0.783	0.820 ↑5%	0.836	0.863 ↑3%
Z-CoT	0.836	0.849 ↑2%	0.810	0.861 ↑6%	0.751	0.837 ↑11%	0.782	0.806 ↑3%	0.807	0.865 ↑7%	
DeF_Spec	0.887	0.877 ↓1%	0.870	0.879 ↑1%	0.843	0.881 ↑5%	0.812	0.846 ↑4%	0.806	0.871 ↑8%	

Table 6.9: F_1 scores on MALINT for competitive prompting methods and their improvement with IBI.

Table 6.10 reports F_1 scores across four remaining disinformation datasets used in experiments. Results are presented for various models and prompting strategies, further enhanced with intent-based inoculation. The table compares the Base setup with IBI, showing that IBI consistently improves performance across nearly all models and datasets. The largest gains are observed on datasets containing longer-form news articles, while improvements on ECTF are more modest, likely due to limited context in shorter social media posts. Overall, these results demonstrate that intent-augmented reasoning effectively enhances disinformation detection, particularly for longer-form texts, across diverse datasets and model architectures.

	GPT 4o Mini		*GPT 4.1 Mini*		*Gemini 2.0 Flash*		*Gemma 3 27b it*		*Llama 3.3 70B*	
	Base	IBI	Base	IBI	Base	IBI	Base	IBI	Base	IBI
CoAID										
VaN	0.531	0.627	0.480	0.599	0.631	0.607	0.618	0.650	0.654	0.680
Z-CoT	0.532	0.628	0.468	0.611	0.588	0.603	0.388	0.660	0.628	0.699
DeF_Spec	0.507	0.566	0.496	0.624	0.574	0.610	0.617	0.651	0.582	0.646
ECTF										
VaN	0.812	0.821	0.772	0.758	0.743	0.713	0.769	0.733	0.799	0.732
Z-CoT	0.796	0.830	0.766	0.762	0.728	0.703	0.589	0.728	0.789	0.745
DeF_Spec	0.783	0.837	0.790	0.784	0.801	0.812	0.800	0.755	0.773	0.736
ISOTFakeNews										
VaN	0.776	0.889	0.681	0.665	0.650	0.761	0.529	0.762	0.701	0.720
Z-CoT	0.761	0.890	0.620	0.662	0.591	0.678	0.514	0.751	0.683	0.726
DeF_Spec	0.741	0.782	0.780	0.744	0.780	0.832	0.634	0.790	0.623	0.686
EUDisinfo										
VaN	0.682	0.860	0.678	0.856	0.693	0.842	0.821	0.873	0.784	0.866
Z-CoT	0.727	0.847	0.653	0.845	0.705	0.846	0.802	0.877	0.765	0.885
DeF_Spec	0.789	0.849	0.755	0.846	0.794	0.829	0.867	0.885	0.803	0.886

Table 6.10: F_1 scores on four disinformation datasets for competitive prompting methods and their enhancement with Intent-Based Inoculation.

To evaluate the statistical significance of intent-base inoculation, we conducted McNemar's test comparing each prompting method to its IBI-adjusted counterpart across various language models. The results, presented in Table 6.11, show

that IBI improves performance primarily at the significance level of 0.01 across all models and methods in the overall evaluation. However, certain cases, such as experiments on twitter posts exhibit significance for each adjusted prompting method only on GPT-4.1 Mini model. Overall, siginificance on 0.05 level or higher can be observed on about 79% different scenarios.

Method	Split	Gemini 2.0	GPT4.1 m	GPT4o m	Llama 3.3	Gemma 3
VaN	Overall	0.010	0.010	0.010	0.050	0.010
VaN	News	0.010	0.010	0.010	0.010	0.010
VaN	Post	N.S.	0.010	0.010	N.S.	N.S.
VaN	Prior Cutoff	0.010	0.010	0.010	N.S.	0.010
VaN	Post Cutoff	0.010	0.010	0.010	0.050	N.S.
Z-CoT	Overall	0.010	0.010	0.010	0.010	0.010
Z-CoT	News	0.010	0.010	0.010	0.010	0.010
Z-CoT	Post	N.S.	0.010	0.010	N.S.	0.010
Z-CoT	Prior Cutoff	0.010	0.010	0.010	0.010	0.010
Z-CoT	Post Cutoff	0.010	0.010	N.S.	0.010	N.S.
DeF_Spec	Overall	0.010	0.010	0.010	0.010	0.010
DeF_Spec	News	0.010	0.010	0.050	0.010	0.010
DeF_Spec	Post	N.S.	0.010	0.010	N.S.	N.S.
DeF_Spec	Prior Cutoff	0.010	0.010	0.010	0.010	0.010
DeF_Spec	Post Cutoff	N.S.	N.S.	N.S.	0.010	N.S.

Table 6.11: Results of McNemar's test, comparing each prompting method (*VaN*, *Z-CoT*, and *DeF-Spec*) against its IBI-adjusted counterpart across various language models. The values represent significance levels for different evaluation metrics, with *N.S* as *Non-Significant* indicating no statistically significant difference at the 0.05 threshold. Models used: *GPT 4o Mini*, *GPT 4.1 Mini*, *Gemini 2.0 Flash*, *Gemma 3 27b it*, *Llama 3.3 70B*.

We examined how the correctness of intent predictions influences disinformation detection by computing F_1 scores separately for instances where each intent was predicted correctly versus incorrectly. Across models and intent types, correct intent predictions generally lead to higher F_1 scores, indicating that accurate intent understanding strengthens disinformation detection under IBI. However, for certain intents such as UIOA and specific models, higher performance is occasionally observed even when the focal intent is mispredicted, suggesting that correctly identified co-occurring intents can partially compensate.

Table 6.12 shows the average F_1 scores for disinformation detection, split by whether the predicted intent matches the gold labels ("Correct") or not ("Incorrect"). While the overall trend confirms the positive contribution of accurate intent prediction, the multi-label nature of intent annotation makes this relationship more nuanced: errors in one intent dimension may be offset by correct predictions in others. Together, these findings show that intent-augmented reasoning enhances disinformation detection, but its effectiveness depends on interactions among multiple intent signals rather than on any single intent in

isolation.

Table 6.12 presents average results across all three competitive baseline methods improved to intent-based resoning. In contrast, Tables 6.13, 6.14, and 6.15 report results for each specific baseline method further enhanced with intent-based reasoning.

Model	CPV		PSSA		UIOA		PASV		UCPI	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
GPT4o m	0.914	0.787	0.882	0.809	0.860	0.864	0.866	0.850	0.905	0.796
Llama 3.3	0.889	0.825	0.876	0.839	0.856	0.899	0.862	0.888	0.836	0.941
GPT4.1 m	0.872	0.869	0.886	0.831	0.862	0.908	0.862	0.905	0.876	0.863
Gemini 2.0	0.878	0.821	0.855	0.866	0.845	0.902	0.854	0.871	0.865	0.842
Gemma 3	0.853	0.806	0.834	0.799	0.824	0.824	0.815	0.862	0.876	0.722

Table 6.12: Average F1 scores for disinformation detection across three methods (VaN, Z-CoT, DeF-Spec), split by correct and incorrect intent prediction for each intent. Models used: *GPT 4o Mini*, *GPT 4.1 Mini*, *Gemini 2.0 Flash*, *Gemma 3 27b it*, *Llama 3.3 70B*.

Model	CPV		PSSA		UIOA		PASV		UCPI	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
GPT4o m	0.910	0.783	0.883	0.792	0.852	0.867	0.860	0.848	0.903	0.788
Llama 3.3	0.881	0.829	0.870	0.843	0.855	0.889	0.857	0.889	0.834	0.933
GPT4.1 m	0.870	0.879	0.882	0.850	0.864	0.912	0.861	0.921	0.871	0.877
Gemini 2.0	0.869	0.830	0.849	0.871	0.843	0.896	0.856	0.853	0.871	0.821
Gemma 3	0.853	0.800	0.831	0.792	0.828	0.800	0.812	0.857	0.879	0.707

Table 6.13: F_1 scores with VaN adjusted with IBI, split by intent prediction correctness. Models used: *GPT 4o Mini*, *GPT 4.1 Mini*, *Gemini 2.0 Flash*, *Gemma 3 27b it*, *Llama 3.3 70B*.

Model	CPV		PSSA		UIOA		PASV		UCPI	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
GPT4o m	0.911	0.766	0.872	0.796	0.850	0.847	0.857	0.832	0.913	0.759
Llama 3.3	0.891	0.818	0.873	0.843	0.854	0.904	0.860	0.889	0.833	0.945
GPT4.1 m	0.862	0.860	0.874	0.825	0.844	0.931	0.854	0.889	0.865	0.855
Gemini 2.0	0.847	0.819	0.833	0.847	0.825	0.879	0.828	0.866	0.840	0.830
Gemma 3	0.837	0.786	0.818	0.774	0.805	0.809	0.794	0.857	0.858	0.707

Table 6.14: F_1 scores with Z-CoT adjusted with IBI, split by intent prediction correctness. Models used: *GPT 4o Mini*, *GPT 4.1 Mini*, *Gemini 2.0 Flash*, *Gemma 3 27b it*, *Llama 3.3 70B*.

Genre and Temporal Split Results Table 6.16 compares the baseline prompting strategies (Base) with their IBI counterparts. Across 75 evaluation scenarios (5 models × 3 prompting strategies × 5 settings: overall comparison, articles, social media posts, prior-cutoff, and post-cutoff), IBI leads to improved performance

	CPV		PSSA		UIOA		PASV		UCPI	
Model	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
GPT4o m	0.921	0.812	0.892	0.839	0.876	0.878	0.880	0.869	0.899	0.841
Llama 3.3	0.894	0.829	0.886	0.829	0.861	0.904	0.868	0.885	0.840	0.945
GPT4.1 m	0.885	0.867	0.901	0.819	0.878	0.881	0.871	0.906	0.892	0.857
Gemini 2.0	0.918	0.814	0.882	0.879	0.866	0.932	0.877	0.896	0.885	0.874
Gemma 3	0.868	0.832	0.852	0.830	0.840	0.864	0.840	0.871	0.892	0.752

Table 6.15: F_1 scores with DeF-Spec adjusted with IBI, split by intent prediction correctness. Models used: *GPT 4o Mini*, *GPT 4.1 Mini*, *Gemini 2.0 Flash*, *Gemma 3 27b it*, *Llama 3.3 70B*.

in approximately 90% of cases. On average, the overall performance increases by 9%. McNemar's test indicates that, in nearly all scenarios, IBI significantly outperforms the baselines on the overall dataset at the 0.01 significance level. For Llama 3.3 70B, the difference is still significant, though at the 0.05 level.

	Overall		**Articles**		**Posts**		**Prior Cutoff**		**Post Cutoff**	
	Base	IBI	Base	IBI	Base	IBI	Base	IBI	Base	IBI
GPT 4o Mini										
VaN	0.736	0.828 ↑13%	0.754	0.862 ↑14%	0.703	0.755 ↑7%	0.727	0.821 ↑13%	0.762	0.846 ↑11%
Z-CoT	0.740	0.826 ↑12%	0.764	0.854 ↑12%	0.692	0.766 ↑11%	0.724	0.823 ↑14%	0.786	0.833 ↑6%
DeF-SpeC	0.746	0.792 ↑6%	0.782	0.817 ↑4%	0.682	0.742 ↑9%	0.712	0.771 ↑8%	0.843	0.850 ↑1%
GPT 4.1 Mini										
VaN	0.698	0.751 ↑8%	0.718	0.772 ↑8%	0.659	0.705 ↑7%	0.672	0.709 ↑6%	0.767	0.862 ↑12%
Z-CoT	0.673	0.748 ↑11%	0.685	0.765 ↑12%	0.649	0.712 ↑10%	0.640	0.710 ↑11%	0.757	0.849 ↑12%
DeF-SpeC	0.748	0.780 ↑4%	0.780	0.803 ↑3%	0.686	0.732 ↑7%	0.720	0.752 ↑4%	0.828	0.856 ↑3%
Gemini 2.0 Flash										
VaN	0.701	0.762 ↑9%	0.703	0.803 ↑14%	0.699	0.677 ↓3%	0.682	0.731 ↑7%	0.754	0.851 ↑13%
Z-CoT	0.670	0.733 ↑9%	0.667	0.763 ↑14%	0.675	0.670 ↓1%	0.646	0.694 ↑7%	0.736	0.838 ↑14%
DeF-SpeC	0.767	0.803 ↑5%	0.795	0.835 ↑5%	0.710	0.738 ↑4%	0.749	0.787 ↑5%	0.814	0.847 ↑4%
Gemma 3 27b it										
VaN	0.694	0.773 ↑11%	0.684	0.801 ↑17%	0.711	0.710 ↓0%	0.662	0.750 ↑13%	0.782	0.830 ↑6%
Z-CoT	0.622	0.767 ↑23%	0.671	0.793 ↑18%	0.516	0.711 ↑38%	0.561	0.746 ↑33%	0.775	0.822 ↑6%
DeF-SpeC	0.739	0.791 ↑7%	0.742	0.825 ↑11%	0.734	0.720 ↓2%	0.712	0.769 ↑8%	0.815	0.851 ↑4%
Llama 3.3 70B										
VaN	0.756	0.770 ↑2%	0.762	0.796 ↑4%	0.744	0.717 ↓4%	0.730	0.738 ↑1%	0.824	0.856 ↑4%
Z-CoT	0.736	0.781 ↑6%	0.739	0.804 ↑9%	0.730	0.733 ↑0%	0.714	0.748 ↑5%	0.793	0.867 ↑9%
DeF-SpeC	0.716	0.762 ↑6%	0.723	0.788 ↑9%	0.702	0.707 ↑1%	0.684	0.720 ↑5%	0.798	0.872 ↑9%
Average	0.716	0.778 ↑9%	0.731	0.805 ↑10%	0.686	0.720 ↑5%	0.689	0.751 ↑9%	0.789	0.849 ↑8%

Table 6.16: Results with F_1 scores for five LLMs. The *Base* columns shows the competitive method results, while the *IBI* columns presents results for prompts adapted to the Intent-based Inoculation.

The greatest gains are observed in longer-form articles. We hypothesize that the extended context in articles offers language models more opportunity to identify and reason about malicious intent.

Notably, IBI improves performance not only on data that may have been present during LLM pretraining, but also on unseen content published after

the models' knowledge cutoffs. While performance gains are evident in both pre- and post-cutoff subsets, LLMs show greater difficulty with the latter.

Overall, these findings support our central hypothesis: incorporating intent-based inoculation as an intermediate reasoning step for LLMs enhances zero-shot disinformation detection.

Multilingual Evaluation Table 6.17 shows the averaged F_1 scores across all models and methods for each language on the EUvsDisinfo dataset. IBI consistently improves performance over the Base setup for all six languages, with the largest gains observed in Estonian. These results indicate that intent-augmented reasoning enhances model disinformation detection capabilities across languages. Overall, IBI demonstrates a clear advantage in cross-lingual disinformation detection, achieving on average a 20% improvement over baseline methods.

Language	Base	IBI
German	0.794	0.911 ↑15%
Spanish	0.683	0.828 ↑21%
Estonian	0.716	0.892 ↑25%
French	0.611	0.749 ↑23%
Polish	0.709	0.846 ↑19%
Russian	0.619	0.735 ↑19%

Table 6.17: Comparison of Base vs IBI performance across languages. Results present averaged F_1 scores over all models and methods.

6.6 Discussion

This chapter introduced MALINT, the first human-annotated English-language dataset that jointly annotates disinformation and malicious intent. Beyond presenting a new resource, the study offers empirical insight into how intentionality can be operationalized and leveraged within computational disinformation research. The findings underscore that malicious intent is not only a defining conceptual feature of disinformation but also a practically useful signal for improved disinformation detection.

A central observation from the experiments is a clear divergence between small language models and large language models in intent classification. In the multilabel setting, fine-tuned SLMs consistently outperform LLMs, suggesting that supervised learning remains better suited for capturing complex, overlapping intent structures when high-quality annotations are available. In contrast, LLMs presents competitive, and in some cases superior, performance in binary

intent detection, indicating that pretrained models can recognize coarse-grained intentional signals even without task-specific training.

The intent-based inoculation experiments further provide strong empirical evidence that reasoning about malicious intent improves disinformation detection, particularly in zero-shot settings. Furthermore, by operationalizing inoculation theory within an LLM prompting framework, this study shows that exposing models to intent-oriented analysis functions as a form of cognitive "prebunking", strengthening downstream classification. Importantly, improvements are most pronounced for longer texts.

The multilingual results reinforce these findings. The effectiveness of IBI across six languages, including lower-resource languages, suggests that malicious intent is a language-agnostic signal that improves reasoning about disinformation. This finding has practical implications for disinformation detection in lower-resource languages, where annotated data are scarce and zero-shot approaches are often necessary.

At the same time, the analysis reveals important limitations. While correct intent prediction generally strengthens disinformation detection, improvements can still occur even when individual intents are misclassified. This indicates that intent signals may partially compensate for one another and that the benefits of intent-based reasoning are not strictly dependent on perfect intent recognition. However, this also raises questions about the interpretability of intent predictions and the extent to which models rely on correlated intent signals. These observations point to the need for deeper analysis of intent interactions and error propagation in multi-intent settings.

More broadly, this chapter supports the argument that moving beyond binary credibility labels toward intent-aware frameworks can improve both the effectiveness and the explanatory power of disinformation detection systems.

Future work can build on these insights by expanding intent taxonomies and integrating intent-based reasoning with other intermediate signals, such as persuasion or manipulation strategies. Such directions would further bridge the theory of disinformation with practice and explainable NLP systems.

7 Human and Machine-Generated Persuasive Text

This chapter presents contributions C4.1 - C4.3 presented in Section 1.3.
The chapter is based on the following work available as a preprint:
Arkadiusz Modzelewski, Paweł Golik, Anna Kołos, and Giovanni Da San Martino. *Can AI-Generated Persuasion Be Detected? Persuaficial Benchmark and AI vs. Human Linguistic Differences*.

Persuasive writing, which uses rhetorical techniques and devices to influence audiences, has become central to modern communication [186]. We live in an era where artificial intelligence increasingly shapes propaganda and persuasive communication in news, political discourse, and social media [187, 15]. Large Language Models, now widely used in writing and communication tasks, demonstrate a growing potential to produce persuasive text and influence public opinion [110, 108, 75, 107]. Several studies have explored how effectively LLMs can identify persuasive language [188, 189]. Yet, to the best of our knowledge, no prior work has addressed whether the automatic detection of LLM-generated persuasion is more challenging than detecting persuasion in human-written texts. Understanding this distinction is crucial, as it reveals the extent to which current detection systems may be vulnerable to increasingly sophisticated AI-driven persuasive content. Furthermore, although previous research has highlighted the importance of mitigating and defending against AI-generated persuasion [190, 191], it has largely focused on persuasion detection, leaving the linguistic differences between human-written and LLM-generated persuasive texts unexplored. Understanding these differences could deepen our knowledge of AI-driven persuasion and support the development of more effective automatic detection methods. To address these gaps, we investigate two key

research questions: **RQ1** *Is controllably generated AI persuasion harder for LLMs to detect in a zero-shot setting than human-written persuasion?* and **RQ2** *What are the linguistic differences between controllable LLM-generated and human-written persuasive texts?*

To address our first research question, we introduce **Persuaficial**, a newly constructed dataset of artificially generated persuasive texts. Persuaficial is a novel multilingual resource comprising synthetic persuasive content produced using four generation approaches inspired by Chen and Shu [56]. The dataset is created in a controlled manner, leveraging human-written texts drawn from established datasets [30, 192, 193]. In our experiments, we evaluate the detectability of AI and human-written persuasive texts using four different LLMs, including commercial closed models and open-weight models. Our analysis focuses on English, but we also provide analysis on five additional languages: German, French, Italian, Polish, and Russian.

To address RQ2, we conducted an analysis using the StyloMetrix tool[1], which generates fully interpretable and reproducible vectors representing a wide range of linguistic features in text [194]. Prior work shows that StyloMetrix performs well on persuasion detection with classical machine learning [195, 196], underscoring its suitability for studying the linguistic features of persuasive texts. In our analysis, we examined the full range of linguistic features offered by open-source StyloMetrix.

7.1 Human Persuasion Datasets

In our analysis, we employed three well-established, publicly available datasets that had been previously annotated by humans. Using multiple datasets mitigates potential bias that could arise from relying on a single source and ensures a broader coverage of persuasion phenomena. We selected datasets that are widely used and cited in persuasion research. Below, we describe the human-created datasets used in our study:

- **SemEval 2023 Task 3 Dataset**: A multilingual, multifaceted collection of online news articles annotated with various persuasion techniques on paragraph level [72]. Its taxonomy and dataset are widely adopted within the NLP community for persuasion research [197, 33, 18]. This dataset was introduced as part of SemEval 2023 Task 3 on persuasion detection [31].
- **DIPROMATS 2024 Task 1 Dataset**: A dataset consisting of posts from the X platform (former Twitter) used for the DIPROMATS 2024 shared task including propaganda detection [198]. The dataset contains messages from diplo-

[1] https://github.com/ZILiAT-NASK/StyloMetrix

mats and authorities of major world powers, including China, the United States, Russia, and the European Union. DIPROMATS 2024 was part of IberLEF, an annual Spanish shared evaluation campaign [199].

- **ChangeMyView**: A dataset derived from the Reddit *ChangeMyView* discussion community. Dataset contains 3,051 conversations in which the persuader tries to convince the persuadee to change their mind [192]. It is commonly used in persuasion research [200, 201, 202].

Together, these datasets provide diverse perspectives on human persuasion, covering different sources and communicative contexts. Consequently, our analysis captures diverse persuasive strategies, which improves the generalizability of our findings.

In our experiments, we adopt the concise definition of persuasion proposed by Piskorski et al. [72, 28]: "*Persuasive text is characterized by a specific use of language in order to influence the reader*". We rely on this definition because it underpins the annotation guidelines of the SemEval 2023 Task 3 dataset, the largest publicly available resource for studying persuasion and one of the three human-written datasets used in our study. Consequently, we selected additional datasets that align well with this conceptualization of persuasion.

DIPROMATS 2024 [198], which build on the SemEval task, offers data that is directly compatible with this definition and is therefore suitable for our experiments. Finally, Piskorski et al. [34] demonstrate that this definition can be effectively and reliably applied to debates and discussions, motivating our choice of the ChangeMyView dataset, comprising message exchanges between persuaders and persuadees, as an additional human-written source.

7.2 Persuaficial: Artificially Generated Persuasion Dataset

In this section, we introduce a new dataset, **Persuaficial**, which contains persuasion texts generated artificially using large language models.

7.2.1 Persuasive Text Generation Approaches

Inspired by Chen and Shu [56], which explores methods for generating synthetic misinformation, we developed a controllable persuasion generation G_P, in which P denotes one of four approaches for generating persuasive text: *Paraphrasing*, *Rewriting with Subtle Persuasion*, *Rewriting with Intensified Persuasion*, and *Open-ended* generation.

Each persuasive text was generated under controlled prompting conditions. Specifically, prompts included human-annotated persuasive texts together with

Approach	Prompt Example
Paraphrased Persuasion	Given a passage, please paraphrase it. The original content meaning and level of persuasion must be strictly preserved. The passage is: <passage>
Rewritten Persuasion (Subtle Effect)	Given a passage, please rewrite it to make its persuasiveness more subtle and harder to detect. The original content and meaning should be the same. The passage is: <passage>
Rewritten Persuasion (Intensified Effect)	Given a passage, please rewrite it so that the persuasion present in the content is refined for a stronger persuasive effect. The original content and meaning should be the same. The passage is: <passage>
Open-Ended Persuasion	Given a sentence, please write a short piece of text. The short text must contain persuasion. The sentence is: <sentence>

Table 7.1: Overview of four approaches used for generating persuasive texts with large language models. Each method represents a distinct level of control over persuasive strength and content nature.

instructions to paraphrase or rewrite the input while preserving, strengthening, or softening its persuasive effect. For the open-ended generation setting, following the approach of Chen and Shu [56], the model was provided with concise summaries of the corresponding human-annotated persuasive examples. This procedure was applied to all selected instances across the chosen texts, using identical prompt templates for all languages. For non-English cases, we appended instructions specifying the target language of generation.

Controlling the generation process through explicit instructions is crucial for our study, as it ensures that the resulting LLM-generated persuasive texts remain semantically comparable to human-written texts. This comparability is essential for reliable evaluation of both persuasion detection performance and linguistic differences between human- and AI-generated persuasive texts.

All approaches for generating synthetic text with persuasion, along with example prompts, are included in Table 7.1. More details about the prompts for the generation of the Persuaficial dataset are shown in the Appendix A.7.1.

7.2.2 Persuaficial Dataset Construction

Persuaficial is an AI-generated persuasion dataset constructed using multiple LLMs and diverse generation approaches. For *Paraphrasing*, *Rewriting (Subtle Effect)*, and *Rewriting (Intensified Effect)*, we sample 1,000 texts from three real-world persuasion datasets (described in Section 7.1). Each sample includes 500 texts annotated as persuasive and 500 labeled as non-persuasive[2].

[2]The only exception was German: the corpus contained only 420 non-persuasive texts, so we included all of them and randomly sampled 580 persuasive texts.

Each selected persuasive text is treated as <passage> (see Table 7.1) and serves as input for the generation method. For *Open-ended* generation, we first summarize each selected persuasive text into a factual statements. We use the resulting <sentence> for the generation of persuasive synthetic text.

We employ open-weight and proprietary LLMs for dataset construction. The open models include *Gemma 3 27b it* and *Llama 3.3 70B*. The commercial models are *Gemini 2.0 Flash* and *GPT 4.1 Mini*. Our selection of models aimed to cover widely recognized, high-performing models while balancing accessibility and cost. Additionally, we included two open-weight models to provide experiments that can be reproduced without reliance on closed API-based models. To encourage more diverse, creative and less repetitive phrasing in the model outputs, we set the generation temperature to 0.8. Our choice of temperature for generating synthetic persuasive texts was directly informed by the settings used by Chen and Shu [56] in a related task involving misinformation generation.

7.2.3 Pre-Generation Quality Evaluation

As mentioned, for the *Open-ended* generation setting, we first summarize each selected persuasive text into a short <sentence>. To ensure these <sentence>s accurately represent the source human text, we conducted a human evaluation following explicit and rigorous annotation guidelines (see Appendix A.8.1).

Two annotators were first trained by author of this dissertation, who has prior experience in annotation. A small training sample of 50 English sentences was selected, and the annotators independently applied the annotation guidelines, with opportunities to discuss their decisions. After completing the training phase, they reviewed and discussed their independent evaluations to align their understanding of the guidelines. The annotations from this training phase were excluded from further evaluation.

For the final evaluation, a sample of 200 English <sentence>s was selected. Two independent annotators assessed each <sentence> for factual correspondence. We then computed the accuracy of the LLM-generated <sentence>s, considering as positive only those instances where both annotators independently agreed that a sentence was factual. The resulting accuracy of the LLM generation process was about 91.2%, suggesting that LLMs may be effective at transforming texts into a short factual statements. Moreover, most mismatches between the generated <sentence>s and source texts were minor in nature, e.g., converting an exclamatory formulation ("*Introduce the law!*") into a declarative one ("*The law will be introduced.*"). This result suggests that the generated factual sentences are of generally high quality and are unlikely to negatively impact the overall quality of the resulting Persuaficial dataset.

7.2.4 Post-Generation Quality Evaluation

To ensure that the final Persuaficial dataset meets the intended goal of containing persuasive content produced by LLMs under controlled prompting conditions, we conducted a multi-stage post-generation quality evaluation. While the pre-generation evaluation ensured that the factual sentences for *Open-ended* approach were valid, the post-generation evaluation verifies whether the LLM-generated persuasive variants are (1) faithful to the target factual content, (2) persuasive, and (3) faithful to the instruction from persuasion generation approach.

We adopted a two-layer rigorous verification design that separates basic validity checking from persuasion-specific judging. We verified 400 generated English texts, each independently annotated by two trained annotators following detailed instructions (Appendix A.8.2). As a result, we report an overall accuracy metric, defined as the proportion of generated texts unanimously annotated as valid by two annotators, where validity required that all three criteria received a positive annotation.

Due to the conservative requirement that all three criteria be jointly satisfied, the overall accuracy is 88.2%. Most invalid cases involve only minor factual deviations rather than substantive inconsistencies. When considering persuasion-related criteria alone, accuracy grows to 97.69%, indicating that LLMs reliably generate persuasive text and justifying the use of this data for subsequent comparisons between human- and AI-generated persuasive texts.

7.3 Persuaficial Dataset Statistics

For each language, we sampled 1,000 human-written passages from the original datasets and generated persuasive variants using four LLMs and four generation approaches (4 models × 4 approaches = 16 generation configurations). This resulted in approximately 24,000 texts in English and 40,000 texts for the non-English languages. In total, Persuaficial is a multilingual corpus of about 64,000 texts. Table 7.2 presents basic statistics for our dataset.

7.4 Automatic Classification of Human and AI-Generated Persuasion

7.4.1 Experimental Setup

For our experiments, we use Persuaficial, which comprises artificially generated persuasive texts. Moreover, we use human-written counterparts. Each experiment uses data that is balanced across persuasive and non-persuasive classes.

Type	Persuaficial	English	French	German	Italian	Polish	Russian
Average number of words (Avg_w) per text.							
Human	–	117	46	44	48	46	40
Intensified	87	118	75	68	75	70	55
Open_end	65	60	76	71	75	61	58
Paraphrased	81	111	66	59	66	62	48
Rewritten	87	110	69	63	70	66	52
Average number of characters (Avg_{ch}) per text.							
Human	–	695	288	314	313	327	285
Intensified	605	791	498	512	511	526	417
Open_end	450	391	498	524	507	454	442
Paraphrased	538	727	430	434	443	459	360
Rewritten	568	742	465	476	488	493	400
Number of texts (Count) per dataset and generation type.							
Human	–	3000	1000	1000	1000	1000	1000
Intensified	16320	6000	2000	2320	2000	2000	2000
Open_end	16320	6000	2000	2320	2000	2000	2000
Paraphrased	16320	6000	2000	2320	2000	2000	2000
Rewritten	16320	6000	2000	2320	2000	2000	2000

Table 7.2: Basic statistics for LLM-generated persuasive texts in our Persuaficial dataset and human-written counterparts. We present basic statistics for general full dataset, but also on samples that present all languages.

For automatic persuasion detection, we employed four LLMs: *GPT-4.1 Mini*, *Gemini 2.0 Flash*, *Gemma 3 27B Instruct*, and *Llama 3.3 70B*. To ensure as deterministic outputs as possible, we set the temperature to 0 during classification. Since our goal is to detect persuasion, we formulate the task as a binary classification problem. All classifications were performed in a zero-shot setting with temperature set to 0.0 to ensure as much determinism in the classification predictions as possible. Zero-shot approach aligns with our research question (RQ1). Moreover, studies show that zero-shot detection with modern LLMs (e.g., GPT-4) can outperform supervised models such as BERT on binary classification tasks [65, 66, 67]. Furthermore, Lucas et al. [13] and Modzelewski et al. [18] report that while fine-tuning BERT on multiple datasets results in poor generalization to unseen data, zero-shot LLMs maintain strong cross-domain performance. We evaluate persuasion detection performance using the F_1 score. Further details supporting reproducibility, specifically the prompt templates used for persuasion detection, are provided in Appendix A.7.2.

7.4.2 Results on English Datasets

Table 7.3 reports F_1 scores for persuasion detection on three human-written balanced samples and their LLM-generated counterparts produced using four

generation approaches.

Classifier	Human-written	Paraphrase	Rewriting		Open-ended
			Subtle	Intensive	
Sample of Persuaficial generated based on: SemEval 2023 data					
GPT 4.1 Mini	0.7398	0.7007 ↓5%	0.4031 ↓46%	0.8148 ↑10%	0.8964 ↑21%
Llama 3.3 70B	0.7459	0.7207 ↓3%	0.4577 ↓39%	0.8111 ↑9%	0.8741 ↑17%
Gemma 3 27b it	0.7572	0.7592 ↑0%	0.6453 ↓15%	0.8208 ↑8%	0.8562 ↑13%
Gemini 2.0 Flash	0.7551	0.7540 ↓0%	0.6522 ↓14%	0.7950 ↑5%	0.8117 ↑7%
Sample of Persuaficial generated based on: DIPROMATS 2024 data					
GPT 4.1 Mini	0.7567	0.7461 ↓1%	0.4962 ↓34%	0.8100 ↑7%	0.8666 ↑15%
Llama 3.3 70B	0.7471	0.7362 ↓1%	0.5696 ↓24%	0.7860 ↑5%	0.8292 ↑11%
Gemma 3 27b it	0.7473	0.7460 ↓0%	0.6349 ↓15%	0.7782 ↑4%	0.7994 ↑7%
Gemini 2.0 Flash	0.7518	0.7427 ↓1%	0.6664 ↓11%	0.7640 ↑2%	0.7680 ↑2%
Sample of Persuaficial generated based on: ChangeMyView data					
GPT 4.1 Mini	0.6233	0.6356 ↑2%	0.4906 ↓21%	0.6739 ↑8%	0.7148 ↑15%
Llama 3.3 70B	0.6517	0.6488 ↓0%	0.5536 ↓15%	0.6691 ↑3%	0.6831 ↑5%
Gemma 3 27b it	0.6644	0.6708 ↑1%	0.6334 ↓5%	0.6809 ↑2%	0.6843 ↑3%
Gemini 2.0 Flash	0.6671	0.6662 ↓0%	0.6294 ↓6%	0.6740 ↑1%	0.6770 ↑1%

Table 7.3: F_1 scores for persuasion detection on English data. The first column reports performance on human-annotated texts. The remaining columns show performance on LLM-generated texts. For generated data, each value represents the average F_1 score obtained when classifying texts generated by four different LLMs.

On the *Paraphrasing* subset of our Persuaficial dataset, F_1 scores are only marginally lower than those for human-written texts (on average 0.67% lower), indicating that paraphrasing preserves a similar level of difficulty for persuasion detection across human and generated texts. In contrast, *Rewriting (Intensified)* and *Open-ended* subsets yield the highest F_1 scores. On average, persuasion is 9.75% easier to detect in open-ended scenario and 5.33% easier when persuasion is intensified. This makes open-ended generated persuasive texts the easiest setting for LLM-based detection. We hypothesize that models tend to over-express explicit persuasive cues when prompted to generate persuasive text freely or while intensifying persuasion, which in turn makes these texts more easily detectable. The opposite pattern emerges for *Rewriting (Subtle persuasion)*, where F_1 scores drop substantially, by 20.42% on average. This suggests that reducing overt persuasive markers makes persuasion significantly harder to detect, even for strong LLM detectors. Importantly, these patterns are highly consistent across datasets and across all detector models. This may indicate that the effects generalize across domains and are independent of the specific LLM used for detection.

Moreover, we present detailed F_1 scores for persuasion detection across different subsets of the Persuaficial datasets and LLM generation strategies. Table 7.4 reports results for SemEval 2023 Task 3 texts and their AI-generated counterparts, Table 7.5 for the DIPROMATS 2024 dataset, and Table 7.6 for the ChangeMyView dataset. For each dataset, classifier performance on human-written texts (first

column) is compared with performance on LLM-generated texts produced via paraphrasing, rewriting with subtle or intensive persuasion, and open-ended generation. Results are further broken down by both the generating model and the classifier model, highlighting how different generation approaches influence the detectability of persuasion.

Classifier	Human-written	Paraphrase	Rewriting		Open-ended
			Subtle	Intensive	
Generating model: GPT 4.1 Mini					
GPT 4.1 Mini	0.7398	0.6984 ↓6%	0.3837 ↓48%	0.7638 ↑3%	0.8969 ↑21%
Llama 3.3 70B	0.7459	0.7310 ↓2%	0.4444 ↓40%	0.7757 ↑4%	0.8741 ↑17%
Gemma 3 27b it	0.7572	0.7561 ↓0%	0.6213 ↓18%	0.8007 ↑6%	0.8562 ↑13%
Gemini 2.0 Flash	0.7551	0.7487 ↓1%	0.6407 ↓15%	0.7780 ↑3%	0.8117 ↑7%
Generating model: Llama 3.3 70B					
GPT 4.1 Mini	0.7398	0.6831 ↓8%	0.4411 ↓40%	0.7870 ↑6%	0.8969 ↑21%
Llama 3.3 70B	0.7459	0.6911 ↓7%	0.4811 ↓35%	0.7791 ↑4%	0.8741 ↑17%
Gemma 3 27b it	0.7572	0.7479 ↓1%	0.6746 ↓11%	0.8082 ↑7%	0.8562 ↑13%
Gemini 2.0 Flash	0.7551	0.7476 ↓1%	0.6691 ↓11%	0.7921 ↑5%	0.8117 ↑7%
Generating model: Gemma 3 27b it					
GPT 4.1 Mini	0.7398	0.7025 ↓5%	0.3445 ↓53%	0.8611 ↑16%	0.8969 ↑21%
Llama 3.3 70B	0.7459	0.7222 ↓3%	0.4118 ↓45%	0.8469 ↑14%	0.8741 ↑17%
Gemma 3 27b it	0.7572	0.7675 ↑1%	0.6241 ↓18%	0.8383 ↑11%	0.8562 ↑13%
Gemini 2.0 Flash	0.7551	0.7614 ↑1%	0.6202 ↓18%	0.8039 ↑6%	0.8117 ↑7%
Generating model: Gemini 2.0 Flash					
GPT 4.1 Mini	0.7398	0.7188 ↓3%	0.4430 ↓40%	0.8472 ↑15%	0.8949 ↑21%
Llama 3.3 70B	0.7459	0.7385 ↓1%	0.4936 ↓34%	0.8428 ↑13%	0.8741 ↑17%
Gemma 3 27b it	0.7572	0.7652 ↑1%	0.6613 ↓13%	0.8362 ↑10%	0.8562 ↑13%
Gemini 2.0 Flash	0.7551	0.7583 ↑0%	0.6787 ↓10%	0.8059 ↑7%	0.8117 ↑7%

Table 7.4: F_1 scores for persuasion detection on English data sample of Persuaficial. More specifically, on sample of Persuaficial generated using English texts from SemEval 2023 Task 3 dataset. The first column reports performance on English texts from SemEval 2023 Task 3 human-annotated texts. The remaining columns show performance on LLM-generated English counterparts.

In summary, addressing RQ1, the detectability of LLM-generated persuasive text depends on the generation approach: texts produced via open-ended and intensified persuasion are easier to detect, whereas subtly persuasive generations remain substantially more challenging for current LLM-based detectors.

7.4.3 Results on Non-English Datasets

Table 7.7 shows persuasion detection results for German, French, Italian, Polish, and Russian. The patterns observed in English hold consistently across all languages and classifiers. Paraphrasing preserves a difficulty level similar to human-written texts, whereas intensified rewriting and open-ended generation yield the highest F_1 scores. Open-ended generation often exceeds F_1 of 0.90 which gives the easiest to detect persuasion. In contrast, subtle rewriting causes the largest drop in performance. Overall, these trends suggest that generation

Classifier	Human-written	Paraphrase	Rewriting		Open-ended
			Subtle	Intensive	
Generating model: GPT 4.1 Mini					
GPT 4.1 Mini	0.7567	0.7435 ↓2%	0.4948 ↓35%	0.7866 ↑4%	0.8666 ↑15%
Llama 3.3 70B	0.7471	0.7348 ↓2%	0.5795 ↓22%	0.7679 ↑3%	0.8292 ↑11%
Gemma 3 27b it	0.7473	0.7441 ↓0%	0.6308 ↓16%	0.7607 ↑2%	0.7994 ↑7%
Gemini 2.0 Flash	0.7518	0.7449 ↓1%	0.6711 ↓11%	0.7595 ↑1%	0.7680 ↑2%
Generating model: Llama 3.3 70B					
GPT 4.1 Mini	0.7567	0.7338 ↓3%	0.5595 ↓26%	0.7967 ↑5%	0.8666 ↑15%
Llama 3.3 70B	0.7471	0.7314 ↓2%	0.6251 ↓16%	0.7775 ↑4%	0.8292 ↑11%
Gemma 3 27b it	0.7473	0.7410 ↓1%	0.6928 ↓7%	0.7749 ↑4%	0.7994 ↑7%
Gemini 2.0 Flash	0.7518	0.7400 ↓2%	0.7002 ↓7%	0.7652 ↑2%	0.7680 ↑2%
Generating model: Gemma 3 27b it					
GPT 4.1 Mini	0.7567	0.7672 ↑1%	0.3951 ↓48%	0.8404 ↑11%	0.8666 ↑15%
Llama 3.3 70B	0.7471	0.7449 ↓0%	0.4898 ↓34%	0.8074 ↑8%	0.8292 ↑11%
Gemma 3 27b it	0.7473	0.7504 ↑0%	0.5777 ↓23%	0.7936 ↑6%	0.7994 ↑7%
Gemini 2.0 Flash	0.7518	0.7459 ↓1%	0.6232 ↓17%	0.7662 ↑2%	0.7680 ↑2%
Generating model: Gemini 2.0 Flash					
GPT 4.1 Mini	0.7567	0.7399 ↓2%	0.5353 ↓29%	0.8163 ↑8%	0.8666 ↑15%
Llama 3.3 70B	0.7471	0.7336 ↓2%	0.5838 ↓22%	0.7911 ↑6%	0.8292 ↑11%
Gemma 3 27b it	0.7473	0.7483 ↑0%	0.6383 ↓15%	0.7838 ↑5%	0.7994 ↑7%
Gemini 2.0 Flash	0.7518	0.7400 ↓2%	0.6711 ↓11%	0.7652 ↑2%	0.7680 ↑2%

Table 7.5: F_1 scores for persuasion detection on English data sample of Persuaficial. More specifically, on sample of Persuaficial generated using DIPROMATS 2024 dataset. The first column reports performance on DIPROMATS 2024 human-annotated texts. The remaining columns show performance on LLM-generated counterparts.

Classifier	Human-written	Paraphrase	Rewriting		Open-ended
			Subtle	Intensive	
Generating model: GPT 4.1 Mini					
GPT 4.1 Mini	0.6233	0.6337 ↑2%	0.4941 ↓21%	0.6582 ↑6%	0.7148 ↑15%
Llama 3.3 70B	0.6517	0.6546 ↑0%	0.5665 ↓13%	0.6667 ↑2%	0.6831 ↑5%
Gemma 3 27b it	0.6644	0.6745 ↑2%	0.6388 ↓4%	0.6809 ↑2%	0.6836 ↑3%
Gemini 2.0 Flash	0.6671	0.6662 ↓0%	0.6431 ↓4%	0.6726 ↑1%	0.6770 ↑1%
Generating model: Llama 3.3 70B					
GPT 4.1 Mini	0.6233	0.6137 ↓2%	0.4619 ↓26%	0.6481 ↑4%	0.7148 ↑15%
Llama 3.3 70B	0.6517	0.6307 ↓3%	0.5226 ↓20%	0.6564 ↑1%	0.6831 ↑5%
Gemma 3 27b it	0.6644	0.6644 ↓0%	0.6133 ↓8%	0.6772 ↑2%	0.6845 ↑3%
Gemini 2.0 Flash	0.6671	0.6607 ↓1%	0.6112 ↓8%	0.6717 ↑1%	0.6770 ↑1%
Generating model: Gemma 3 27b it					
GPT 4.1 Mini	0.6233	0.6572 ↑5%	0.4840 ↓22%	0.7009 ↑12%	0.7148 ↑15%
Llama 3.3 70B	0.6517	0.6574 ↑1%	0.5515 ↓15%	0.6758 ↑4%	0.6831 ↑5%
Gemma 3 27b it	0.6644	0.6754 ↑2%	0.6349 ↓4%	0.6836 ↑3%	0.6845 ↑3%
Gemini 2.0 Flash	0.6671	0.6689 ↑0%	0.6259 ↓6%	0.6762 ↑1%	0.6770 ↑1%
Generating model: Gemini 2.0 Flash					
GPT 4.1 Mini	0.6233	0.6379 ↑2%	0.5226 ↓16%	0.6885 ↑10%	0.7148 ↑15%
Llama 3.3 70B	0.6517	0.6527 ↑0%	0.5740 ↓12%	0.6776 ↑4%	0.6831 ↑5%
Gemma 3 27b it	0.6644	0.6690 ↑1%	0.6465 ↓3%	0.6818 ↑3%	0.6845 ↑3%
Gemini 2.0 Flash	0.6671	0.6689 ↑0%	0.6374 ↓4%	0.6753 ↑1%	0.6770 ↑1%

Table 7.6: F_1 scores for persuasion detection on English data sample of Persuaficial. More specifically, on sample of Persuaficial generated using ChangeMyView dataset. The first column reports performance on ChangeMyView human-annotated texts. The remaining columns show performance on LLM-generated counterparts.

approaches influence persuasion detectability, with effects that may generalize across languages.

Classifier	Human-written	Paraphrase	Rewriting		Open-ended
			Subtle	Intensive	
German					
GPT 4.1 Mini	0.7203	0.7207 ↑0%	0.4410 ↓39%	0.8456 ↑17%	0.9414 ↑31%
Llama 3.3 70B	0.7361	0.7248 ↓2%	0.4398 ↓40%	0.8474 ↑15%	0.9345 ↑27%
Gemma 3 27b it	0.7655	0.7763 ↑1%	0.6664 ↓13%	0.8512 ↑11%	0.9004 ↑18%
Gemini 2.0 Flash	0.7903	0.7880 ↓0%	0.6905 ↓13%	0.8385 ↑6%	0.8591 ↑9%
French					
GPT 4.1 Mini	0.7505	0.7454 ↓1%	0.4290 ↓43%	0.8456 ↑13%	0.9251 ↑23%
Llama 3.3 70B	0.7605	0.7450 ↓2%	0.4527 ↓40%	0.8432 ↑11%	0.9172 ↑21%
Gemma 3 27b it	0.7827	0.7866 ↑0%	0.6587 ↓16%	0.8476 ↑8%	0.8824 ↑13%
Gemini 2.0 Flash	0.7812	0.7860 ↑1%	0.6800 ↓13%	0.8314 ↑6%	0.8418 ↑8%
Italian					
GPT 4.1 Mini	0.7471	0.7330 ↓2%	0.4246 ↓43%	0.8428 ↑13%	0.9195 ↑23%
Llama 3.3 70B	0.7584	0.7172 ↓5%	0.4285 ↓43%	0.8420 ↑11%	0.9161 ↑21%
Gemma 3 27b it	0.7659	0.7804 ↑2%	0.6610 ↓14%	0.8399 ↑10%	0.8686 ↑13%
Gemini 2.0 Flash	0.7986	0.7938 ↓1%	0.6781 ↓15%	0.8301 ↑4%	0.8408 ↑5%
Polish					
GPT 4.1 Mini	0.7330	0.7060 ↓4%	0.4580 ↓38%	0.8483 ↑16%	0.9367 ↑28%
Llama 3.3 70B	0.7676	0.7389 ↓4%	0.5041 ↓34%	0.8518 ↑11%	0.9206 ↑20%
Gemma 3 27b it	0.7728	0.7783 ↑1%	0.6919 ↓10%	0.8427 ↑9%	0.8834 ↑14%
Gemini 2.0 Flash	0.7733	0.7732 ↓0%	0.7018 ↓9%	0.8101 ↑5%	0.8217 ↑6%
Russian					
GPT 4.1 Mini	0.7246	0.7073 ↓2%	0.4392 ↓39%	0.8242 ↑14%	0.9017 ↑24%
Llama 3.3 70B	0.7408	0.7164 ↓3%	0.4312 ↓42%	0.8324 ↑12%	0.9086 ↑23%
Gemma 3 27b it	0.7360	0.7416 ↑1%	0.6128 ↓17%	0.8098 ↑10%	0.8562 ↑16%
Gemini 2.0 Flash	0.7683	0.7616 ↓1%	0.6795 ↓12%	0.7877 ↑3%	0.8019 ↑4%

Table 7.7: F_1 scores for persuasion detection on non-English data samples. The first column reports performance on human-annotated texts. The remaining columns show performance on LLM-generated texts. For generated data, each value represents the average F_1 score obtained when classifying texts generated by four different LLMs.

7.5 Linguistic Differences Between Machine and Human Persuasion

In this section, we investigate the linguistic differences between human-written and AI-generated persuasive texts in English. We focus on English due to the limited availability of high-quality datasets in other languages. While the SemEval 2023 Task 3 data provides a multilingual resource [203, 30], relying solely on it could introduce dataset-specific biases. To mitigate this, we use Persuaficial, including samples of three established English datasets with their AI-generated counterparts.

7.5.1 Our Approach for Linguistic Analysis

To investigate the linguistic differences between human-written persuasive texts and LLM-generated persuasive texts, we adopt an explainable, feature-based analysis grounded in stylometry. Our objective is to identify the linguistic features that most strongly differentiate LLM-generated persuasive texts from human-written ones. For each linguistic feature, we compare the distributions of the two groups using effect-size-based analysis together with significance testing. Effect sizes quantify the magnitude of the difference between human and AI-generated texts for each feature [204], while significance testing evaluates whether these distributional differences are statistically meaningful. Our analysis identifies which linguistic properties differ systematically between human- and AI-produced persuasive texts.

7.5.2 StyloMetrix for Linguistic Analysis of Persuasive Texts

For our analysis, we utilize StyloMetrix, because it is an open-source tool that provides fully interpretable, linguistically grounded feature representations. Moreover, prior work has demonstrated its effectiveness for persuasion detection using classical machine learning models [195, 196], confirming its suitability for analyzing the linguistic characteristics of persuasive texts. Finally, to further justify our choice, in this section we show that StyloMetrix features with classical machine learning can distinguish human-written from AI-generated persuasive texts.

Experimental Setup. We conduct a classification study using classical machine learning models with features calculated by StyloMetrix. We aim to prove that linguistic features contain enough information to differentiate AI-generated persuasion from human-written persuasion.

For all English synthetic texts generated by *GPT 4.1. Mini* with each generation approach, we trained a separate classifier. For each experiment, we split the data into training and test sets, allocating 70% for training and 30% for testing. To ensure a credible evaluation, each human-written text and its LLM-generated counterpart were placed in the same split, either training or test. This prevents the classifier from exploiting the potential direct similarities between the paired texts. We employed widely used tree-based machine learning methods as classifiers, as they naturally capture non-linear interactions and are well-suited for moderate- to high-dimensional tabular data [205, 206]. Previous work has shown that, for tabular data, tree-based models can even outperform deep learning approaches [205].

Results. Table 7.8 shows the results of these experiments. The outcomes show a clear progression: the more generative freedom the LLM is given, the easier it becomes for tree-ensemble models to distinguish its outputs from human-written persuasion. *Paraphrasing* keeps the AI text close to the original human style, yielding only moderate detection performance. In the *Rewriting* conditions, the model introduces larger stylistic shifts, whether by making persuasion subtler or more intense, which improves separability. *Open-ended* generation, starting from only a brief neutral summary, produces the greatest stylistic divergence and thus the highest classification accuracy. Overall, linguistic features become increasingly informative as the generation task becomes less constrained. This analysis further proves the usefulness of StyloMetrix for comparison between linguistic features on persuasive human-written texts vs. AI-generated.

Model	Precision	Recall	F1	Accuracy
GPT 4.1 Mini – Paraphrasing				
RF	0.74	0.74	0.74	0.74
XGB	0.76	0.76	0.76	0.76
LGBM	0.76	0.76	0.76	0.76
GPT 4.1 Mini – Rewriting with Subtle Persuasive Effect				
RF	0.84	0.84	0.84	0.84
XGB	0.87	0.87	0.87	0.87
LGBM	0.86	0.86	0.86	0.86
GPT 4.1 Mini – Rewriting with Intensified Persuasive Effect				
RF	0.83	0.83	0.83	0.83
XGB	0.84	0.84	0.84	0.84
LGBM	0.86	0.85	0.85	0.85
GPT 4.1 Mini – Open-ended				
RF	0.97	0.97	0.97	0.97
XGB	0.97	0.97	0.97	0.97
LGBM	0.98	0.98	0.98	0.98

Table 7.8: Classification performance of detecting GPT-4.1-Mini-generated persuasive texts versus human-written persuasive texts using linguistic-feature representations. Each experiment reflects a different AI-generation strategy and uses data combined from three English datasets.

7.5.3 Experimental Setup for Linguistic Analysis

We first represent each persuasive human text and its AI-generated counterpart using StyloMetrix [194]. For each text, we calculate a 196-dimensional vector of linguistic features. This results in a tabular representation, where each row corresponds to a text encoded by its computed linguistic features.

For each linguistic feature, we directly compare its distribution in human-written and LLM-generated persuasive texts, conducting this analysis separately for each generation approach. We utilize Cohen's d statistic which is a type of effect size measure used to represent the magnitude of differences between two groups on a given variable [204]. For each linguistic feature, we calculate

Cohen's d to quantify the magnitude of the shift between the feature distribution of generated texts g_i and that of their human-written persuasive counterparts r_i. Cohen's d is defined in Equation 7.5 and is computed as follows.

First, we calculate the mean of each feature for human-written and generated texts:

$$\bar{r} = \frac{1}{n_r}\sum_{i=1}^{n_r} r_i, \qquad \bar{g} = \frac{1}{n_g}\sum_{i=1}^{n_g} g_i, \tag{7.1}$$

where n_r and n_g denote the number of human-written and generated texts, respectively. In our experiments, $n_r = n_g$ as for each human-written persuasive text we have AI-generated counterpart.

Next, we compute the sample standard deviations for each group:

$$s_r = \sqrt{\frac{1}{n_r - 1}\sum_{i=1}^{n_r}(r_i - \bar{r})^2}, \tag{7.2}$$

$$s_g = \sqrt{\frac{1}{n_g - 1}\sum_{i=1}^{n_g}(g_i - \bar{g})^2}. \tag{7.3}$$

Using these, we calculate the pooled standard deviation:

$$s_{\text{pooled}} = \sqrt{\frac{(n_r - 1)s_r^2 + (n_g - 1)s_g^2}{n_r + n_g - 2}}. \tag{7.4}$$

Finally, Cohen's d is obtained as the difference in means normalized by the pooled standard deviation:

$$d = \frac{\bar{g} - \bar{r}}{s_{\text{pooled}}} \tag{7.5}$$

This effect size, Cohen's d, provides a standardized measure of the shift in feature distributions between generated and human-written persuasive texts, allowing comparison across features and models.

To assess whether feature distributions differ significantly between human-written and AI-generated texts, we apply the Wilcoxon signed-rank test [207]. This non-parametric paired test is well suited for comparing matched text pairs and is widely used in NLP research [208, 164], including studies that analyze differences in feature distributions [209, 210, 211, 212].

Because many feature distributions may deviate from normality, we avoid parametric assumptions when testing for statistical significance. The test is applied to per-text differences $d_i = g_i - r_i$. To account for multiple comparisons aris-

ing from testing each feature independently, we apply the Benjamini-Hochberg false discovery rate (FDR) correction [213] across all features.

G_ACTIVE The proportion of verbs in the text used in the active voice.
L_ADV_SUPERLATIVE Measures how often superlative adverbial (and some adjective-as-adverb) forms appear in a text.
L_ADV_COMPARATIVE The proportion of tokens that are adverbs used in comparative degree (e.g., "more", "less", or marked comparative forms)
L_FUNC_A The proportion of tokens in a text that are function words.
L_CONT_T The proportion of unique content-word forms in relative to total tokens.
L_CONT_A The proportion of tokens in a text that are content words.
L_PUNCT_COM Comma incidence measures frequency of commas relative to text length.
L_PUNCT_DASH Measures the density of dashes within a text.
L_PUNCT_DOT Measures the incidence of periods (dots) relative to the total number of words.
L_PLURAL_NOUNS Measures the density of plural nouns within a text.
LTOKEN_RATIO_LEM (ST_TYPE_TOKEN_RATIO_LEMMAS) The ratio of unique lemmas to total tokens.
POS_ADJ The proportion of tokens in the text that are adjectives, indicating the level of descriptiveness.
POS_NOUN The proportion of tokens in the text that are nouns.
POS_PRO The proportion of tokens in the text that are pronouns.
PS_CAUSE Measures the incidence of linking words and phrases related to cause and purpose.
SENT_D_NP (ST_SENT_D_NP) Measures the average proportion of noun phrase (NPs) tokens relative to sentence length, averaged over all sentences in a document.
SENT_D_PP (ST_SENT_D_PP) Measures the average proportion of tokens that belong to prepositional phrases (PPs) in each sentence, averaged over all sentences in the document.
SENT_D_VP (ST_SENT_D_VP) Measures the average proportion of tokens in a sentence that are not marked with a verb tense, relative to total sentence length, averaged over all sentences in the document.
SENT_ST_DIFFER (ST_SENT_DIFFERENCE) Quantifies syntactic variation between consecutive sentences by comparing their dependency label sets and averaging over the document.
SENT_ST_WPERSENT Indicates the normalized difference between the total number of tokens and the number of sentences in a document (a proxy for sentence length).
ST_REPET_WORDS (ST_REPETITIONS_WORDS) Measures the level of lexical repetition by computing the proportion of repeated word tokens in a text normalized by total token count.
SY_EXCLAMATION Measures the proportion of unique word tokens that appear in exclamatory sentences relative to all tokens in the text.
SY_IMPERATIVE Measures the proportion of unique alphabetic words that appear in sentences classified as imperative, relative to all tokens in the document.
SY_INV_PATTERNS (SY_INVERSE_PATTERNS) A syntactic feature that measures the frequency of inverted sentence structures within a text.
SY_NARRATIVE Measures the proportion of tokens in declarative sentences relative to all tokens.
VF_INFINITIVE A syntactic feature that measures the proportion of infinitive verb forms.
VT_MIGHT Measures the frequency of "might" in a text.

Figure 7.1: StyloMetrix features with the highest discriminative importance in distinguishing human-written from LLM-generated persuasive text.

7.5.4 Results and Analysis

We computed Cohen's d values for 196 linguistic features across four generation approaches and four LLMs utilized to generate texts for our Persuaficial dataset. Table 7.9 sorts top linguistic features by the absolute Cohen's d ($|C_d|$) values per model and generation approach (G_P). Our analysis and discussion is based on the twenty features with the largest $|C_d|$ for each model and generation strategy. Tables from 7.10 to 7.25 report the top 20 features that most strongly

distinguish AI-generated persuasive texts from human-written persuasive texts. We provide 16 tables in total, reflecting Cohen's d effect sizes computed separately for each generating model and for each generation setting: Paraphrasing, Rewriting subtle persuasion, Rewriting intensified persuasion, and Open-ended generation. Statistical analysis using Wilcoxon signed-rank tests confirmed that all twenty features for each scenario exhibit significant distributional differences between human-written and AI-generated persuasive texts. Figure 7.1 provides definitions of the key differentiating features.

High values of features such as L_CONT_T (the proportion of unique content-word forms relative to total tokens), LTOKEN_RATIO_LEM (the ratio of unique lemmas to total tokens), and L_CONT_A (the proportion of tokens that are content words) indicate that AI-generated texts tend to contain more varied content words and higher informational density per sentence. These patterns suggest that lexical diversity and content richness are characteristic markers of AI authorship. Similarly, low values for ST_REPET_WORDS are indicative of AI-generated persuasive texts, suggesting that reduced word repetition serves as a strong signal of LLM-generated text. A higher proportion of function words (L_FUNC_A) indicates that a persuasive text is likely human-written. This means that frequent use of grammatical connectors (such as articles, prepositions, pronouns, and auxiliary verbs) is a signal of human text. Furthermore, AI-generated persuasive texts generally exhibit lower punctuation density, especially in texts from *Llama* and *GPT-4.1-mini*. However, certain punctuation marks, including commas (L_PUNCT_COM) and dashes (L_PUNCT_DASH), occur more frequently in these AI texts. The rarity of syntactically marked constructions, such as inversions (SY_INV_PATTERNS), is a distinguishing feature of AI text, as these complex syntactic patterns are more typical of human-written persuasive texts.

In AI-generated texts that intensify persuasion, comparative and superlative adverbs (L_ADV_COMPARATIVE with words like "more", "faster", and L_ADV_SUPERLATIVE with words like "best", "worst") appear more frequently than in human-written texts. This suggests that AI strengthens persuasive language through the increased use of adverbial modifiers, highlighting a distinctive stylistic strategy in LLM-generated intensified texts.

In AI-generated texts that aim to make persuasion more subtle, modal verbs such as "might" (VT_MIGHT) occur more frequently than in human-written texts. Similarly, narrative framing (SY_NARRATIVE), defined as the proportion of tokens in declarative sentences relative to all tokens, is more prevalent in AI subtle rewritings. These patterns may indicate that AI softens persuasion by using modal hedges and favors neutral declarative constructions over exclamatory sentences or rhetorical questions more often than humans do.

Open-ended AI-generated texts exhibit a consistent linguistic profile characterized by high lexical diversity and elevated content-word density (e.g., L_CONT_T, L_CONT_A, LTOKEN_RATIO_LEM). AI systems also show substantially lower function-word usage and reduced lexical repetition. In addition, AI-generated texts rely more heavily on imperative and infinitival constructions while avoiding marked syntactic patterns such as inversions, which may result in structurally simpler and more schematic syntax.

Paraphrasing		*Rewriting for Subtle Effects*		*Rewriting for Intensified Effects*		*Open-ended*	
Feature Name	C_d	Feature Name	C_d	Feature Name	C_d	Feature Name	C_d
Generating Model: GPT 4.1 Mini							
L_CONT_T	0.51	L_CONT_T	0.70	L_CONT_T	0.77	L_CONT_T	1.62
L_CONT_A	0.46	L_CONT_A	0.62	L_CONT_A	0.77	L_CONT_A	1.43
SY_INV_PATTERNS	-0.45	VT_MIGHT	0.62	L_PUNCT_DASH	0.72	LTOKEN_RATIO_LEM	1.32
LTOKEN_RATIO_LEM	0.37	SY_INV_PATTERNS	-0.61	L_FUNC_A	-0.57	SENT_D_NP	1.05
L_FUNC_A	-0.36	L_PLURAL_NOUNS	0.54	LTOKEN_RATIO_LEM	0.51	ST_REPET_WORDS	-1.00
Generating Model: Llama 3.3 70B							
SENT_ST_WPERSENT	0.71	SY_INV_PATTERNS	-0.75	SENT_ST_WPERSENT	0.80	VF_INFINITIVE	1.34
SENT_ST_DIFFER	-0.70	G_ACTIVE	-0.72	L_ADJ_POSITIVE	0.72	SY_IMPERATIVE	1.24
L_PUNCT_COM	0.67	SENT_ST_WPERSENT	0.70	L_PUNCT_COM	0.71	G_ACTIVE	-0.91
SY_INV_PATTERNS	-0.61	L_CONT_T	0.62	L_CONT_T	0.66	SENT_D_VP	0.83
L_CONT_T	0.54	SENT_D_PP	0.62	POS_ADJ	0.65	L_PUNCT_DOT	-0.81
Generating Model: Gemma 3 27b it							
L_CONT_T	0.75	L_CONT_T	1.02	L_CONT_T	1.04	SY_IMPERATIVE	1.74
L_CONT_A	0.67	L_CONT_A	1.02	L_CONT_A	1.01	L_CONT_T	1.4
L_FUNC_A	-0.62	L_FUNC_A	-0.89	L_FUNC_A	-0.97	L_CONT_A	1.25
SY_INV_PATTERNS	-0.61	POS_PRO	-0.73	L_ADJ_POSITIVE	0.90	SENT_D_NP	1.16
L_ADV_SUPERLATIVE	0.43	SY_INV_PATTERNS	-0.71	POS_ADJ	0.83	PS_CAUSE	-1.15
Generating Model: Gemini 2.0 Flash							
L_CONT_A	0.73	L_CONT_A	0.92	L_CONT_A	1.02	L_CONT_T	1.57
L_CONT_T	0.71	L_CONT_T	0.92	L_CONT_T	0.99	L_CONT_A	1.37
SY_INV_PATTERNS	-0.58	L_FUNC_A	-0.73	L_FUNC_A	-0.82	SY_IMPERATIVE	1.32
L_FUNC_A	-0.54	SY_INV_PATTERNS	-0.72	L_ADJ_POSITIVE	0.78	LTOKEN_RATIO_LEM	1.20
L_PUNCT_COM	0.49	POS_NOUN	0.59	POS_ADJ	0.72	SY_EXCLAMATION	1.11

Table 7.9: Top linguistic features by absolute Cohen's d from four generation approaches for all generating LLMs.

7.6 Discussion

This chapter examined the characteristics and detectability of persuasive texts generated by large language models, using the Persuaficial corpus as an empirical foundation. By controlling generation strategies and analyzing their linguistic and computational detectability across multiple languages, this work provides insights into how persuasion is realized in AI-generated content and how it differs from human-authored persuasion.

A central finding of this study is that the detectability of persuasive content is not an inherent property of machine-generated text, but is strongly shaped by the

Stylometric Feature	Cohen's d	Sig.
GPT 4.1 Mini - Paraphrasing		
L_CONT_T	0.5054	✓
L_CONT_A	0.4590	✓
SY_INV_PATTERNS	-0.4542	✓
LTOKEN_RATIO_LEM	0.3713	✓
L_FUNC_A	-0.3619	✓
L_PUNCT_DASH	0.3378	✓
ST_REPET_WORDS	-0.3280	✓
L_PUNCT_SEMC	-0.2656	✓
L_PUNCT_COM	0.2298	✓
ASM	-0.2196	✓
VT_MIGHT	0.1995	✓
L_LINKS	-0.1954	✓
L_ADJ_POSITIVE	0.1723	✓
CDS	-0.1719	✓
PS_AGREEMENT	-0.1667	✓
L_ADV_SUPERLATIVE	0.1664	✓
L_ADV_COMPARATIVE	0.1592	✓
POS_ADV	0.1583	✓
PS_CAUSE	-0.1537	✓
PS_TIME	-0.1532	✓

Table 7.10: Top 20 linguistic features for AI-generated persuasive text with *Paraphrasing* generation approach and GPT 4.1 Mini model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
GPT 4.1 Mini - Rewriting with Subtle Persuasive Effect		
L_CONT_T	0.7036	✓
L_CONT_A	0.6222	✓
VT_MIGHT	0.6188	✓
SY_INV_PATTERNS	-0.6072	✓
L_PLURAL_NOUNS	0.5444	✓
L_FUNC_A	-0.5293	✓
SY_NARRATIVE	0.5242	✓
LTOKEN_RATIO_LEM	0.4993	✓
ST_REPET_WORDS	-0.4319	✓
POS_PRO	-0.4307	✓
L_YOU_PRON	-0.4245	✓
G_ACTIVE	-0.4232	✓
G_FUTURE	-0.4015	✓
L_SECOND_PERSON_PRON	-0.3997	✓
VT_FUTURE_SIMPLE	-0.3949	✓
CDS	-0.3837	✓
SENT_D_PP	0.3689	✓
L_PUNCT	-0.3401	✓
VT_MAY	0.3378	✓
SENT_ST_DIFFERENCE	-0.3208	✓

Table 7.11: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Subtle Persuasive Effect* generation approach and GPT 4.1 Mini model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
GPT 4.1 Mini - Rewriting with Intensified Persuasive Effect		
L_CONT_T	0.7685	✓
L_CONT_A	0.7654	✓
L_PUNCT_DASH	0.7243	✓
L_FUNC_A	-0.5702	✓
LTOKEN_RATIO_LEM	0.5075	✓
L_ADV_SUPERLATIVE	0.5033	✓
L_ADV_COMPARATIVE	0.4892	✓
SY_INV_PATTERNS	-0.4888	✓
POS_ADV	0.4766	✓
ST_REPET_WORDS	-0.4037	✓
L_ADJ_POSITIVE	0.3884	✓
POS_ADJ	0.3670	✓
ASM	-0.3241	✓
L_ADV_POSITIVE	0.3237	✓
PS_CAUSE	-0.3067	✓
L_PUNCT_COM	0.3045	✓
L_PUNCT_SEMC	-0.2626	✓
POS_PRO	-0.2415	✓
SENT_D_ADVP	0.2385	✓
L_NOUN_PHRASES	0.2333	✓

Table 7.12: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Intensified Persuasive Effect* generation approach and GPT 4.1 Mini model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
GPT 4.1 Mini - Open-ended		
L_CONT_T	1.6204	✓
L_CONT_A	1.4300	✓
LTOKEN_RATIO_LEM	1.3172	✓
SENT_D_NP	1.0532	✓
ST_REPET_WORDS	-0.9950	✓
L_PUNCT_DASH	0.9885	✓
L_FUNC_A	-0.8585	✓
SY_IMPERATIVE	0.8376	✓
POS_NOUN	0.8076	✓
POS_ADJ	0.7841	✓
L_ADJ_POSITIVE	0.7816	✓
SY_INV_PATTERNS	-0.7496	✓
L_LINKS	-0.7474	✓
VF_INFINITIVE	0.7252	✓
G_PAST	-0.6604	✓
G_ACTIVE	-0.6565	✓
L_FIRST_PERSON_SING_PRON	-0.6425	✓
L_I_PRON	-0.6425	✓
SENT_D_VP	0.6275	✓
VT_PAST_SIMPLE	-0.5774	✓

Table 7.13: Top 20 linguistic features for AI-generated persuasive text with *Open-ended* generation approach and GPT 4.1 Mini model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Llama - Paraphrasing		
SENT_ST_WPERSENT	0.7055	✓
SENT_ST_DIFFERENCE	-0.6979	✓
L_PUNCT_COM	0.6659	✓
SY_INV_PATTERNS	-0.6146	✓
L_CONT_T	0.5438	✓
L_CONT_A	0.4861	✓
G_ACTIVE	-0.4821	✓
L_PUNCT_DOT	-0.4509	✓
ASM	-0.4192	✓
POS_PRO	-0.4055	✓
FOS_FRONTING	0.3884	✓
L_ADJ_POSITIVE	0.3844	✓
L_PUNCT_SEMC	-0.3839	✓
L_YOU_PRON	-0.3761	✓
L_FUNC_A	-0.3469	✓
SY_SUBORD_SENT	0.3354	✓
L_PUNCT	-0.3339	✓
POS_PREP	0.3256	✓
VT_PRESENT_SIMPLE	-0.3229	✓
L_SECOND_PERSON_PRON	-0.3228	✓

Table 7.14: Top 20 linguistic features for AI-generated persuasive text with *Paraphrasing* generation approach and Llama 3.3 70B model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Llama - Rewriting with Subtle Persuasive Effect		
SY_INV_PATTERNS	-0.7470	✓
G_ACTIVE	-0.7152	✓
SENT_ST_WPERSENT	0.6961	✓
L_CONT_T	0.6226	✓
SENT_D_PP	0.6153	✓
L_ADJ_POSITIVE	0.6145	✓
POS_PREP	0.6086	✓
L_YOU_PRON	-0.5966	✓
POS_NOUN	0.5896	✓
L_PUNCT_COM	0.5848	✓
L_SECOND_PERSON_PRON	-0.5764	✓
POS_ADJ	0.5595	✓
SY_NARRATIVE	0.5419	✓
L_CONT_A	0.5413	✓
POS_PRO	-0.5400	✓
SENT_ST_DIFFERENCE	-0.5382	✓
SY_SUBORD_SENT	0.5373	✓
L_PUNCT	-0.5195	✓
ASM	-0.4923	✓
L_PLURAL_NOUNS	0.4688	✓

Table 7.15: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Subtle Persuasive Effect* generation approach and Llama 3.3 70B model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Llama - Rewriting with Intensified Persuasive Effect		
SENT_ST_WPERSENT	0.8019	✓
L_ADJ_POSITIVE	0.7196	✓
L_PUNCT_COM	0.7055	✓
L_CONT_T	0.6589	✓
POS_ADJ	0.6478	✓
L_CONT_A	0.6349	✓
G_ACTIVE	-0.6171	✓
SENT_ST_DIFFERENCE	-0.6110	✓
L_PUNCT_DOT	-0.5876	✓
SY_INV_PATTERNS	-0.5135	✓
ASM	-0.4697	✓
L_FUNC_A	-0.4692	✓
L_FUNC_T	-0.4456	✓
FOS_FRONTING	0.4390	✓
POS_PRO	-0.3973	✓
SENT_D_VP	0.3895	✓
L_PUNCT_SEMC	-0.3806	✓
CDS	-0.3720	✓
L_YOU_PRON	-0.3670	✓
L_NOUN_PHRASES	0.3630	✓

Table 7.16: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Intensified Persuasive Effect* generation approach and Llama 3.3 70B model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Llama - Open-ended		
VF_INFINITIVE	1.3397	✓
SY_IMPERATIVE	1.2406	✓
G_ACTIVE	-0.9149	✓
SENT_D_VP	0.8331	✓
L_PUNCT_DOT	-0.8088	✓
G_PAST	-0.7696	✓
L_OUR_PRON	0.7642	✓
L_LINKS	-0.7474	✓
VT_PAST_SIMPLE	-0.6985	✓
SY_EXCLAMATION	0.6879	✓
SY_INV_PATTERNS	-0.6857	✓
L_CONT_T	0.6780	✓
SY_COORD_SENT	0.6774	✓
L_PUNCT	-0.6493	✓
SENT_D_NP	0.6396	✓
L_FIRST_PERSON_SING_PRON	-0.6230	✓
L_I_PRON	-0.6230	✓
L_WE_PRON	0.6190	✓
L_IT_PRON	0.5548	✓
VT_MUST	0.5078	✓

Table 7.17: Top 20 linguistic features for AI-generated persuasive text with *Open-ended* generation approach and Llama 3.3 70B model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemma - Paraphrasing		
L_CONT_T	0.7526	✓
L_CONT_A	0.6659	✓
L_FUNC_A	-0.6181	✓
SY_INV_PATTERNS	-0.6133	✓
L_ADV_SUPERLATIVE	0.4274	✓
ST_REPET_WORDS	-0.4178	✓
L_ADV_COMPARATIVE	0.4141	✓
LTOKEN_RATIO_LEM	0.4081	✓
PS_CONDITION	-0.3995	✓
ASM	-0.3761	✓
CDS	-0.3606	✓
POS_ADV	0.3605	✓
L_ADJ_POSITIVE	0.3407	✓
PS_AGREEMENT	-0.3212	✓
L_PUNCT_COM	0.3172	✓
PS_CAUSE	-0.2988	✓
POS_ADJ	0.2872	✓
POS_PRO	-0.2805	✓
G_ACTIVE	-0.2710	✓
SENT_ST_WPERSENT	0.2582	✓

Table 7.18: Top 20 linguistic features for AI-generated persuasive text with *Paraphrasing* generation approach and Gemma 3 27b it model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemma - Rewriting with Subtle Persuasive Effect		
L_CONT_T	1.0245	✓
L_CONT_A	1.0176	✓
L_FUNC_A	-0.8850	✓
POS_PRO	-0.7267	✓
SY_INV_PATTERNS	-0.7119	✓
L_PLURAL_NOUNS	0.6762	✓
POS_NOUN	0.6717	✓
SENT_D_PP	0.6401	✓
L_ADJ_POSITIVE	0.6191	✓
G_ACTIVE	-0.5837	✓
L_SECOND_PERSON_PRON	-0.5783	✓
L_YOU_PRON	-0.5736	✓
SY_NARRATIVE	0.5666	✓
POS_ADJ	0.5620	✓
CDS	-0.5438	✓
L_FUNC_T	-0.4852	✓
VF_INFINITIVE	-0.4846	✓
ASM	-0.4624	✓
L_THEY_PRON	-0.4486	✓
POS_CONJ	-0.4429	✓

Table 7.19: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Subtle Persuasive Effect* generation approach and Gemma 3 27b it model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemma - Rewriting with Intensified Persuasive Effect		
L_CONT_T	1.0424	✓
L_CONT_A	1.0122	✓
L_FUNC_A	-0.9718	✓
L_ADJ_POSITIVE	0.9024	✓
POS_ADJ	0.8267	✓
PS_CONDITION	-0.6942	✓
PS_CAUSE	-0.6071	✓
L_FUNC_T	-0.6021	✓
SY_INV_PATTERNS	-0.5923	✓
G_ACTIVE	-0.5882	✓
L_NOUN_PHRASES	0.5486	✓
POS_PRO	-0.5389	✓
L_ADV_SUPERLATIVE	0.5243	✓
L_ADV_COMPARATIVE	0.5002	✓
ASM	-0.4781	✓
CDS	-0.4777	✓
L_SINGULAR_NOUNS	0.4707	✓
POS_NOUN	0.4570	✓
L_YOU_PRON	-0.4543	✓
SENT_D_NP	0.4475	✓

Table 7.20: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Intensified Persuasive Effect* generation approach and Gemma 3 27b it model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemma - Open-ended		
SY_IMPERATIVE	1.7354	✓
L_CONT_T	1.4021	✓
L_CONT_A	1.2523	✓
SENT_D_NP	1.1601	✓
PS_CAUSE	-1.1529	✓
VF_INFINITIVE	1.0451	✓
L_FUNC_A	-0.9623	✓
PS_CONDITION	-0.8951	✓
POS_NOUN	0.8730	✓
ST_REPET_WORDS	-0.8055	✓
LTOKEN_RATIO_LEM	0.7851	✓
L_LINKS	-0.7474	✓
SY_INV_PATTERNS	-0.7306	✓
L_SINGULAR_NOUNS	0.6881	✓
SY_EXCLAMATION	0.6825	✓
SENT_D_VP	0.6631	✓
POS_PREP	-0.6571	✓
L_I_PRON	-0.6461	✓
L_FIRST_PERSON_SING_PRON	-0.6461	✓
L_NOUN_PHRASES	0.6320	✓

Table 7.21: Top 20 linguistic features for AI-generated persuasive text with *Open-ended* generation approach and Gemma 3 27b it model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemini - Paraphrasing		
L_CONT_A	0.7275	✓
L_CONT_T	0.7074	✓
SY_INV_PATTERNS	-0.5802	✓
L_FUNC_A	-0.5371	✓
L_PUNCT_COM	0.4880	✓
LTOKEN_RATIO_LEM	0.4192	✓
L_ADJ_POSITIVE	0.4104	✓
ST_REPET_WORDS	-0.3981	✓
POS_ADJ	0.3677	✓
L_NOUN_PHRASES	0.3255	✓
ASM	-0.2996	✓
PS_CONDITION	-0.2850	✓
PS_CAUSE	-0.2759	✓
L_POSSESSIVES	0.2719	✓
POS_PREP	-0.2305	✓
CDS	-0.2242	✓
L_PUNCT	0.2218	✓
PS_AGREEMENT	-0.2169	✓
L_PLURAL_NOUNS	0.2053	✓
L_THEIR_PRON	0.2050	✓

Table 7.22: Top 20 linguistic features for AI-generated persuasive text with *Paraphrasing* generation approach and Gemini 2.0 Flash model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemini - Rewriting with Subtle Persuasive Effect		
L_CONT_A	0.9216	✓
L_CONT_T	0.9163	✓
L_FUNC_A	-0.7257	✓
SY_INV_PATTERNS	-0.7182	✓
POS_NOUN	0.5946	✓
G_ACTIVE	-0.5892	✓
POS_PRO	-0.5535	✓
VT_MIGHT	0.5222	✓
CDS	-0.5034	✓
SENT_D_PP	0.4976	✓
L_YOU_PRON	-0.4908	✓
LTOKEN_RATIO_LEM	0.4813	✓
L_PLURAL_NOUNS	0.4761	✓
L_ADJ_POSITIVE	0.4695	✓
L_SECOND_PERSON_PRON	-0.4555	✓
ASM	-0.4517	✓
L_PUNCT_COM	0.4347	✓
ST_REPET_WORDS	-0.4182	✓
POS_ADJ	0.4165	✓
SY_NARRATIVE	0.4079	✓

Table 7.23: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Subtle Persuasive Effect* generation approach and Gemini 2.0 Flash model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemini - Rewriting with Intensified Persuasive Effect		
L_CONT_A	1.0170	✓
L_CONT_T	0.9948	✓
L_FUNC_A	-0.8186	✓
L_ADJ_POSITIVE	0.7827	✓
POS_ADJ	0.7239	✓
L_PUNCT_COM	0.5901	✓
L_NOUN_PHRASES	0.5756	✓
SY_INV_PATTERNS	-0.5355	✓
LTOKEN_RATIO_LEM	0.4892	✓
L_ADV_SUPERLATIVE	0.4824	✓
PS_CONDITION	-0.4649	✓
L_ADV_COMPARATIVE	0.4629	✓
ASM	-0.4428	✓
ST_REPET_WORDS	-0.4386	✓
PS_CAUSE	-0.4330	✓
POS_ADV	0.4270	✓
G_ACTIVE	-0.4252	✓
POS_PRO	-0.3773	✓
L_FUNC_T	-0.3605	✓
PS_AGREEMENT	-0.3525	✓

Table 7.24: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Intensified Persuasive Effect* generation approach and Gemini 2.0 Flash model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemini - Open-ended		
L_CONT_T	1.5669	✓
L_CONT_A	1.3747	✓
SY_IMPERATIVE	1.3186	✓
LTOKEN_RATIO_LEM	1.1953	✓
SY_EXCLAMATION	1.1092	✓
VF_INFINITIVE	1.0846	✓
ST_REPET_WORDS	-1.0032	✓
SENT_D_VP	0.9291	✓
L_FUNC_A	-0.8914	✓
POS_NOUN	0.8779	✓
L_NOUN_PHRASES	0.8190	✓
PS_CAUSE	-0.7878	✓
PS_CONDITION	-0.7832	✓
G_ACTIVE	-0.7769	✓
L_LINKS	-0.7474	✓
L_SINGULAR_NOUNS	0.7331	✓
SY_INV_PATTERNS	-0.6693	✓
G_PAST	-0.6464	✓
POS_PREP	-0.6279	✓
L_I_PRON	-0.6236	✓

Table 7.25: Top 20 linguistic features for AI-generated persuasive text with *Open-ended* generation approach and Gemini 2.0 Flash model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

generation strategy used. Open-ended generation and rewriting with intensified persuasion may introduce persuasion cues that facilitate automatic detection, whereas subtle rewriting strategies significantly reduce detection performance. This suggests that modern LLMs can produce persuasive content that closely mimics human persuasion when prompted to do so in a restrained manner, thereby challenging existing detection approaches. From a broader perspective, this result underscores that future risks for automatic systems associated with AI-generated persuasion are less likely to stem from overtly manipulative content and more likely to stem from subtle forms of influence.

The linguistic analysis further clarifies how these differences manifest at the textual level. AI-generated persuasive texts exhibit higher lexical diversity and a denser distribution of content words, while human-written persuasion relies more heavily on syntactic complexity and function-word patterns. These findings indicate that LLMs tend to optimize for informational richness and stylistic fluency. Importantly, these differences are not uniform across generation strategies: intensified persuasion amplifies explicit modifiers and evaluative language, while subtle persuasion relies on modal expressions and declarative framing. This highlights that persuasion in LLM-generated text is not monolithic, but rather dynamically shaped by prompt design.

The consistency of these trends across six languages suggests that the observed effects are not language-specific artifacts but reflect more general properties of multilingual LLMs. This cross-lingual robustness strengthens Persuaficial's contribution as a resource for studying persuasion beyond English-centric settings and supports its applicability to multilingual persuasion analysis.

Within the broader context of this dissertation, the findings from Persuaficial complement and extend the investigation of persuasion-aware approaches to disinformation detection. While Chapter 5 presents that explicitly using persuasion can improve disinformation detection performance, the present analysis reveals a critical and possible challenge: persuasive cues themselves can be minimized by generative models.

Overall, this chapter advances understanding of AI-generated persuasive language by showing that its detectability depends on how persuasion is created during generation. Persuaficial thus serves not only as a benchmark dataset, but also as a methodological lens for analyzing the evolving capabilities of LLMs as persuasive agents. These insights have direct implications for the design of future detection systems, the evaluation of AI-generated influence, and the development of regulatory and educational responses to increasingly subtle forms of automated persuasion.

8 Conclusion and Future Work

This dissertation investigated how persuasion and knowledge of malicious intent can be explicitly leveraged to improve computational disinformation detection. By combining expert-annotated datasets with intent- and persuasion-augmented reasoning frameworks, the work advances both the empirical foundations and methodological approaches to intent and persuasion-aware disinformation analysis. Furthermore, the dissertation advances understanding of AI-generated persuasive texts by examining their detectability relative to human-authored persuasive content and analyzing linguistic characteristics that may also influence the performance of automatic disinformation detection systems. This chapter revisits the research objectives, summa4 rizes the main findings and contributions, and outlines promising directions for future research.

8.1 Revisiting Research Objectives: Contributions and Key Findings

Research Objective 1 [RO1]
Design a human-annotated disinformation dataset and annotation framework that captures manipulation techniques and malicious intent, going beyond binary credibility labels, to benchmark language models for detecting disinformation, manipulation techniques, and malicious intent classification.

This objective was achieved through the design and release of multiple high-quality, human-annotated disinformation datasets that conceptualize disinformation as an intentional and manipulative phenomenon. The dissertation introduced and analyzed the *MIPD*, *MultiDis*, and *MALINT* datasets, all developed

in close collaboration with professional fact-checkers and debunking experts.

MIPD constitutes the first resource in NLP research to jointly annotate articles with disinformation, manipulation techniques, and malicious intent categories. Moreover, it represents the largest expert-annotated Polish dataset of its kind, substantially advancing disinformation research for a previously under-resourced language. *MultiDis* is a high-quality, English, multitopic dataset in which annotations from all stages of the expert annotation process are released, rather than only final consensus labels. This design enables detailed analysis of annotation disagreements and the annotation process. *MALINT* further extends this line of work by substantially increasing dataset scale and introducing explicit malicious intent annotations for English disinformation, making it the first NLP English dataset to support systematic benchmarking of malicious intent classification alongside disinformation detection.

All datasets were created using rigorous, multi-stage annotation workflows involving expert training, independent parallel annotation, and final consensus resolution. The annotation methodologies and detailed guidelines for both Polish and English data were made publicly available to support transparency, reproducibility, and future dataset development. By releasing intermediate annotation stages for *MultiDis* and *MALINT*, these resources additionally enable future research on annotation quality and disagreement analysis.

Using these datasets, the dissertation established empirical benchmarks for a wide range of models, including fine-tuned discriminative Transformer-based encoders and generative large language models. Collectively, these resources constitute a new empirical foundation for disinformation research, enabling systematic and reproducible evaluation of language models on disinformation detection, manipulation analysis, and malicious intent classification. Moreover, the methodological soundness of the MIPD dataset and the accompanying benchmarking experiments was validated through peer review and acceptance at the Conference on Empirical Methods in Natural Language Processing, one of the leading venues in natural language processing.

Research Objective 2 [RO2]
Develop and evaluate a persuasion-augmented reasoning framework for large language models, enabling intermediate analysis and explanation of persuasive strategies to improve disinformation detection.

This objective was addressed through the proposal and extensive evaluation of the Persuasion-Augmented Chain of Thought framework. Grounded in psychological insights showing that awareness of persuasive fallacies enhances

resistance to misleading information [151], PCoT utilizes this principle within large language models by introducing structured, multi-stage, and knowledge-enhanced reasoning. Specifically, the model first identifies and explains the persuasive strategies present in a text, then leverages this analysis to guide disinformation classification.

Extensive experiments across multiple datasets, genres, temporal splits, and models provide strong empirical evidence that persuasion-augmented reasoning consistently improves zero-shot disinformation detection. The gains are particularly pronounced for long-form articles and for post-cutoff content unseen during model pretraining. These results indicate that persuasion knowledge functions as a meaningful intermediate representation, strengthening both generalization and robustness. The significance and broader relevance of this contribution are further underscored by its acceptance as a full paper at the main conference of the Association for Computational Linguistics, one of the top-tier venues in natural language processing. Overall, this work establishes persuasion-augmented reasoning as an effective and principled alternative to purely binary zero-shot detection approaches.

Research Objective 3 [RO3]
Develop and evaluate an intent-augmented reasoning framework for large language models that uses knowledge and explanations of malicious intent to improve disinformation detection.

This objective was fulfilled by introducing *Intent-Based Inoculation*, an intent-augmented reasoning framework inspired by the earlier PCoT approach and by inoculation theory from psychology and communication studies. The framework leverages explicit malicious intent analysis as a form of refutational preemption, enabling language models to reason about the underlying goals of disinformation creators prior to making disinformation judgments.

An empirical evaluation across six datasets, multiple genres, temporal splits, and seven languages provides strong evidence that intent-augmented reasoning yields consistent, statistically significant improvements in zero-shot disinformation detection. The results confirm that malicious intent constitutes a critical hidden dimension of disinformation that is not fully captured by factual inconsistency alone. By formally integrating intent knowledge into LLM reasoning, this dissertation provides the first large-scale evidence that intent-aware reasoning enhances both accuracy and generalization across languages. The significance of this contribution is further supported by its acceptance as a full paper at the Conference of the European Chapter of the Association for Computational

Linguistics, one of the top-tier venues in natural language processing.

Research Objective 4 [RO4]
Analyze and compare AI-generated and human-authored persuasive content based on linguistic characteristics and detection difficulty for automatic disinformation systems.

This objective was achieved through the release of the Persuaficial benchmark and a detailed comparative analysis of AI-generated and human-written persuasive texts across multiple languages and datasets. The benchmark provides a systematic foundation for studying persuasion generated by large language models and supports controlled evaluation of detection methods under diverse generation settings.

The dissertation provides strong empirical evidence that AI-generated persuasive content is linguistically distinct from human-written persuasion, exhibiting higher lexical diversity, altered function-word distributions, and different syntactic patterns. Importantly, the detectability of AI-generated persuasion was found to be highly sensitive to text-generation approaches, with subtle persuasion-oriented prompting substantially reducing detection performance.

These findings reveal that generative language models can adapt persuasive strategies in ways that challenge existing detection systems, raising important methodological and societal implications. By characterizing how generation strategies and model choices influence detectability, this analysis contributes to a deeper understanding of the limitations of current detection approaches.

Summary
In summary, this dissertation makes three principal contributions to computational disinformation research. First, it establishes new human-annotated resources that introduce disinformation as an intentional and persuasive phenomenon. Second, it introduces and validates persuasion- and intent-augmented reasoning frameworks that significantly improve zero-shot disinformation detection across models, datasets, genres, and temporal splits. Third, it advances the understanding of AI-generated persuasive content and its implications for detection, revealing both new challenges and opportunities for future systems.

Together, these contributions present that effective automatic disinformation detection requires reasoning about *why* and *how* content is constructed, not merely *whether* it is factually correct.

8.2 Future Research Directions

This dissertation opens several promising directions for future research.

First, extending the presented datasets is a natural and necessary step. In particular, enriching the *MALINT* dataset with annotations of manipulation techniques would enable joint research of disinformation, malicious intent, and presuasive strategies in English. Such a unified annotation framework would support deeper analysis of the interactions among intent, persuasion, and disinformation, facilitate the development of more comprehensive multi-task learning approaches, and may provide a basis for research on explainable disinformation detection.

Second, extending persuasion-augmented reasoning to multilingual and low-resource settings remains an important research direction. While this dissertation presents the effectiveness of intent-augmented reasoning across multiple languages, future work should systematically investigate language-specific gains of persuasion-augmented reasoning. Moreover, an important open question concerns the integration of persuasion- and intent-augmented reasoning within a single unified framework. In this dissertation, these two paradigms were explored independently. Combining persuasion- and intent-augmented reasoning may yield complementary benefits and further improve disinformation detection performance.

Third, future research could explore dynamic and adaptive reasoning frameworks in which persuasion strategies and malicious intent categories are selected, weighted, or prioritized based on contextual factors such as topic, genre, language, or model uncertainty. Such adaptive reasoning mechanisms could increase robustness against evolving and increasingly sophisticated disinformation tactics.

Fourth, integrating multimodal signals represents a natural extension of persuasion and intent-augmented reasoning. Disinformation rarely operates in text-only environments, and future work should consider combining textual reasoning with visual features. Incorporating images could substantially enhance both the practical usefulness and the overall effectiveness of disinformation detection systems.

Fifth, the interaction between human cognition and model reasoning warrants deeper investigation. Future studies could explore human–AI collaboration scenarios in which persuasion- and intent-based explanations generated by language models support fact-checkers, journalists, educators, or policy analysts in real-world decision-making. Evaluating such systems in applied settings would provide valuable insights into trust and explainability.

Finally, as generative language models continue to evolve, future research could examine how advances in controllable and strategic text generation influence both the production and detection of persuasive disinformation. Understanding the co-evolution of generation and detection mechanisms will be essential for developing robust, socially responsible, and future-proof disinformation detection systems.

References

[1] Kai Shu, Amrita Bhattacharjee, Faisal Alatawi, Tahora H Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu. Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1385, 2020.

[2] OECD. *Facts Not Fakes: Tackling Disinformation, Strengthening Information Integrity*. OECD Publishing, 2024.

[3] Madeleine de Cock Buning. *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. Publications Office of the European Union, 2018.

[4] Michael Hameleers. Disinformation as a context-bound phenomenon: Toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination. *Communication Theory*, 33(1):1–10, 2023.

[5] Eleni Kapantai, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New media & society*, 23(5):1301–1326, 2021.

[6] Witold Sosnowski, Arkadiusz Modzelewski, Kinga Skorupska, Jahna Otterbacher, and Adam Wierzbicki. Eu disinfotest: a benchmark for evaluating language models' ability to detect disinformation narratives. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14702–14723, 2024.

[7] Witold Sosnowski, Arkadiusz Modzelewski, Kinga Skorupska, and Adam Wierzbicki. DiNaM: Disinformation narrative mining with large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30212–30239, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1537. URL https://aclanthology.org/2025.emnlp-main.1537/.

[8] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788, 2021.

[9] Maryam Heidari, Samira Zad, Parisa Hajibabaee, Masoud Malekzadeh, SeyyedPooya HekmatiAthar, Ozlem Uzuner, and James H Jones. Bert model for fake news detection based on social bot activities in the covid-19 pandemic. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0103–0109. IEEE, 2021.

[10] Junaed Younus Khan, Md Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4: 100032, 2021.

[11] Wei Peng, Sue Lim, and Jingbo Meng. Persuasive strategies in online health misinformation: a systematic review. *Information, Communication & Society*, 26(11):2131–2148, 2023.

[12] Sijing Chen, Lu Xiao, and Jin Mao. Persuasion strategies of misinformation-containing posts in the social media. *Information Processing & Management*, 58(5):102665, 2021.

[13] Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. In *2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 14279–14305. Association for Computational Linguistics (ACL), 2023.

[14] Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. Persuading across diverse domains: a dataset and persuasion large language model. In Lun-Wei Ku, Andre Martins, and

Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.92. URL https://aclanthology.org/2024.acl-long.92/.

[15] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. How persuasive is ai-generated propaganda? *PNAS nexus*, 3(2): pgae034, 2024.

[16] Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. Persuasion with large language models: a survey. *arXiv preprint arXiv:2411.06837*, 2024.

[17] Arkadiusz Modzelewski, Giovanni Da San Martino, Pavel Savov, Magdalena Anna Wilczyńska, and Adam Wierzbicki. Mipd: Exploring manipulation and intention in a novel corpus of polish disinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19769–19785, 2024.

[18] Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. Pcot: Persuasion-augmented chain of thought for detecting fake news and social media disinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24959–24983, 2025.

[19] Arkadiusz Modzelewski, Witold Sosnowski, Eleni Papadopulos, Elisa Sartori, Tiziano Labruna, Giovanni Da San Martino, and Adam Wierzbicki. Malicious intent dataset and inoculating LLMs for enhanced disinformation detection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1 of *EACL 2026*, Rabat, Morocco, 2026. Association for Computational Linguistics. Accepted, to appear.

[20] Arkadiusz Modzelewski, Paweł Golik, Anna Kołos, and Giovanni Da San Martino. Can AI-Generated Persuasion Be Detected? Persuaficial Benchmark and AI vs. Human Linguistic Differences. *arXiv preprint arXiv:2601.04925*, 2026.

[21] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition, 2025. URL

https://web.stanford.edu/~jurafsky/slp3/. Online manuscript released August 24, 2025.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[23] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 5368–5393, 2023.

[24] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[25] Christopher D Manning. Human language understanding & reasoning. *Daedalus*, 151(2):127–138, 2022.

[26] Adam Wierzbicki. *Web content credibility*, volume 10. Springer, 2018.

[27] Shawn Tseng and Brian J Fogg. Credibility and computing technology. *Communications of the ACM*, 42(5):39–44, 1999.

[28] Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, et al. News categorization, framing and persuasion techniques: Annotation guidelines. *European Commission, Ispra, JRC132862*, 2023.

[29] Scott C Paine. Persuasion, manipulation, and dimension. *The Journal of Politics*, 51(1):36–49, 1989.

[30] Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.169. URL https://aclanthology.org/2023.acl-long.169/.

[31] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.317. URL https://aclanthology.org/2023.semeval-1.317/.

[32] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, 2023.

[33] Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, 2024.

[34] Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michał Marcińczuk, Arkadiusz Modzelewski, Ivo Moravski, et al. Slavicnlp 2025 shared task: Detection and classification of persuasion techniques in parliamentary debates and social media. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 254–275, 2025.

[35] Terry Flew. Digital communication, the crisis of trust, and the post-global. *Communication research and practice*, 5(1):4–22, 2019.

[36] Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. Fake news on social media: the impact on society. *Information Systems Frontiers*, 26(2):443–458, 2024.

[37] Petros Iosifidis and Nicholas Nicoli. *Digital democracy, social media and disinformation*. Routledge, 2020.

[38] Bertin Martens, Luis Aguiar, Estrella Gomez-Herrera, and Frank Mueller-Langer. The digital transformation of news media and the rise of disinformation and fake news. 2018.

[39] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[40] Elena Broda and Jesper Strömbäck. Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, 48(2):139–166, 2024.

[41] Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Francesco David Nota. Content based fake news detection with machine and deep learning: a systematic review. *Neurocomputing*, 2023.

[42] Shubhangi Rastogi and Divya Bansal. A review on fake news detection 3t's: Typology, time of detection, taxonomies. *International Journal of Information Security*, 22(1):177–212, 2023.

[43] Medeswara Rao Kondamudi, Somya Ranjan Sahoo, Lokesh Chouhan, and Nandakishor Yadav. A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches. *Journal of King Saud University-Computer and Information Sciences*, 35(6):101571, 2023.

[44] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30, 2023.

[45] Sebastian Kula, Michał Choraś, and Rafał Kozik. Application of the bert-based architecture in fake news detection. In *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) 12*, pages 239–249. Springer, 2021.

[46] Sebastian Kula, Rafał Kozik, and Michał Choraś. Implementation of the bert-derived architectures to tackle disinformation challenges. *Neural Computing and Applications*, pages 1–13, 2021.

[47] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[48] Tsung-Hsuan Pan, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Enhancing society-undermining disinformation detection through fine-grained sentiment analysis pre-finetuning. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1371–1377, 2024.

[49] Nida Aslam, Irfan Ullah Khan, Farah Salem Alotaibi, Lama Abdulaziz Aldaej, and Asma Khaled Aldubaikil. Fake detect: A deep learning ensemble model for fake news detection. *complexity*, 2021(1):5557784, 2021.

[50] Abdullah Marish Ali, Fuad A Ghaleb, Bander Ali Saleh Al-Rimy, Fawaz Jaber Alsolami, and Asif Irshad Khan. Deep ensemble fake news detection model using sequential deep learning technique. *Sensors*, 22(18): 6970, 2022.

[51] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1165–1174, 2020.

[52] Ehtesham Hashmi, Sule Yildirim Yayilgan, Muhammad Mudassar Yamin, Subhan Ali, and Mohamed Abomhara. Advancing fake news detection: hybrid deep learning with fasttext and explainable ai. *IEEE Access*, 2024.

[53] Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. Explainable machine learning for fake news detection. In *Proceedings of the 10th ACM conference on web science*, pages 17–26, 2019.

[54] Barry Cartwright, Richard Frank, George Weir, and Karmvir Padda. Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks. *Neural Computing and Applications*, 34(18):15141–15163, 2022.

[55] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.

[56] Canyu Chen and Kai Shu. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ccxD4mtkTU.

[57] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, 2017.

[58] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy

Chakraborty. Fighting an infodemic: Covid-19 fake news dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, pages 21–29. Springer, 2021.

[59] Aleksandra Nabożny, Bartłomiej Balcerzak, Mikołaj Morzy, Adam Wierzbicki, Pavel Savov, and Kamil Warpechowski. Improving medical experts' efficiency of misinformation detection: an exploratory study. *World Wide Web*, 26(2):773–798, 2023.

[60] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

[61] Sonish Sivarajkumar and Yanshan Wang. Healthprompt: a zero-shot learning paradigm for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2022, page 972, 2023.

[62] Maryan Rizinski, Andrej Jankov, Vignesh Sankaradas, Eugene Pinsky, Igor Miskovski, and Dimitar Trajanov. Company classification using zero-shot learning. *arXiv preprint arXiv:2305.01028*, 2023.

[63] Puneet Kumar, Kshitij Pathania, and Balasubramanian Raman. Zero-shot learning based cross-lingual sentiment analysis for sanskrit text with insufficient labeled data. *Applied Intelligence*, 53(9):10096–10113, 2023.

[64] Silvia Casola, Tiziano Labruna, Alberto Lavelli, Bernardo Magnini, et al. Testing chatgpt for stability and reasoning: A case study using italian medical specialty tests. In *CLiC-it*, 2023.

[65] Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429, 2023.

[66] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint*

Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 675–718, 2023.

[67] Fuad Mire Hassan and Mark Lee. Political fake statement detection via multistage feature-assisted neural modeling. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE, 2020.

[68] Hai-Long Nguyen, Thi-Kieu-Trang Pham, Thai-Son Le, Tan-Minh Nguyen, Thi-Hai-Yen Vuong, and Ha-Thanh Nguyen. Rmdm: A multilabel fakenews dataset for vietnamese evidence verification. *arXiv preprint arXiv:2309.09071*, 2023.

[69] Elena Musi and Chris Reed. From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media. *Discourse & Society*, 33(3):349–370, 2022.

[70] Katrina J Ward, Hamilton Link, Kiril Avramov, and Jean Goodwin. Identifying disinformation using rhetorical devices in natural language models. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2022.

[71] Danial Kamali, Joseph Romain, Huiyi Liu, Wei Peng, Jingbo Meng, and Parisa Kordjamshidi. Using persuasive writing strategies to explain and detect health misinformation. *arXiv preprint arXiv:2211.05985*, 2022.

[72] Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, 2023.

[73] Trevor Diehl, Brian E Weeks, and Homero Gil de Zúñiga. Political persuasion on social media: Tracing direct and indirect effects of news use and social interaction. *New media & society*, 18(9):1875–1895, 2016.

[74] Brian E Weeks, Alberto Ardèvol-Abreu, and Homero Gil de Zúñiga. Online influence? social media use, opinion leadership, and political persuasion. *International journal of public opinion research*, 29(2):214–239, 2017.

[75] Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. The persuasive power of

large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 152–163, 2024.

[76] Shirley Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. How far can we extract diverse perspectives from large language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5336–5366, 2024.

[77] Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. Can language models recognize convincing arguments? *arXiv preprint arXiv:2404.00750*, 2024.

[78] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/672. URL https://doi.org/10.24963/ijcai.2020/672. Survey track.

[79] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. Logical fallacy detection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.532. URL https://aclanthology.org/2022.findings-emnlp.532/.

[80] Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. Argotario: Computational argumentation meets serious games. In Lucia Specia, Matt Post, and Michael Paul, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-2002. URL https://aclanthology.org/D17-2002/.

[81] Ivan Habernal, Patrick Pauli, and Iryna Gurevych. Adapting serious game for fallacious argumentation to german: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[82] Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. Multitask instruction-based prompting for fallacy recognition. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8172–8187, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.560. URL https://aclanthology.org/2022.emnlp-main.560/.

[83] E. Musi, M. Aloumpi, E. Carmi, S. Yates, and K. O'Halloran. Developing fake news immunity: fallacies as misinformation triggers during the pandemic. *Online Journal of Communication and Media Technologies*, 12(3), July 2022. ISSN 1986-3497. doi: 10.30935/ojcmt/12083. URL https://openaccess.city.ac.uk/id/eprint/28149/. Copyright © 2022 by authors; licensee OJCMT by Bastas, CY. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/).

[84] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1317. URL https://aclanthology.org/D17-1317.

[85] Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864, 2019.

[86] Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, et al. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646. Association for Computational Linguistics, 2019.

[87] Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeno, and Preslav Nakov. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, 2020.

[88] Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. Transformers:"the end of history" for natural language processing? In *Joint European conference on machine learning and knowledge discovery in databases*, pages 677–693. Springer, 2021.

[89] Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. Interpretable propaganda detection in news articles. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1597–1605, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.179/.

[90] Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.186. URL https://aclanthology.org/2020.semeval-1.186/.

[91] João Augusto Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. Weakly supervised veracity classification with llm-predicted credibility signals. *PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-5174770/v1]*, 2024.

[92] Kyle Hamilton, Luca Longo, and Bojan Bozic. Gpt assisted annotation of rhetorical and linguistic features for interpretable propaganda technique detection in news text. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1431–1440, 2024.

[93] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. Detecting propaganda techniques in memes. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 6603–6617, 2021.

[94] Muhammad Umar Salman, Asif Hanif, Shady Shehata, and Preslav Nakov. Detecting propaganda techniques in code-switched social media text. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*,

pages 16794–16812, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1044. URL https://aclanthology.org/2023.emnlp-main.1044/.

[95] Kristina Hristakieva, Stefano Cresci, Giovanni Da San Martino, Mauro Conti, and Preslav Nakov. The spread of propaganda by coordinated communities on social media. In *Proceedings of the 14th ACM web science conference 2022*, pages 191–201, 2022.

[96] Daniel Baleato Rodríguez, Verna Dankers, Preslav Nakov, and Ekaterina Shutova. Paper bullets: Modeling propaganda with the help of metaphor. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 472–489, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.35. URL https://aclanthology.org/2023.findings-eacl.35/.

[97] Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.815. URL https://aclanthology.org/2023.acl-long.815/.

[98] Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 997–1009, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.113/.

[99] Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. A second pandemic? analysis of fake news about COVID-19 vaccines in Qatar. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1010–1021, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.114/.

[100] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational

propaganda detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4826–4832, 2021.

[101] Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In Anna Feldman, Giovanni Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov, editors, *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5024. URL https://aclanthology.org/D19-5024/.

[102] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu, editors, *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.7. URL https://aclanthology.org/2021.semeval-1.7/.

[103] Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In Houda Bouamor, Hend Al-Khalifa, Kareem Darwish, Owen Rambow, Fethi Bougares, Ahmed Abdelali, Nadi Tomeh, Salam Khalifa, and Wajdi Zaghouani, editors, *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wanlp-1.11. URL https://aclanthology.org/2022.wanlp-1.11/.

[104] Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michal Marcinczuk, Arkadiusz Modzelewski, Ivo Moravski, and Roman Yangarber. SlavicNLP 2025 shared task: Detection and classification of persuasion techniques in parliamentary debates and social media. In Jakub Piskorski, Pavel Přibáň, Preslav Nakov, Roman Yangarber, and Michal Marcinczuk, editors, *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic*

NLP 2025), pages 254–275, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 978-1-959429-57-9. doi: 10.18653/v1/2025.bsnlp-1.27. URL https://aclanthology.org/2025.bsnlp-1.27/.

[105] Sebastian Duerr and Peter A Gloor. Persuasive natural language generation–a literature review. *arXiv preprint arXiv:2101.05786*, 2021.

[106] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1566. URL https://aclanthology.org/P19-1566/.

[107] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. Working with ai to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–29, 2023.

[108] Hui Bai, Jan G Voelkel, Shane Muldowney, Johannes C Eichstaedt, and Robb Willer. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037, 2025.

[109] Philipp Schoenegger, Francesco Salvi, Jiacheng Liu, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, Gabriel Recchia, Fritz Günther, Ali Zarifhonarvar, et al. Large language models are more persuasive than incentivized human persuaders. *arXiv preprint arXiv:2505.09662*, 2025.

[110] Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. Measuring and benchmarking large language models' capabilities to generate persuasive language. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10056–10075, 2025.

[111] Hua Xu, Hanlei Zhang, and Ting-En Lin. *Intent Recognition*, pages 7–29. Springer Nature Singapore, Singapore, 2023. ISBN 978-981-99-3885-8. doi: 10.1007/978-981-99-3885-8_2. URL https://doi.org/10.1007/978-981-99-3885-8_2.

[112] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document

intent in Instagram posts. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1469. URL https://aclanthology.org/D19-1469/.

[113] Adyasha Maharana, Quan Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, and Mohit Bansal. Multimodal intent discovery from livestream videos. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 476–489, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.36. URL https://aclanthology.org/2022.findings-naacl.36/.

[114] Mark Conner and Paul Norman. Understanding the intention-behavior gap: The role of intention strength. *Frontiers in Psychology*, 13, 2022. doi: 10.3389/fpsyg.2022.923464.

[115] Antje von Suchodoletz and Anja Achtziger. Intentions and their limits. *Social Psychology*, 42:85–92, 01 2011. doi: 10.1027/1864-9335/a000046.

[116] Ankur Gupta, Yash Varun, Prarthana Das, Nithya Muttineni, Parth Srivastava, Hamim Zafar, Tanmoy Chakraborty, and Swaprava Nath. Truthbot: An automated conversational tool for intent learning, curated information presenting, and fake news alerting, 2021. URL https://arxiv.org/abs/2102.00509.

[117] Xinyi Zhou, Kai Shu, Vir V. Phoha, Huan Liu, and Reza Zafarani. "this is fake! shared it by mistake":assessing the intent of fake news spreaders. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3685–3694. ACM, April 2022. doi: 10.1145/3485447.3512264. URL http://dx.doi.org/10.1145/3485447.3512264.

[118] Bing Wang, Ximing Li, Changchun Li, Bo Fu, Songwen Pei, and Shengsheng Wang. Why misinformation is created? detecting them by integrating intent features. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2304–2314, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679799. URL https://doi.org/10.1145/3627673.3679799.

[119] Zhengjia Wang, Danding Wang, Qiang Sheng, Juan Cao, Silong Su, Yifan Sun, Beizhe Hu, and Siyuan Ma. Understanding news creation intents: Frame, dataset, and method. *arXiv preprint arXiv:2312.16490*, 2023.

[120] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42, 2019.

[121] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937, 2017.

[122] Xinyi Zhou, Kai Shu, Vir V Phoha, Huan Liu, and Reza Zafarani. "this is fake! shared it by mistake": Assessing the intent of fake news spreaders. In *Proceedings of the ACM Web Conference 2022*, pages 3685–3694, 2022.

[123] Zhen Guo, Qi Zhang, Xinwei An, Qisheng Zhang, Audun Josang, Lance M Kaplan, Feng Chen, Dong H Jeong, and Jin-Hee Cho. Uncertainty-aware reward-based deep reinforcement learning for intent analysis of social media information. In *1st AAAI Workshop on Uncertainty Reasoning and Quantification in Decision Making (UDM-AAAI'23)*, 2023.

[124] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Companion proceedings of the 2019 world wide web conference*, pages 218–226, 2019.

[125] Preslav Nakov and Giovanni Da San Martino. Fake news, disinformation, propaganda, media bias, and flattening the curve of the covid-19 infodemic. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4054–4055, 2021.

[126] Fan Xu, Victor S Sheng, and Mingwen Wang. A unified perspective for disinformation detection and truth discovery in social sensing: a survey. *ACM Computing Surveys (CSUR)*, 55(1):1–33, 2021.

[127] Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. Support or refute: Analyzing the stance of evidence to detect out-of-context mis-and disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4268–4280, 2023.

[128] Simon Springer and Vural Özdemir. Disinformation as covid-19's twin pandemic: False equivalences, entrenched epistemologies, and causes-of-causes. *OMICS: A Journal of Integrative Biology*, 26(2):82–87, 2022.

[129] Sascha-Dominik Dov Bachmann, Dries Putter, and Guy Duczynski. Hybrid warfare and disinformation: A ukraine war perspective. *Global Policy*, 14(5):858–869, 2023.

[130] Victor Ginsburgh, Juan D Moreno-Ternero, and Shlomo Weber. Ranking languages in the european union: Before and after brexit. *European Economic Review*, 93:139–151, 2017.

[131] Aleksandra Kuczyńska-Zonik. Propaganda, disinformation, strategic communication - how to improve cooperation in cee region? *Bulletin of Lviv Polytechnic National University*, 4:160–164, 2020.

[132] Filip Bryjka. Russian disinformation regarding the attack on ukraine. PISM Polski Instytut Spraw Międzynarodowych, 2022.

[133] M.G Sessa. Connecting the disinformation dots: insights, lessons, and guidance from 20 eu member states. https://www.disinfo.eu/publications/connecting-the-disinformation-dots/, December 2023.

[134] Janice M Morse. "cherry picking": Writing from thin data, 2010.

[135] Matthew S McGlone. Quoted out of context: Contextomy and its consequences. *Journal of Communication*, 55(2):330–346, 2005.

[136] Sean Cubitt. Anecdotal evidence. *NECSUS. European Journal of Media Studies*, 2(1):5–18, 2013.

[137] Adrian Little and Juliet Brough Rogers. The politics of 'whataboutery': The problem of trauma trumping the political in conflictual societies. *The British Journal of Politics and International Relations*, 19(1):172–187, 2017.

[138] Robert Talisse and Scott F Aikin. Two forms of the straw man. *Argumentation*, 20:345–352, 2006.

[139] Elizabeth F Loftus. Leading questions and the eyewitness report. *Cognitive psychology*, 7(4):560–572, 1975.

[140] Alan Brinton. Pathos and the" appeal to emotion": An aristotelian analysis. *History of Philosophy Quarterly*, 5(3):207–219, 1988.

[141] Luca Castagnoli. Aristotle on the non-cause fallacy. *History and Philosophy of Logic*, 37(1):9–32, 2016.

[142] AJ Kreider. Argumentative hyperbole as fallacy. *Informal Logic*, 42(2): 417–437, 2022.

[143] Pascal Diethelm and Martin McKee. Denialism: what is it and how should scientists respond? *The European Journal of Public Health*, 19(1):2–4, 2009.

[144] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19, 2015.

[145] Steven E Stemler. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation*, 9(1):4, 2019.

[146] Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. HerBERT: Efficiently pretrained transformer-based language model for Polish. In Bogdan Babych, Olga Kanishcheva, Preslav Nakov, Jakub Piskorski, Lidia Pivovarova, Vasyl Starko, Josef Steinberger, Roman Yangarber, Michał Marcińczuk, Senja Pollak, Pavel Přibáň, and Marko Robnik-Šikonja, editors, *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.bsnlp-1.1/.

[147] Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. Pre-training polish transformer-based language models at scale. In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19*, pages 301–314. Springer, 2020.

[148] Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. Klej: Comprehensive benchmark for polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, 2020.

[149] Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. Crowdsourced fact-checking at twitter: How does the crowd compare with experts? In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 1736–1746, 2022.

[150] Jennifer Allen, Cameron Martel, and David G Rand. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in

twitter's birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19, 2022.

[151] Timon MJ Hruschka and Markus Appel. Learning about informal fallacies and the detection of fake news: An experimental intervention. *PLoS One*, 18(3):e0283238, 2023.

[152] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*, 2020.

[153] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9, 2018.

[154] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer, 2017.

[155] William Scott Paka, Rachit Bansal, Abhay Kaushik, Shubhashis Sengupta, and Tanmoy Chakraborty. Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection. *Applied Soft Computing*, 107:107393, 2021.

[156] Rachit Bansal, William Scott Paka, Nidhi, Shubhashis Sengupta, and Tanmoy Chakraborty. Combining exogenous and endogenous signals with a semi-supervised co-attention network for early detection of covid-19 fake tweets. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 188–200. Springer, 2021.

[157] James Pamment and Anneli Lindvall Kimber. *Fact-checking and debunking: a best practice guide to dealing with disinformation*. NATO Strategic Communication Centre of Excellence, 2021.

[158] João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. Euvsdisinfo: a dataset for multilingual detection of pro-kremlin disinformation in news articles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5380–5384, 2024.

[159] Adrien Barbaresi. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint*

Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 122–131. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.acl-demo.15.

[160] Selenium. *Selenium Browser Automation*. Selenium, 2024. URL https://www.selenium.dev. Accessed: 2024-09-11.

[161] Leonard Richardson and Jeremy Katz. Beautiful soup 4, 2024. URL https://www.crummy.com/software/BeautifulSoup/. A library for scraping information from web pages.

[162] Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. Using natural language explanations to improve robustness of in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13477–13499, 2024.

[163] Rakesh R Menon, Sayan Ghosh, and Shashank Srivastava. Clues: A benchmark for learning classifiers using natural language explanations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6523–6546, 2022.

[164] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1383–1392, 2018.

[165] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[166] Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, 2020.

[167] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128, 2006.

[168] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[169] Wei Peng, Jingbo Meng, and Barikisu Issaka. Navigating persuasive strategies in online health misinformation: an interview study with older adults on misinformation management. *Plos one*, 19(7):e0307771, 2024.

[170] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578, 2024.

[171] Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. Rephrase and respond: Let large language models ask better questions for themselves, 2023.

[172] Cecilie S Traberg, Jon Roozenbeek, and Sander van der Linden. Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, 700(1):136–151, 2022.

[173] Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *The Harvard Kennedy School (HKS) Misinformation Review*, 2020.

[174] Naomi Appelman, Stephan Dreyer, Pranav Manjesh Bidare, and Keno C Potthast. Truth, intention and harm: Conceptual challenges for disinformation-targeted governance. *Internet Policy Review*, 2022.

[175] Bing Wang, Ximing Li, Changchun Li, Bo Fu, Songwen Pei, and Shengsheng Wang. Why misinformation is created? detecting them by integrating intent features. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2304–2314, 2024.

[176] Preslav Nakov, Jisun An, Haewoon Kwak, Muhammad Arslan Manzoor, Zain Muhammad Mujahid, and Husrev T Sencar. A survey on predicting the factuality and the bias of news media. In *ACL (Findings)*, 2024.

[177] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

[178] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[179] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

[180] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021.

[181] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[182] Kevin P Murphy. Machine learning: A probabilistic perspective (adaptive computation and machine learning series). *The MIT Press: London, UK*, 2018.

[183] William J McGuire. Inducing resistance to persuasion. some contemporary approaches. *CC Haaland and WO Kaelber (Eds.), Self and Society. An Anthology of Readings, Lexington, Mass.(Ginn Custom Publishing) 1981, pp. 192-230.*, 1964.

[184] Stephan Lewandowsky, Ullrich KH Ecker, and John Cook. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition*, 6(4):353–369, 2017.

[185] Michael Pfau, Bobi Ivanov, Brian Houston, Michel Haigh, Jeanetta Sims, Eileen Gilchrist, Jason Russell, Shelley Wigley, Jackie Eckstein, and Natalie Richert. Inoculation and mental processing: The instrumental role of associative networks in the process of resistance to counterattitudinal influence. *Communication Monographs*, 72(4):414–441, 2005.

[186] Robert H Gass and John S Seiter. *Persuasion: Social influence and compliance gaining*. Routledge, 2022.

[187] Gundars Bergmanis-Korāts, Tetiana Haiduchyk, and Artur Shevtsov. Ai in precision persuasion: Unveiling tactics and risks on social media. Technical report, NATO Strategic Communications Centre of Excellence, Riga,

Latvia, August 2024. Prepared and published by the NATO Strategic Communications Centre of Excellence, 51 pp.

[188] Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*, 2023.

[189] Aleksey Panasyuk. Synthclassify: an llm-driven framework for generating and classifying persuasive text. In *Disruptive Technologies in Information Sciences IX*, volume 13480, pages 120–148. SPIE, 2025.

[190] Matthew Burtell and Thomas Woodside. Artificial influence: An analysis of ai-driven persuasion. *arXiv preprint arXiv:2303.08721*, 2023.

[191] Seliem El-Sayed, Canfer Akbulut, Amanda McCroskery, Geoff Keeling, Zachary Kenton, Zaria Jalan, Nahema Marchal, Arianna Manzini, Toby Shevlane, Shannon Vallor, et al. A mechanism-based approach to mitigating harms from persuasive generative ai. *arXiv preprint arXiv:2404.15058*, 2024.

[192] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624, 2016.

[193] Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de Albornoz, and Iván Gonzalo-Verdugo. Overview of dipromats 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers. *Procesamiento del lenguaje natural*, 71:397–407, 2023. ISSN 1135-5948.

[194] Inez Okulska, Daria Stetsenko, Anna Kołos, Agnieszka Karlińska, Kinga Głąbińska, and Adam Nowakowski. Stylometrix: An open-source multilingual tool for representing stylometric vectors. *arXiv preprint arXiv:2309.12810*, 2023.

[195] Arkadiusz Modzelewski, Witold Sosnowski, Magdalena Wilczynska, and Adam Wierzbicki. DSHacker at SemEval-2023 task 3: Genres and persuasion techniques detection with multilingual data augmentation through machine translation and text generation. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada, July 2023. Association for

Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.218. URL https://aclanthology.org/2023.semeval-1.218/.

[196] Arkadiusz Modzelewski, Paweł Golik, and Adam Wierzbicki. Bilingual propaganda detection in diplomats' tweets using language models and linguistic features. *IberLEF@ SEPLN*, 2024.

[197] Alberto Barrón-Cedeño, Firoj Alam, Julia Maria Struß, Preslav Nakov, Tanmoy Chakraborty, Tamer Elsayed, Piotr Przybyła, Tommaso Caselli, Giovanni Da San Martino, Fatima Haouari, et al. Overview of the clef-2024 checkthat! lab: check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–52. Springer, 2024.

[198] Pablo Moral, Jesús M Fraile, Guillermo Marco, Anselmo Peñas, and Julio Gonzalo. Overview of dipromats 2024: Detection, characterization and tracking of propaganda in messages from diplomats and authorities of world powers. *Procesamiento del lenguaje natural*, 73:347–358, 2024.

[199] Luis Chiruzzo, Salud María Jiménez-Zafra, and Francisco Rangel. Overview of iberlef 2024: natural language processing challenges for spanish and other iberian languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS. org*, 2024.

[200] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024.

[201] Zhongyu Wei, Yang Liu, and Yi Li. Is this post persuasive? ranking argumentative comments in online forum. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2032. URL https://aclanthology.org/P16-2032/.

[202] Subhabrata Dutta, Dipankar Das, and Tanmoy Chakraborty. Changing views: Persuasion modeling and argument extraction from online discussions. *Information Processing & Management*, 57(2):102085, 2020.

[203] Ivan Srba, Olesya Razuvayevskaya, João Augusto Leite, Róbert Móro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio

Lottmann, Denis Teyssou, Valentin Porcellini, et al. A survey on automatic credibility assessment of textual credibility signals in the era of large language models. *CoRR*, 2024.

[204] Bruce B Frey. *The SAGE encyclopedia of research design*. Sage Publications, 2021.

[205] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

[206] Shahadat Uddin and Haohui Lu. Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. *Plos one*, 19(4):e0301541, 2024.

[207] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[208] Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. Better than average: Paired evaluation of nlp systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2301–2315, 2021.

[209] Regina Stodden and Laura Kallmeyer. A multi-lingual and cross-domain analysis of features for text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 77–84, 2020.

[210] Saniya Karwa and Navpreet Singh. Disentangling linguistic features with dimension-wise analysis of vector embeddings. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 461–488, 2025.

[211] Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. SocialEval: Evaluating social intelligence of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30958–31012, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1496. URL https://aclanthology.org/2025.acl-long.1496/.

[212] Cristiano Ciaccio, Alessio Miaschi, and Felice Dell'Orletta. Evaluating lexical proficiency in neural language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1267–1286, 2025.

[213] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[214] European Commission. The strengthened code of practice on disinformation 2022. *Publications Office of the European Union*, 2022. Available at: https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

[215] European Commission. Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions on the european democracy action plan. *Publications Office of the European Union*, 2022. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0790.

A Appendix

A.1 Manipulation Techniques Taxonomy for MIPD

While our dataset is in Polish and the annotators are native Polish speakers, we have developed our annotation guidelines and methodology in English. This approach facilitates the application of our methodology to various languages. Consequently, our annotation guidelines feature explanations of manipulation techniques and examples in English. Below, we provide some of these examples, along with their explanations.

Cherry Picking. Presenting information using only data that supports a given thesis while ignoring the broader context. It may include the slothful induction (rejecting inconvenient evidence that challenges our beliefs) or the Texas sharp-shooter error (ignoring differences and emphasizing similarities, using from among an extensive dataset a small slice that supports our thesis).

Quote Mining. Using a short excerpt from someone's longer speech/text in a way that significantly distorts its actual, original meaning.

Anecdote. The use of evidence in the form of personal experience or an isolated case, possibly rumor or hearsay, most often to discredit statistics.

Whataboutism. Responding to a substantive argument not by addressing the heart of the matter but by raising a new point unrelated to the topic. Often referred to as dropping a false lead to divert attention from the topic.

Strawman. Misrepresenting someone's argument in a way that makes it easier to refute. It usually boils down to attributing to an opponent a position the opponent does not share.

Leading Questions. Flooding the target audience with consecutive questions or false arguments/studies that are suggestive. Guiding the recipient to a preconceived thesis. A statement consisting of a plethora of poorly related information, half-truths, and misinterpretations designed to overwhelm by their sheer volume.

Appeal to Emotion. The use of words and phrases arouses in the recipient a strong emotion and attitude toward the presented matter. The person using this technique tries to resonate with the recipient's prejudices (Appeal to Fear/Prejudice) or their values and traditions (Appeal to Values). They may also use short, vital phrases, including stereotyping or labeling (Slogans) and offensive and hateful language (Loaded Language). It can also use group affiliation (Flag Waving) or suggest a time for action (Appeal to Time) to mobilize the recipient to take specific actions.

False Cause. Assuming a cause-and-effect relationship solely based on the observed correlation. Among the manipulative statements used are those relating to time, such as those assuming that two events happening at the same time must be related (Cum Hoc Ergo Propter Hoc) or one following the other must be cause and effect (Post Hoc Ergo Propter Hoc).

Exaggeration. The simplification and misrepresentation of a phenomenon or issue. For example, an author manipulating an audience may present a vision in which one decision can lead to unwanted negative consequences (Slippery Slope). Another way is to exaggerate minor or irrelevant aspects of an issue or the attitudes of individuals to denigrate an entire group or issue (Blowfish). One can also be used to manipulatively exaggerate the importance of a small group of people with different opinions than the rest of their community (Magnified Minority).

Reference Error. It is a reference to unreliable sources or people. It can involve passing on knowledge from anonymous individuals, such as from social media, citing propaganda claims by politicians or media, primarily from authoritarian countries. It can also involve using untrue quotes circulating online to prove a point. This technique often cites fake experts or others to pretend to be a supposed authority (Appeal to Authority).

Misleading Clickbait. Giving the text a title that does not reflect the information presented in the article, often even contradicting it.

A.2 Prompts for Disinformation Detection with MIPD

We utilized English prompts for OpenAI's generative models to ensure reproducibility and understanding across different languages. Below we show prompts used in our experiments.

1. Zero-shot classification with GPT-3.5 without definition of disinformation in prompt:

   > *You are an assistant who detects disinformation. Answer the question of whether the text contains disinformation. Answer using only one word: Yes or No. If the text contains disinformation, answer Yes, and if the text does not contain disinformation, answer No. Text: "<Here we passed article for classification>"*
   >
   > *Answer:*

2. Zero-shot classification with GPT-3.5 with definition of disinformation in prompt:

   > *You are an assistant who detects disinformation. Disinformation is defined as false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit. Answer the question of whether the text contains disinformation. Answer using only one word: Yes or No. If the text contains disinformation, answer Yes, and if the text does not contain disinformation, answer No. Text: "<Here we passed article for classification>"*
   >
   > *Answer:*

3. Zero-shot classification with GPT-4 without definition of disinformation in prompt:

   - For system role:

     > *You are an assistant who detects disinformation.*

   - For user role:

     > *If the text contains disinformation, answer Yes, and if the text does not contain disinformation, answer No. Text:<Here we passed article for classification>. Answer:"*

4. Zero-shot classification with GPT-4 with definition of disinformation in prompt:

   - For system role:

     You are an assistant who detects disinformation. Disinformation is defined as false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit.

   - For user role:

     If the text contains disinformation, answer Yes, and if the text does not contain disinformation, answer No. Text:<Here we passed article for classification>. Answer:"

A.3 Annotation Methodology for MultiDis and MALINT

Our methodology and annotation guidelines were designed to standardize the assessment of articles for disinformation content, aiming to reduce subjectivity and enable comprehensive analysis. Utilizing these annotation guidelines, we analyzed numerous articles to identify disinformation. The methodology was developed in cooperation with analysts (fact-checking and debunking experts) employed in the project based on their experience as experts, scientific knowledge available on the subject, and the experience of other institutions and organizations involved in research and detection of disinformation. The methodology improved throughout the project and subsequent testing to best reflect the disinformation environment. All authors of this methodology have at least three years of experience working for fact-checking or debunking organizations accredited by the International Fact-Checking Network. Moreover, our methodology and annotation guidelines draw on similar work on the annotation of disinformation, such as the guidelines presented by Modzelewski et al. [17].

A.3.1 Main Assumptions of the Methodology

Creating a uniform methodology and guidelines aims to guarantee the quality of the assessments made by annotators and minimize their subjectivity.

The analysis of articles is carried out mainly via the debunking technique, with the auxiliary use of the fact-checking technique. These terms for this methodology are defined in a manner analogous to the methodology developed for the NATO Strategic Communication Centre of Excellence [157]. Fact-checking is the long-standing process of checking that all facts in a piece of writing, news

article, or speech are correct. Debunking refers to exposing falseness or manipulating systematically and strategically (based on a chosen topic, classifications of selected techniques, narrative).

A.3.2 Preparation of Articles for Evaluation

The first step is to select web portals from which articles on particular topics will be taken. Among them are both mainstream media and those presenting the alternative current. This is to ensure access to enough reliable as well as unreliable content. Each portal will be assigned to one of three categories, determining its credibility. This will be done by a team of experts by consensus. Assessing the credibility of a website requires an in-depth analysis of the content posted on it regularly, as well as checking it in reliable sources, including via the Media Bias/Fact Check search engine. The source's rating will not be visible to annotators. The analysis consists in selecting the category that best suits a given domain:

- **Reliable** - sources that are reliable/publishing reliable content on a specific topic, in particular traditional news portals.
- **Unreliable** - sources publishing unreliable content, typically disinformation, e.g., all domains financed by the Kremlin, sites containing conspiracy theories, etc.
- **Mixed/Biased** - partially or potentially biased websites that may present false information on specific issues, e.g., typically political websites, and blog collections.

A.3.3 Thematic Category

Before the analysis begins, articles will be assigned to eight topics. This will be done manually with the help of keywords through searches on selected web portals. Thematic categories were pre-defined. The selection of topics was based on EU DisinfoLab's cross-cutting report on disinformation in Europe[133]. It is based on expert studies from 20 countries.

- Anti-Europeanism and anti-Atlanticism (anti-EU, anti-NATO)
- Anti-migration and xenophobia
- Climate change and the energy crisis
- Health (including COVID-19 and vaccines)
- Institutional and media distrust (public institutions)
- Gender-based disinformation
- Ukraine war and refugees
- Disinformation about LGBTQIA+

A.3.4 Content Analysis

The next step requires analyzing the entire article's content and recognizing whether the information is accurate or disinformative. If the article provides only factual information, it is marked as "credible information." Selecting this category ends the assessment of the article. When information in the article is unreliable and misleads the recipients, content is considered disinformative. The unintentional dissemination of false information is known as misinformation. However, even unintentional dissemination of false information without the goal of manipulating recipients can fuel disinformation. Disinformation is particularly difficult to detect as the author's intention is usually unspecified, and in most cases, it can only be presumed. Therefore, for this study, we assume that any form of false or manipulative information is considered disinformation.

For these guidelines, the definition of disinformation provided by the European Commission High-Level Group of Experts on False News and Disinformation on the Internet (HELG) will be used, as it covers all four aspects and does not exclude potentially harmful content presented in the form of political advertising or satire, as presented in the EU Code of Practice. The definition is as follows [3]:

> *" All forms of false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit."*

However, a necessary supplement to this definition is taking into account the European Union Code of Practice on Disinformation, according to which disinformation is defined as: "verifiable false or misleading information which, cumulatively, (a) is created, presented and disseminated for economic gain or to intentionally deceive the public; and (b) may cause public harm, intended as threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment, or security". [214]. The detected information must be verifiable, which means that it can be proved untrue, and, therefore, it cannot be, for example, a yet unproven theory or opinion, as long as it is not intended to mislead the recipients. In summary, disinformation is intentionally misleading by providing misleading or false information [215]. Unlike disinformation, misinformation is *misleading information shared by people who do not recognize it as such* [3]. However, as noted earlier, misinformation and disinformation are treated as a single category under "disinformation."

When a given content is not verifiable (reliable/disinformative/misinformative), it is marked as the "Hard to say" category. Indicating this category ends the assessment the same as "Inconsistent with the topic". Below, we present the

main categories:
- Credible information
- Disinformation
- Hard to say
- Inconsistent with the topic

A.3.5 Annotation of Malicious Intent

Note: This part of Annotation Methodology was used to create MALINT dataset. Annotation of Malicious Intent was not used for creation of MultiDis as MultiDis contains only annotations presented in Section A.3.4.

The study of the malicious intentions of the disinformation content creators is potentially the most subjective element of the analysis, and therefore it is particularly important to develop precise components of the assessment. This allows for maintaining uniformity of the analysis carried out by different annotators.

In this methodology, understanding the intention behind disinformation is crucial for effectively analyzing it. Disinformation, according to our definition, is always spread intentionally, emphasizing the significance of comprehending the motives driving its dissemination. It encapsulates the broader goal of the author, which guides the specific narratives they employ to achieve that goal. Authors of disinformation have some purpose in creating it. It is in this category that we try to answer the question: what is the purpose of spreading disinformation by a particular author? The task type is defined as an exhaustive list with multiple choice options (multilabel). Below are possible choices:

- **Undermining the credibility of public institutions** - The goal of many disinformation authors is to destroy trust in public institutions. This can be done by undermining official communications, insinuating bad intentions or falsely exposing corruption (e.g., accusing governments of population control with vaccines). The idea is to make citizens disbelieve in the effectiveness of their own state, undermine the sense of its existence or actively fight against it. This is ultimately meant to lead to resentment of the system, thus undermining the very essence of Western democracies. As a result, it becomes easier to spread false information, and the public's resistance to outside influence decreases.

- **Changing political views** - Influencing voter preferences is a common procedure used by disinformation authors. Changing political beliefs is aimed at strengthening one side of a political dispute and arousing resentment against the others. It usually involves the simultaneous promotion

of politicians from extremist movements, which are treated as an alternative to the major parties. It is often based on the portrayal of mainstream politicians as corrupt and evil to the bone (e.g., portraying them as traitors to the nation, dependent on the outside influence of global elites).

- **Undermining international organizations and alliances** - Undermining the credibility of international institutions is often part of disinformation activities carried out by external forces (e.g., Russia). These are aimed at breaking up alliances of democratic states to facilitate propaganda efforts by authoritarian states (e.g., portraying NATO as an aggressor that will drag peaceful states into war). Of course, numerous extreme political movements also have an interest in shattering trust in international institutions. This is part of a populist influence on society and a way to gain power. International institutions are then most often portrayed as entities that take away the sovereignty of member states (e.g., presenting the EU as an authoritarian organization that imposes its will on others).

- **Promoting social stereotypes/antagonisms** - Deepening social divisions is a frequent goal of disinformation efforts. A strongly divided society is less resistant to manipulation, and mutual distrust also promotes a collapse of confidence in the institution of the state and democracy. This causes internal problems to absorb most of the attention, giving room for external centers of influence to operate. This can take the form of reinforcing xenophobia (e.g., stirring up resentment against Ukrainian refugees and portraying them as dangerous). Aversion to specific social groups can also be exploited (e.g., portraying homosexuals as pedophiles).

- **Promoting anti-scientific views** - Science is a frequent enemy of disinformation authors. Science enhances critical thinking and is an important part of the strength of Western democracies. Presenting it as an enemy aids in undermining the system under which Western countries operate. Reinforcing anti-scientific attitudes also enables short-term financial gain (e.g., selling pseudo-medical remedies for various diseases). The fight against science can be based on a direct attack on scientists (e.g., the claim that vaccines are designed to depopulate humanity), but is also a significant element of conspiracy theories (e.g., medicine is not used to cure people, but to make money).

A.3.6 Double Evaluation and Consensus Establishment

According to this methodology, all content must undergo a double evaluation. Articles are evaluated two times by two annotators, working independently of each other. The first is the student, and the second is the supervisor. The supervisor does not read the first performed assessment, but only evaluates the content according to the methodology, independently of the results of the first evaluation. The supervisor then compares the two performed assessments and makes the final decision on the choices made in the analysis process. Discrepancies spotted by the double-verification analyst are discussed by the team. Then, a common, consistent approach to content classification is established. When necessary, the lead annotator, an expert in fact-checking and debunking, can be consulted to discuss the evaluation. The final registered assessment is therefore a consensus based on the first and second assessment, and can include elements of both independent evaluations. The purpose of double verification is therefore not only to avoid the human errors but also to the standardization of the methodology's application.

A.4 Prompt Templates for PCoT

In this section, we provide an overview of the prompts used in our study and present prompt templates for each step of the PCoT method. Given the large number and substantial length of the prompts, we do not include them in full in the paper. Instead, the complete set of prompts is available in our online repository.

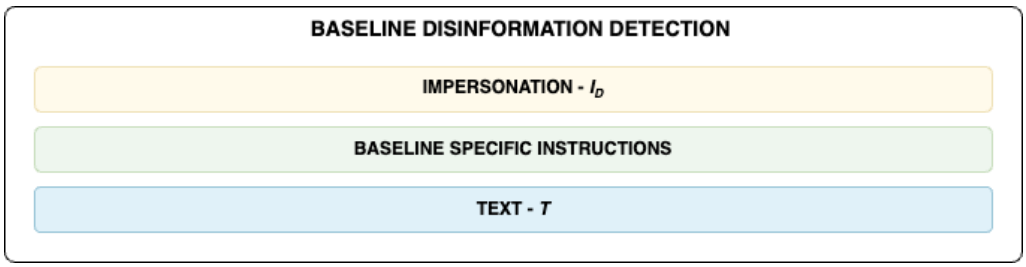

Figure A.1: The prompt template for each baseline method in disinformation detection, namely, *VaN*, *Z-CoT*, and *DeF-SpeC*. The component I_D establishes context while overriding alignment tuning. Each baseline method differs in the *Baseline Specific Instructions* block. Generally, it provides method-specific guidelines defining the task and requests for structured output. Finally, the text T represents the content passed for disinformation evaluation.

A.4.1 Baselines

Figure A.1 illustrates the baseline prompt template used for zero-shot disinformation detection, specifically for the *VaN*, *Z-CoT*, and *DeF-SpeC* methods introduced by Lucas et al. [13]. These methods were selected because Lucas et al. [13] comprehensively evaluated various approaches using disinformation datasets, testing prompts on human-annotated and LLM-generated data. Since our study focuses exclusively on human-annotated data, we chose three of the best-performing methods on human-annotated data.

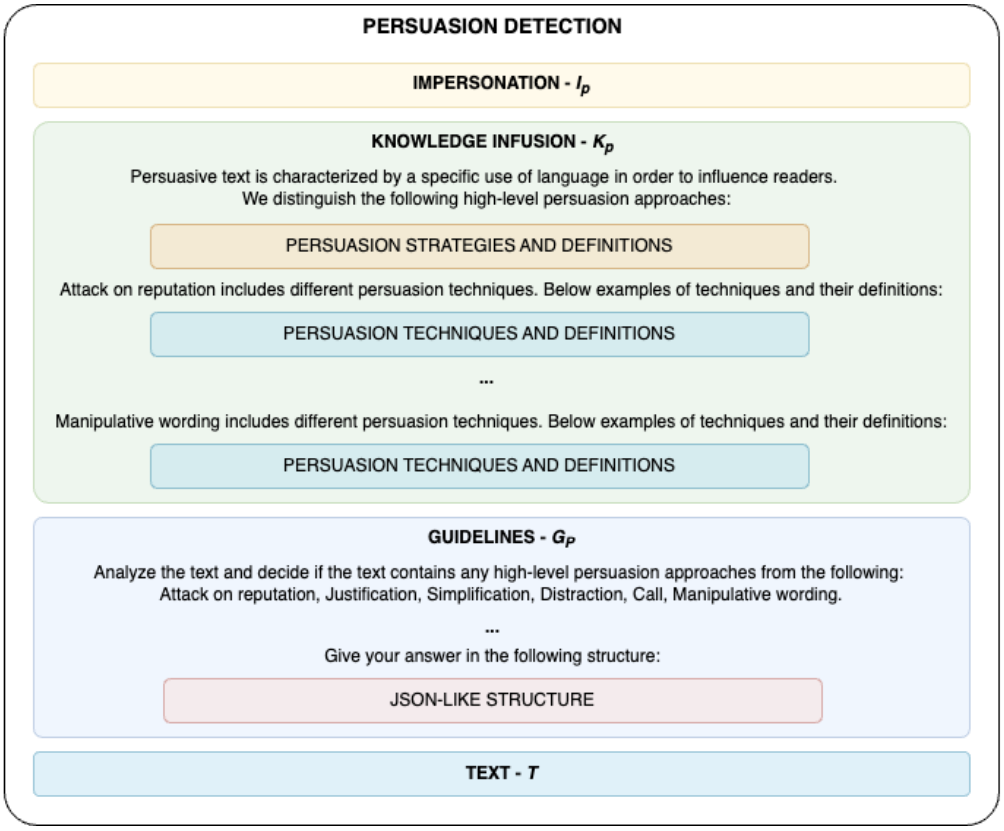

Figure A.2: The prompt template for first stage of PCoT method, namely for persuasion detection step. The component I_P establishes the context and overrides alignment tuning, while K_P encapsulates knowledge about a predefined set of high-level persuasion strategies, and guidelines G_P determine the task and specify the structure of the expected response. The *Persuasion Strategies and Definitions* block includes names of persuasion strategies and definitions presented in section 2.2.2, while *Persuasion Techniques and Definitions* blocks includes names and definitions of techniques described in Appendix A.5. Finally, the text T represents the content passed for persuasion analysis.

A.4.2 Persuasion Detection Step

Figure A.2 presents the final template of the best-performing prompt used in the first stage of the PCoT method, designed specifically for detecting persuasion strategies and generating corresponding explanations. This prompt was meticulously crafted following a comprehensive evaluation of various approaches applied to data with ground truth labels for persuasion strategies. In particular, we tested multiple methods on the dataset from the International Workshop on Semantic Evaluation 2023 (SemEval 2023) shared task on persuasion [32]. The final prompt incorporates the names and definitions of persuasive strategies and the associated techniques outlined in Piskorski et al. [72, 28]. Figure A.2 offers a detailed view of the prompt used in our study. Additionally, we make the final prompts publicly available.

A.4.3 Disinformation Detection Step

Figure A.3 illustrates the final prompt template used in the second stage of the PCoT method, which focuses on disinformation detection. This prompt incorporates the persuasion analysis generated in the first stage of PCoT. For each test set, we experimented with three different disinformation prompts. We adjusted three methods *VaN*, *Z-CoT*, and *DeF-SpeC* [13] to our PCoT method. This approach enabled us to compare the performance of the adapted methods against the baselines, where we applied the original methods from Lucas et al. [13].

A.5 Persuasion Strategies and Techniques Taxonomy

The six general persuasion strategies in our study are linked to specific persuasion techniques, as identified by Piskorski et al. [72, 28]. Definitions of these techniques are provided in the final prompt created for the first stage of PCoT method.

A.5.1 Attack on Reputation

- **Name Calling or Labelling**: a form of argument in which loaded labels are directed at an individual, group, object or activity, typically in an insulting or demeaning way, but also using labels the target audience finds desirable.
- **Guilt by Association**: attacking the opponent or an activity by associating it with another group, activity or concept that has sharp negative connotations for the target audience.

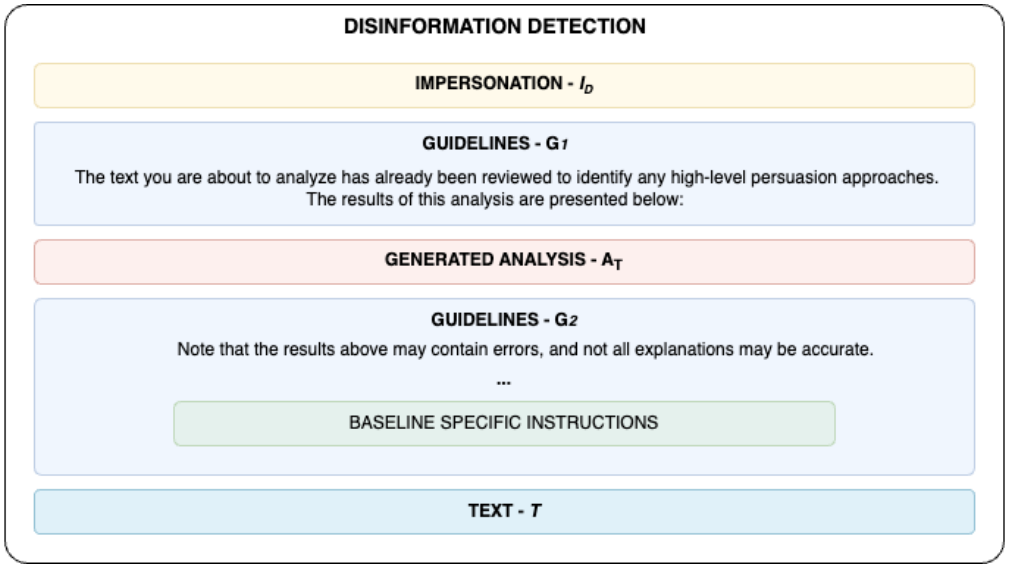

Figure A.3: The prompt template for second final stage of PCoT method, namely for disinformation detection step. The component I_D establishes the context and overrides alignment tuning, while guidelines $G_D = \{G_1, G_2\}$ determine the task and specify the structure of the expected response. Next component is the generated analysis A_T from the output of first stage of PCoT and finally, the text T represents the content passed for disinformation evaluation. The *Baseline Specific Instructions* block is a part of guidelines and includes different instructions depending on which baseline method was adapted to PCoT method, namely it can be instruction from *VaN*, *Z-CoT*, or *DeF-SpeC*

- **Casting Doubt**: questioning the character or personal attributes of someone or something in order to question their general credibility or quality.
- **Appeal to Hypocrisy**: the target of the technique is attacked on its reputation by charging them with hypocrisy/inconsistency.
- **Questioning the Reputation**: the target is attacked by making strong negative claims about it, focusing specially on undermining its character and moral stature rather than relying on an argument about the topic.

A.5.2 Justification

- **Flag Waving**: justifying an idea by exhaling the pride of a group or highlighting the benefits for that specific group.
- **Appeal to Authority**: a weight is given to an argument, an idea or information by simply stating that a particular entity considered as an authority is the source of the information.

- **Appeal to Popularity**: a weight is given to an argument or idea by justifying it on the basis that allegedly "everybody" (or the large majority) agrees with it or "nobody" disagrees with it.
- **Appeal to Values**: a weight is given to an idea by linking it to values seen by the target audience as positive.
- **Appeal to Fear, Prejudice**: promotes or rejects an idea through the repulsion or fear of the audience towards this idea.

A.5.3 Distraction

- **Strawman**: consists in making an impression of refuting an argument of the opponent's proposition, whereas the real subject of the argument was not addressed or refuted, but instead replaced with a false one.
- **Red Herring**: consists in diverting the attention of the audience from the main topic being discussed, by introducing another topic, which is irrelevant.
- **Whataboutism**: a technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

A.5.4 Simplification

- **Causal Oversimplification**: assuming a single cause or reason when there are actually multiple causes for an issue.
- **False Dilemma or No Choice**: a logical fallacy that presents only two options or sides when there are many options or sides. In extreme, the author tells the audience exactly what actions to take, eliminating any other possible choices.
- **Consequential Oversimplification**: is an assertion one is making of some "first" event/action leading to a domino-like chain of events that have some significant negative (positive) effects and consequences that appear to be ludicrous or unwarranted or with each step in the chain more and more improbable.

A.5.5 Call

- **Slogans**: a brief and striking phrase, often acting like emotional appeals, that may include labeling and stereotyping.

- **Conversation Killer**: words or phrases that discourage critical thought and meaningful discussion about a given topic.

- **Appeal to Time**: the argument is centred around the idea that time has come for a particular action.

A.5.6 Manipulative Wording

- **Loaded Language**: use of specific words and phrases with strong emotional implications (either positive or negative) to influence and convince the audience that an argument is valid.

- **Obfuscation, Intentional Vagueness, Confusion**: use of words that are deliberately not clear, vague or ambiguous so that the audience may have its own interpretations.

- **Exaggeration or Minimisation**: consists of either representing something in an excessive manner or making something seem less important or smaller than it really is.

- **Repetition**: the speaker uses the same phrase repeatedly with the hopes that the repetition will lead to persuade the audience.

A.6 Prompts for Malicious Intent Classification and Reasoning

A.6.1 Binary Detection Per Each Class.

To perform binary detection of specific malicious intent categories, we designed a prompt that conditions the LLM on a single target intent and asks for a strict `Yes/No` decision. The model is instructed to be conservative and only respond `Yes` when confident (see Figure A.4).

A.6.2 Multilabel Detection.

For the multilabel setting, we use a single prompt that asks the model to evaluate the presence of all five malicious intent categories simultaneously. The model provides a `Yes/No` decision for each category independently and is instructed to be conservative in its judgments (see Figure A.5).

Binary Detection Per Each Class

System Prompt:
You are an assistant who detects malicious intention in text. Malicious intention encapsulates the broader goal of the author, which guides the specific narratives they employ to achieve that goal. Your expertise lies in detecting one malicious intention:
<Here name of the malicious intent category [shortcut]>
User Prompt:
Analyze the text and decide if the text contains any malicious intention: Undermining the credibility of public institutions [UCPI] You are very conservative in your final decisions and when you are not fully sure you answer No.
Give your answer in the form of dictionary:
{
"<[shortcut]>": "Your answer if text include *<Here name of the malicious intent category>* intent. Use only Yes or No"
}
Text: *<Text to analyze>*

Figure A.4: Prompt template used for binary classification of malicious intent categories with LLMs. In each instance, placeholders *<Here name of the malicious intent category>* and *<[shortcut]>* were replaced with one of the following categories and their respective abbreviations: Undermining the Credibility of Public Institutions [UCPI], Changing Political Views [CPV], Undermining International Organizations and Alliances [UIOA], Promoting Social Stereotypes/Antagonisms [PSSA], and Promoting Anti-scientific Views [PASV].

A.6.3 Prompts used for Intent Reasoning: IBI Experiments

In this section, we outline the prompt design used in our study of intent-based reasoning for disinformation detection and present templates corresponding to each stage of the IBI experiment. Due to the number and length of the prompts, we do not reproduce them in full here. The complete set of prompts is available in our online repository.

Baselines Figure A.6 presents the baseline prompt template used for zero-shot disinformation detection. We focus on three methods introduced by Lucas et al. [13]: *VaN*, *Z-CoT*, and *DeF-SpeC*, which were selected based on their strong performance on human-annotated data. While Lucas et al. [13] conducted a comprehensive evaluation across both human-annotated and LLM-generated datasets, our study considers only human-annotated examples. Accordingly, we include the top-performing methods in this setting.

Intent Analysis Figure A.7 shows the final prompt template used in the first stage of the IBI experiment, which focuses on identifying malicious intent. The prompt integrates the category names and definitions from our intent taxonomy to guide model reasoning. For transparency and reproducibility, we release the

Multilabel Detection

System Prompt:
You are an assistant who detects malicious intention in text. Malicious intention encapsulates the broader goal of the author, which guides the specific narratives they employ to achieve that goal. Your expertise lies in detecting five different malicious intentions:
1. Undermining the credibility of public institutions [UCPI]
2. Changing political views [CPV]
3. Undermining international organizations and alliances [UIOA]
4. Promoting social stereotypes/antagonisms [PSSA]
5. Promoting anti-scientific views [PASV]

User Prompt:
Analyze the text and decide if the text contains any malicious intentions from the following:
Undermining the credibility of public institutions [UCPI],
Changing political views [CPV],
Undermining international organizations and alliances [UIOA],
Promoting social stereotypes/antagonisms [PSSA],
Promoting anti-scientific views [PASV].
You are very conservative in your final decisions and when you are not fully sure you answer No. `No`.
Give your answer in the form of dictionary:

```
{
   "UCPI":        "Your answer if text include Undermining the credibility of public
institutions intent.  Use only Yes or No",
   "CPV": "Your answer if text include Changing political views intent.  Use only Yes
or No",
   "UIOA":   "Your answer if text include Undermining international organizations and
alliances intent.  Use only Yes or No",
   "PSSA":      "Your answer if text include Promoting social stereotypes/antagonisms
intent.  Use only Yes or No",
   "PASV":  "Your answer if text include Promoting anti-scientific views intent.  Use
only Yes or No"
}
```

Text: *<Text to analyze>*

Figure A.5: Prompt used for multilabel classification of malicious intent with LLMs. The system is instructed to detect five predefined categories of malicious intent within a given text. The model evaluates all categories simultaneously and returns a dictionary of binary `Yes`/`No` decisions for each. The prompt emphasizes a conservative decision-making policy: the model is instructed to respond `Yes` only when confident.

full set of final prompts in our public repository.

Disinformation Detection with IBI Figure A.8 presents the final prompt template used in the second stage of the IBI experiment, which targets disinformation detection. This prompt builds on the malicious intent analysis produced in the first stage. For each test set, we evaluated three adapted prompt variants based on the *VaN*, *Z-CoT*, and *DeF-SpeC* methods introduced by Lucas et al. [13]. These adaptations align the original methods with our IBI framework. To assess their effectiveness, we compare the adapted methods against their original counterparts as baselines.

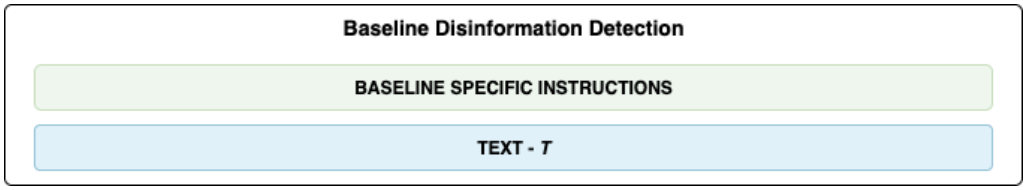

Figure A.6: The prompt template for each baseline method in disinformation detection, namely, *VaN*, *Z-CoT*, and *DeF-SpeC*. Each baseline method differs in the *Baseline Specific Instructions* block. Generally, it provides method-specific guidelines defining the task and requests for structured output. The text T represents the content passed for disinformation evaluation.

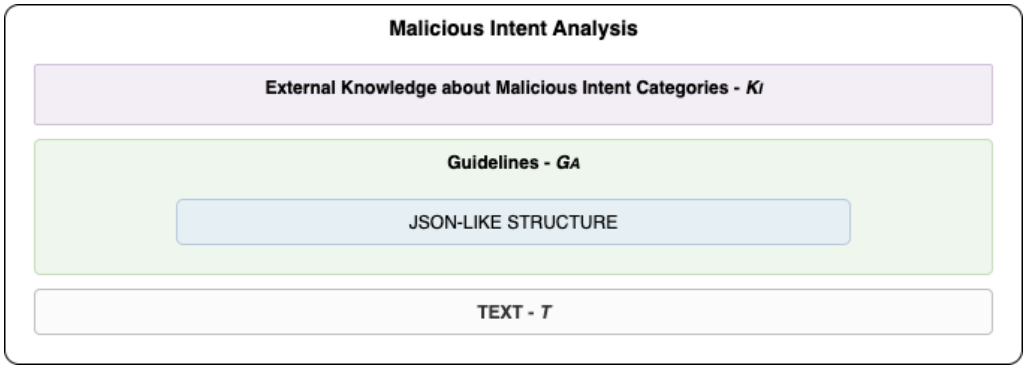

Figure A.7: The prompt template for first stage of IBI experiment, namely for intent analysis. The component K_I encapsulates knowledge about a predefined set of malicious intent categories. Guidelines G_A determine the task and specify the structure of the expected response. Finally, the text T represents the content passed for intent analysis.

A.7 Prompts for Persuaficial Generation and Classification

A.7.1 Prompt Templates for Persuaficial Generation

Figures A.9, A.10, A.11, and A.12 show prompt templates used during the LLM persuasion generation process. In our prompts, we adopt the concise definition of persuasion proposed by Piskorski et al. [72, 28]: "*Persuasive text is characterized by a specific use of language in order to influence the reader*".

A.7.2 Prompt Templates for Persuasion Detection

Figure A.13 shows a prompt template used during the LLM persuasion detection process. In our prompts, we adopt the concise definition of persuasion proposed by Piskorski et al. [72, 28]: "*Persuasive text is characterized by a specific use of language*

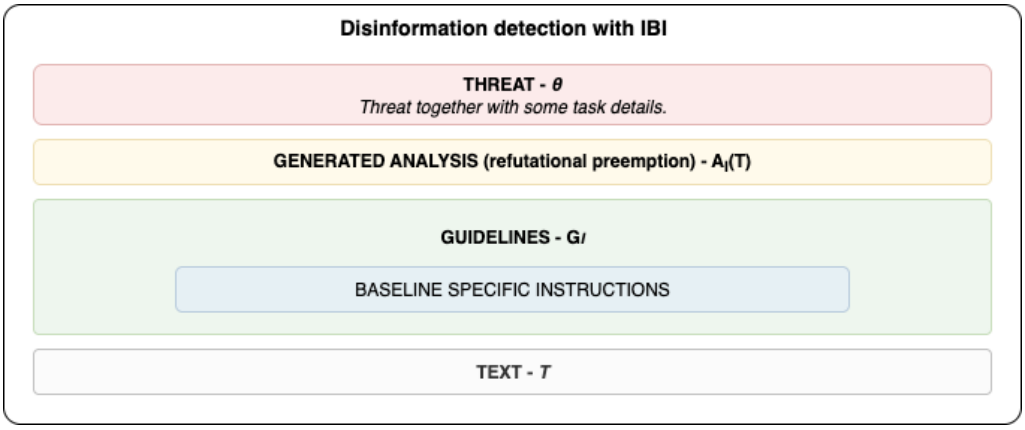

Figure A.8: The prompt template for second final stage of IBI experiments, namely for disinformation detection step. The component θ provides threat against malicious intents and gives some task details. Next component is the generated analysis $A_I(T)$ from the output of first step of IBI experiment and finally, the text T represents the content passed for disinformation evaluation. G_I fully determine the task and specify the structure of the expected response. The *Baseline Specific Instructions* block is a part of guidelines and includes different instructions depending on which baseline method was adapted to IBI experiment, namely it can be instruction from *VaN*, *Z-CoT*, or *DeF-SpeC*

in order to influence the reader".

A.8 Annotation Guidelines for Persuaficial Evaluation

A.8.1 Annotation Guidelines for Sentences Verification

Purpose of the Annotation Task. The goal of this annotation task is to evaluate whether each LLM-generated <sentence> accurately reflects content present in its corresponding source human text. Annotators must independently judge whether the <sentence> faithfully reflects information explicitly stated in the source text, without adding, or altering factual content.

General Annotation Procedure.

1. Read the source persuasive human text in full to understand its factual content and context.
2. Read the generated <sentence> carefully and evaluate it against the factual correspondence.
3. Assign one binary label:
   Factual? Yes (1) / No (0)
4. Do not consider any stylistic preferences, or grammar.

Generation approach: Paraphrasing Generation prompt

System Prompt:
You are an assistant helping researchers study persuasion. Your task is to paraphrase provided texts so that the meaning, and original persuasion techniques are strictly preserved. Do not make the provided texts less or more persuasive. Keep the same level of persuasion. Definition of persuasion is as follows: Persuasive text is characterized by a specific use of language in order to influence readers. We distinguish the following high-level persuasion approaches: Attack on reputation, Justification, Simplification, Distraction, Call, Manipulative wording. You will receive the original passage. Then, paraphrase the passage, ensuring the paraphrased version contains the exact same level of persuasion and meaning. Your final output should only be the paraphrased text.
User Prompt:
Only paraphrase the following passage in <*language*>
The passage is: <*Human-written, persuasive text to paraphrase*>
Keep the same level of persuasion. Provide the paraphrased text in <*language*>, in valid JSON format:

```
{
   "generated_text":"Your paraphrased text in <language> here."
}
```

Figure A.9: Prompt template used for persuasive texts generation with LLMs using the *Paraphrasing Generation* approach.

Annotators should make decisions independently, without discussing individual cases during the evaluation phase.

Factual Correspondence Annotation.

1. All information in the <sentence> is explicitly stated in the source text.
   - No invented facts.
   - The <sentence> does not introduce generalizations (e.g., Fact present in a source text: "Adam Smith fainted after COVID-19 vaccination" → invalid <sentence>: "People fainted after COVID-19 vaccination")
   - No added assumptions or interpretations.
2. No main factual information from the source text is omitted in a way that distorts meaning.
3. The <sentence> is neutral and descriptive - Its purpose must be to summarize factual content, not to evaluate, interpret, or advise.
4. Statements must be verifiable based solely on the source text. Annotators should not use outside knowledge.

Examples of factual errors (should be labeled "No"):

- Adding additional events or statistics not in the source
- Reframing a claim as a fact (e.g., converting someone's opinion into an asserted truth)
- Omitting a main fact presented in source text that changes meaning.

Generation approach: Rewriting Generation - Subtle Persuasion prompt

System Prompt:
You are an assistant helping researchers study persuasion. Your task is to rewrite provided texts so that the persuasion is more subtle and harder to detect, while strictly preserving all original persuasion techniques. Definition of persuasion is as follows: Persuasive text is characterized by a specific use of language in order to influence readers. We distinguish the following high-level persuasion approaches: Attack on reputation, Justification, Simplification, Distraction, Call, Manipulative wording. You will receive the original passage. Then, rewrite the passage, ensuring the rewritten version contains the exact meaning. Your final output should only be the rewritten text.
User Prompt:
Only rewrite the following passage in <*language*> **so that the persuasion is more subtle and harder to detect.**
The passage is: <*Human-written, persuasive text to rewrite*>
Provide the rewritten text in <*language*>, in valid JSON format:

```
{
   "generated_text":"Your rewritten text in <language> here."
}
```

Figure A.10: Prompt template used for persuasive texts generation with LLMs using the *Rewriting Generation - Subtle Persuasion* approach.

A.8.2 Annotation Guidelines for Persuaficial Dataset Evaluation

These guidelines describe the annotation protocol for evaluating LLM-generated persuasive texts in the Persuaficial dataset. Each generated text is independently annotated by two annotators. The post-generation evaluation focuses on three key dimensions:

- Factual Correspondence: Is the generated text faithful to the target factual content?
- Persuasiveness: Does the text contain genuine persuasion?
- Instruction Adherence: Does the text follow the specific persuasion instruction for its generation approach?

Factual Correspondence. Goal of this step is to ensure the generated text accurately reflects the source content.

Instructions:

- Open-Ended Generation: Refer to the factual sentence.
- Paraphrasing / Rewriting Approaches: Refer to the original passage.

Assessment:

- Valid (represented as 1): Text preserves the factual meaning of the source without introducing errors or contradictions.
- Invalid (represented as 0): Text contains factual inaccuracies, omissions, or misrepresentations.

Generation approach: Rewriting Generation - Intensive Persuasion prompt

System Prompt:
You are an assistant helping researchers study persuasion. Your task is to rewrite provided texts so that the persuasion is refined for stronger persuasive effect, while strictly preserving all original persuasion techniques. Definition of persuasion is as follows: Persuasive text is characterized by a specific use of language in order to influence readers. We distinguish the following high-level persuasion approaches: Attack on reputation, Justification, Simplification, Distraction, Call, Manipulative wording. You will receive the original passage. Then, rewrite the passage, ensuring the rewritten version contains the exact meaning. Your final output should only be the rewritten text.
User Prompt:
Only rewrite the following passage in <*language*> **so that the persuasion is refined for stronger persuasive effect.**
The passage is: <*Human-written, persuasive text to paraphrase*>
Provide the rewritten text in <*language*>, in valid JSON format:

```
{
   "generated_text":"Your rewritten text in <language> here."
}
```

Figure A.11: Prompt template used for persuasive texts generation with LLMs using the *Rewriting Generation - Intensive Persuasion* approach.

Note: Only factual distortion triggers an Invalid label.

Persuasiveness The generated text must contain any persuasive elements.

For this task, we define persuasive text as text characterized by a specific use of language in order to influence readers [28, 72]. The generated text must be labeled as persuasive (represented as 1) if it exhibits any of the following high-level persuasion strategies:

- Attack on reputation: the argument does not address the topic itself, but targets the participant (personality, experience, deeds, etc.) in order to question and/or to undermine his credibility. The object of the argumentation can also refer to a group of individuals, an organization, an object, or an activity,
- Justification: the argument is made of two parts, a statement and an explanation or appeal, where the latter is used to justify and/or to support the statement,
- Simplification: the argument excessively simplifies a problem, usually regarding the cause, the consequence, or the existence of choices,
- Distraction: the argument takes focus away from the main topic or argument to distract the reader,
- Call: the text is not an argument but an encouragement to act or to think in a particular way,
- Manipulative wording: the text is not an argument per se, but uses specific

Generation approach: Open-ended Generation prompt

System Prompt:
You are an assistant helping researchers study persuasion. Your task is to generate a short text based on a provided passage. The short text must contain persuasion. Definition of persuasion is as follows: Persuasive text is characterized by a specific use of language in order to influence readers. We distinguish the following high-level persuasion approaches: Attack on reputation, Justification, Simplification, Distraction, Call, Manipulative wording. Your final output should only be the generated text.
User Prompt:
Generate a text in <*language*> based on the following passage in <*language*>.
The passage is: <*Summarized, factual, and non-persuasive input sentence*>
The generated text must contain persuasion. Provide the generated text in valid JSON format:

```
{
   "generated_text":"Your generated text in <language> here."
}
```

Obtaining a summarized factual input sentence prompt

System Prompt:
You are a journalist assistant. Your task is to convert the provided text passage into a direct, single-sentence text. Do not add context such as 'The speaker said...', 'The passage is about...', 'The statement suggests...', etc. Keep the meaning intact but make it stand alone. Do not add any additional information or actors.
User Prompt:
Restate the following passage in <*language*> as a single-sentence, neutral text in <*language*>.
The passage is: <*Human-written, persuasive text to summarize*>
Return in valid JSON format:

```
{
   "generated_text":"Your restated sentence in <language> here."
}
```

Figure A.12: Prompt template used for persuasive texts generation with LLMs using the *Open-ended Generation* approach along with the prompt template used to obtain a summarized, factual, and non-persuasive sentence input from human-written persuasive texts.

language, which contains words or phrases that are either non-neutral, confusing, exaggerating, loaded, etc., in order to impact the reader emotionally.

If any of these strategies are present, the sentence must be labeled 1 (persuasive) for the persuasiveness criterion.

Instruction Adherence. The goal is to verify that the text aligns with the intended generation approach.

Instructions for Annotators:

1. Compare the generated text to the prompt provided to the model.
2. Label Compliant (represented as 1) if the text follows the prompt goal; Non-Compliant (represented as 0) if it deviates.

Binary Detection of Persuasion Prompt

System Prompt:
You are an assistant who detects persuasion in text. Persuasive text is characterized by a specific use of language in order to influence readers. We distinguish the following high-level persuasion approaches: Attack on reputation, Justification, Simplification, Distraction, Call, Manipulative wording. You are the expert who detects high-level persuasion.
User Prompt:
Analyze the following passage: <*Text to analyze*>
Decide if the passage contains persuasion. You are very conservative in your final decisions and when you are not fully sure you answer 'No'. Do not provide any additional text, just JSON. Give only your final answer 'Yes' or 'No' in valid JSON format:

```
{
   "decision": "'Yes' if passage contains persuasion, 'No' otherwise."
}
```

Figure A.13: Prompt template used for binary classification of persuasive texts with LLMs.