POLISH-JAPANESE ACADEMY
OF INFORMATION TECHNOLOGY

Informatyka

mgr inż. Witold Sosnowski

Classifying and Mining Disinformation Narratives With Large Language Models

Doctoral Dissertation

Promotor:
prof. dr hab. Adam Wierzbicki
Promotor pomocniczy:
dr Kinga Skorupska

2026

Abstract

This dissertation explores how disinformation narratives can be conceptualized, classified, and mined with Large Language Models (LLMs). While existing computational approaches to disinformation predominantly operate at the item level, analyzing individual claims, posts, or articles in isolation, this work elevates the unit of analysis to the narrative level. This shift enables more stable measurement of how disinformation emerge and spread over time and across sources. The dissertation proposes three components that evaluate LLMs on disinformation narrative classification and enable narrative mining at both global and outlet level.

First, it introduces EUDisinfoTest, a narrative-level benchmark that assesses whether LLMs can distinguish disinformation narratives from credible narratives across multiple topical domains. The benchmark explicitly tests robustness to persuasive reframing by providing controlled rhetorical variants aligned with Logos, Ethos, and Pathos, enabling systematic analysis of how rhetorical strategy affects model decisions.

Second, it proposes DiNaM, a scalable narrative-mining pipeline that constructs a European-wide narrative map from large collections of fact-checking articles. DiNaM extracts verified false information instances, represents them in semantic embedding spaces, clusters them into coherent groups, and generates narratives that are explicitly grounded within each cluster.

Third, it presents DiNO, an outlet-level framework that anchors narrative mining in databases of verified false claims to identify, consolidate, and characterize disinformation narratives expressed in news coverage. DiNO enables comparative analysis of narrative prevalence and framing across outlets and over time.

Experimental results show that the proposed, implemented and evaluated mining pipelines, which leverage LLMs at key stages, successfully recover coherent global maps of disinformation narratives and derive complementary outlet level narrative profiles narrowed down to specific news outlets. Additionally, the evaluation reveals that contemporary Large Language Models can achieve strong performance on narrative-level classification; however, their judgements remain sensitive to rhetorical variation, indicating that persuasive framing can systematically influence model decisions even when the underlying factual content remains unchanged.

Streszczenie

Niniejsza rozprawa bada, w jaki sposób narracje dezinformacyjne mogą być definiowane, klasyfikowane oraz automatycznie wydobywane z wykorzystaniem dużych modeli językowych. Podczas gdy większość podejść obliczeniowych analizuje dezinformację na poziomie pojedynczych treści (twierdzenia, posty, artykuły), w tej pracy podstawową jednostką analizy staje się narracja dezinformacyjna. Takie ujęcie pozwala stabilniej opisywać i mierzyć, jak dezinformacja rozprzestrzenia się w czasie i pomiędzy różnymi źródłami. Rozprawa wprowadza trzy komponenty wspierające ewaluację modeli w klasyfikacji narracji dezinformacyjnych oraz ich wydobywanie zarówno globalnie, jak i na poziomie pojedynczych mediów.

Po pierwsze, rozprawa przedstawia EUDisinfoTest, benchmark do oceny zdolności dużych modeli językowych do rozróżniania narracji dezinformacyjnych od wiarygodnych w wielu domenach. Narzędzie dodatkowo testuje wpływ, jaki strategie retoryczne (Logos, Etos, Patos) wywierają na decyzje klasyfikacyjne modeli.

Po drugie, praca proponuje DiNaM, metodę, która rekonstruuje globalny zbiór narracji dezinformacyjnych na podstawie dużych korpusów artykułów fact-checkingowych. Metoda identyfikuje zweryfikowane przypadki fałszywych informacji, grupuje treści o zbliżonym znaczeniu w spójne zbiory, a następnie generuje narracje na podstawie wspólnego rdzenia każdej grupy, tworząc tematyczną mapę narracji wraz z analizą trendów w czasie.

Po trzecie, rozprawa prezentuje DiNO, metodę, która skupia się na analizie pojedynczych źródeł medialnych. W przeciwieństwie do metody DiNaM, która rekonstruuje narracje na podstawie artykułów weryfikujących fakty, DiNO wykorzystuje katalog fałszywych twierdzeń jako punkt odniesienia do identyfikacji podobnych treści w artykułach danego źródła, a następnie grupuje je w spójne klastry i na tej podstawie generuje narracje dezinformacyjne.

Wyniki badań pokazują, że zaproponowane metody wydobywania narracji przy wykorzystaniu dużych modeli językowych skutecznie odtwarzają zbiory narracji dezinformacyjnych zarówno w skali globalnej, jak i na poziomie pojedynczych źródeł medialnych. Ponadto, modele tego typu osiągają wysoką skuteczność w klasyfikacji narracji jako dezinformacyjne lub wiarygodne, jednak ich oceny pozostają podatne na wpływ strategii retorycznych.

Contents

Chapter 1

Introduction

1.1 Motivations

Over the past decade, disinformation has shifted from sporadic falsehoods and isolated "fake news" stories to a structural feature of the contemporary information environment. Social media platforms, partisan outlets and coordinated campaigns jointly shape public discourse [201, 29]. This "information disorder" undermines trust in democratic institutions and news media, distorts electoral debates, and creates tangible risks in areas such as public health and security [201, 13]. Large-scale empirical studies further show that false and misleading content can spread faster and farther than accurate information online [197], particularly when it aligns with existing beliefs and partisan identities [116]. In response, governments, civil-society organisations and research consortia have invested in large-scale fact-checking networks and observatories dedicated to detecting, documenting and analysing disinformation [146, 201].

A growing body of work and operational experience suggests that contemporary disinformation often propagates through recurring narratives rather than only through isolated false claims. Here, a disinformation narrative can be understood as "the clear message that emerges from a consistent set of contents that can be demonstrated as false using the fact-checking methodology" [146]. Reflecting this shift, institutional monitoring initiatives routinely track, summarise, and compare narratives as a primary unit for describing disinformation trends and campaigns [7, 79, 62]. However, despite the centrality of narratives in practice, the computational landscape remains uneven. Existing narrative mining methods can recover latent structure from large corpora, but they are typically general purpose and veracity agnostic. As a result, they can extract narratives without supporting the practitioner requirement of identifying disinformation narratives as units whose status is evidence grounded and defensible [20, 18]. Some misinformation

focused workflows combine clustering or topic modeling with expert review to consolidate content into narrative like categories, yet these approaches commonly depend on substantial human in the loop analysis. They also do not amount to a widely established end to end computational pipeline for mining and validating disinformation narratives at scale [115, 181, 172].

Moreover, both classical rhetoric and modern persuasion research highlight that the impact of a message depends not only on its propositional content but also on how it mobilises logical reasoning, perceived credibility and authority, and emotions and values [95, 97]. Research in misinformation shows that such rhetorical strategies systematically shape how audiences perceive the intent, credibility and trustworthiness of political and health-related messages [57]. Yet most benchmarks for automated disinformation classification do not explicitly control for or annotate these persuasive dimensions, making it difficult to understand where current models are robust or vulnerable [184]. This issue becomes particularly important at the narrative-level, where the same underlying message can be re-expressed through different rhetorical strategies to persuade different audiences or evade moderation.

Large Language Models further amplify both the risks and opportunities of this landscape. On the one hand, LLMs enable low-cost, high-volume generation of tailored, multilingual and stylistically sophisticated text, which can be misused to produce persuasive disinformation at scale [75]. On the other hand, frontier models such as GPT-4 achieve strong performance on a broad range of classification and reasoning benchmarks and are increasingly proposed as tools for fact-checking assistance, content moderation and narrative analysis [5, 106]. At the same time, current evidence about how reliably LLMs can distinguish disinformation narratives from credible narratives, especially under variation in rhetorical strategy, topic and stance, remains limited [184]. Without rigorous narrative-level evaluation, deploying these models inside narrative mining and monitoring systems risks producing unstable narrative maps and misleading source profiles.

Building on this context, the thesis is motivated by three interrelated gaps in current research and practice.

First, we lack rigorous narrative focused benchmarks that evaluate whether LLMs can detect disinformation at the level of narratives and whether their judgements remain stable when the same narrative content is expressed through different rhetorical strategies.

Second, we lack methods that turn large collections of fact checking articles into a coherent and reusable global map of disinformation narratives that recur

across countries and domains and evolve over time.

Third, we lack scalable approaches that identify which disinformation narratives appear in specific news outlets and how those outlet specific narrative profiles differ across sources and over time.

Overall, the thesis aims to bridge the gap between item level disinformation classification and narrative-level understanding by providing tools and evidence for evaluation, global mapping, and outlet level analysis of disinformation narratives.

1.2 Research Objectives

Building on the gaps outlined above, this thesis pursues four concrete research objectives.

O1: Conceptual foundations for disinformation narrative classification and mining.
To develop a clear conceptualisation of disinformation narratives and of the associated tasks of narrative classification and narrative mining. The aim is to bridge expert-oriented definitions with task setups suitable for computational modelling (Chapter 3).

O2: Benchmarking LLMs on disinformation narrative classification.
To develop a benchmark for evaluating how LLMs distinguish credible narratives from disinformation narratives across topics and rhetorical strategies. This objective focuses on analysing model performance, robustness and systematic biases in narrative-level classification. It is addressed through the EUDisinfoTest benchmark (Chapter 4).

O3: Constructing a global map of disinformation narratives from fact-checking articles.
To develop methods for constructing a global map of disinformation narratives using fact checking articles as a primary data source. The goal is to formalise narrative mining for fact checking corpora and to characterise which narratives recur within a region spanning multiple countries, such as the European Union, and how they evolve over time. It is addressed through the DiNaM framework (Chapter 5).

O4: Analysing outlet-level realisations of disinformation narratives.
To extend disinformation narrative mining to news outlets, providing a local,

outlet specific perspective that complements the global mapping derived from fact checking corpora. The objective is to characterise which disinformation narratives are present in particular outlets and how they differ across outlets. It is addressed through the DiNO framework (Chapter 6).

1.3 Overview of Contributions and Publications

This thesis makes nine contributions (C1–C9) spanning Information and Telecommunication Technologies and Political Science. Several contributions have been peer-reviewed and published; the thesis integrates these results and adds additional analyses, methodological detail, and cross-chapter synthesis.

Contributions in the field of Information and Telecommunication Technologies

C1: EUDisinfoTest, a benchmark for LLM-based disinformation narrative classification.
The thesis presents EUDisinfoTest, a large-scale benchmark grounded in expert-curated European disinformation narratives. It includes systematically constructed rhetorical variants for each narrative, enabling fine-grained evaluation of LLM performance and robustness to persuasive framing (Chapter 4). A peer reviewed version appears in [184].

C2: Empirical analysis of LLM robustness and biases under rhetorical strategies.
Using EUDisinfoTest, the thesis provides a systematic empirical study of how contemporary LLMs behave when confronted with narrative-level disinformation in different rhetorical variants, identifying consistent biases and failure modes. In particular, the analysis shows how appeals to authority (ethos) and emotional framing (pathos) can systematically alter model judgements, even when the underlying factual content is held constant. These findings contribute to the broader literature on LLM risks and limitations (Section 4.7). This contribution is reported in [184].

C3: Three-stage framework for disinformation narrative mining.
The thesis presents a general three-stage framework for disinformation narrative mining that directly mirrors the defining properties of disinformation narratives. The framework (i) extracts verifiably false information instances

from a corpus, (ii) clusters these instances into semantically coherent groups, and (iii) derives narratives that summarise each cluster together with their support sets. This abstraction provides a unifying perspective for different narrative-mining instantiations, clarifies how narrative-level representations can be linked back to concrete content items, and offers a modular basis for subsequent algorithm design and evaluation (Section 3.2.2). A peer reviewed version appears in [185].

C4: Method for LLM-based extraction of false information from fact-checking articles.

The thesis presents a novel LLM-based method for extracting false information from fact-checking articles, formulated as a multi-step procedure. The method (i) extracts candidate false statements, (ii) verifies that each candidate is a clean, misleading formulation of the false claim (free of debunking/corrective language), and (iii) selectively refines failing cases to remove debunking language while preserving the misleading claim. This approach mitigates LLMs' tendency to entangle claims with corrective context and outperforms competing methods across models (Sections 5.2.2 and 5.4.2). This contribution is reported in [185].

C5: DiNaM, a framework for global mining of disinformation narratives from fact-checking articles.

Building on this extraction method (C4) and the general three-stage framework in C3, the thesis presents DiNaM, an instantiation of disinformation narrative mining for fact-checking articles. DiNaM aggregates false-information instances across many fact-checkers and topics to construct a global view of the disinformation narrative landscape, identifying the main narratives that recur across countries, domains and time. In evaluation against general-purpose narrative mining methods, DiNaM consistently outperforms these baselines and recovers more meaningful disinformation narratives (Chapter 5). A peer-reviewed version appears in [185].

C6: Method for extracting false information from news articles.

The thesis introduces a separate method for extracting false claims from news articles. This method (i) segments articles into candidate spans, (ii) matches these spans to large collections of verified false claims, and (iii) uses entailment-style verification to retain only segments that actually support a false claim. It yields high-precision anchors for disinformation narratives directly in

news text and is tailored to the noisier, non-annotated setting of outlet content, complementing the fact-checking-oriented extraction procedure in C4 (Sections 6.2.1 and 6.4.1).

C7: DiNO, a framework for mining disinformation narratives from news outlets.
Extending narrative mining beyond fact-checking corpora, the thesis proposes DiNO, an outlet-level instantiation of the narrative-mining framework for news articles. DiNO combines large collections of verified false claims with a collection of news articles to map which disinformation narratives appear in a given outlet. Experiments against narrative-mining baselines and expert assessment show that DiNO reliably recovers coherent, meaningful narratives from news outlets (Chapter 6).

Peer reviewed publications underpinning this thesis

The dissertation is primarily based on two peer-reviewed EMNLP publications (CORE A*, 140 ministerial points in the Polish national evaluation system), and it is further supported by ongoing work currently under review at ACL ARR.

EUDisinfoTest: Benchmarking LLMs for Disinformation Narrative Classification [184] (EMNLP 2024, CORE A*, 140 ministerial points)
This paper introduces EUDisinfoTest, a benchmark for evaluating LLMs' ability to distinguish between credible and disinformation narratives, both in their original form and under controlled rhetorical variants (logos/ethos/pathos) that test robustness to persuasive framing.

DiNaM: Mining Disinformation Narratives from Fact-Checking Articles [185] (EMNLP 2025, CORE A*, 140 ministerial points)
This paper demonstrates scalable global narrative discovery from large, multi-source fact-checking corpora, enabling topic-resolved and temporal analyses of narrative dynamics.

DiNO: Disinformation Narrative Observer (under review)
This paper extends disinformation narrative mining to news outlets. It introduces an end-to-end method for extracting narratives from news articles and evaluates the extracted narratives with a reference narratives dataset, showing clear improvements over competitive narrative mining approaches.

These publications provide the methodological and empirical backbone of the thesis. EUDisinfoTest supplies a benchmarked evaluation lens for narrative-level classification, while DiNaM provides scalable narrative mining machinery that is later extended to outlet data, with DiNO extending the extraction and evaluation pipeline directly to news articles.

Overview and Mapping

For ease of reference, Table 1.1 summarises how the research objectives, main chapters and contributions relate to one another, and indicates where contributions correspond to peer-reviewed publications.

Objective	Chapter	Contributions	Publication(s)
O1	Chapter 3	C3	[184, 185]
O2	Chapter 4	C1, C2	[184]
O3	Chapter 5	C4, C5	[185]
O4	Chapter 6	C6, C7	under review

Table 1.1: Mapping between research objectives, chapters and main contributions, with related publications.

1.4 Other Publications and Scientific Background

I have co-authored several publications and participated in research projects that shaped the methodological toolkit used throughout this dissertation.

1.4.1 Publications beyond the dissertation core

The publications listed below are not the main backbone of the dissertation chapters, yet they contributed key ideas, experimental skills, and reusable components that directly supported my work on narrative-based disinformation analysis.

Workshop papers. Taking part in international workshops helped me develop strong, systematic testing practices, which later proved essential for building and validating narrative benchmarks and narrative-mining methods.

- **Persuasion techniques detection (SemEval).** In SemEval-2023 Task 3, we developed the DSHacker system for multilingual detection of genres and persuasion techniques, with a particular focus on data augmentation via machine translation and text generation. [122]
- **Check-worthiness detection (CLEF CheckThat!).** In the CheckThat! 2023 lab, we addressed check-worthiness detection in a multilingual and multigenre setting, using a unified multilingual modelling approach and augmentation based on generative models. [122]

Related peer-reviewed publications on disinformation classification. I also co-authored publications that complement this dissertation by persuasion techniques and malicious intent as explicit signals for improving disinformation classification, particularly in zero-shot settings.

- **Persuasion-Augmented Chain-of-Thought (ACL, CORE A*, 200 ministerial points).** We introduced a persuasion-augmented reasoning framework for detecting fake news and social media disinformation in zero-shot settings, and released additional datasets to evaluate generalization beyond model knowledge cutoffs. [123]
- **MALicious INTent (MALINT): Dataset and Inoculating LLMs for Enhanced Disinformation Detection (EACL 2026, CORE A; 140 ministerial points).** We introduce MALINT, the first expert-annotated English dataset that captures disinformation together with malicious intent. We benchmark 12 language models on binary and multi-label intent classification and propose intent-based inoculation (intent-augmented reasoning) to improve disinformation detection. [1]

Peer-review publications outside the disinformation domain. I also contributed to research on general NLP methodology and information extraction. These publications strengthened the technical foundation of my doctoral work, especially in representation learning and evaluation.

- **Representation learning and metric-learning-inspired objectives.** We studied SoftTriple-style losses for supervised fine-tuning of language

[1] Arkadiusz Modzelewski, Witold Sosnowski, Eleni Papadopulos, Elisa Sartori, Tiziano Labruna, Adam Wierzbicki, Giovanni Da San Martino. Accepted to EACL 2026, to appear.

models and proposed a loss formulation aimed at improving classification robustness, with particular benefits in low-resource regimes. [183]

- **Fine-grained information extraction benchmarks.** I co-authored work introducing a benchmark for fine-grained recognition of complex food entities in recipes, aimed at supporting ingredient substitution. [102]

1.4.2 Research projects and collaboration context

My doctoral research was developed in an applied research environment, with close ties to disinformation practitioners and externally funded projects.

National and European projects.

- **InfoTester (NCBR).** I participated in work on AI methods for detection of disinformation, aimed at supporting analysts and volunteers in credibility assessment workflows.
- **Infotester4Education (Erasmus+).** I contributed to developing educational methods and digital tools that support media literacy and help users recognize and counter disinformation.
- **Link4Skills (Horizon Europe).** Within Link4Skills, I worked on an AI-driven chatbot concept addressing migration-related disinformation, with an emphasis on responsible design.

1.5 Thesis Structure

The thesis is organized as follows. Chapters 2–6 address the research objectives introduced above, moving from conceptual foundations through benchmarking to narrative mining on fact-checking corpora and news outlets, before Chapter 7 synthesises the overall findings.

- **Chapter 2** reviews related literature on disinformation and disinformation narratives, persuasive and rhetorical strategies, and the use of LLMs for disinformation narratives classification and analysis. It concludes by identifying the key knowledge gaps that motivate the subsequent chapters.

- **Chapter 3** presents the conceptual and operational framework for disinformation narratives developed in this thesis. It formalises the notions of disinformation narratives, disinformation narrative classification and disinformation narrative mining.
- **Chapter 4** introduces EUDisinfoTest and reports experiments benchmarking a range of LLMs on disinformation narrative classification. It describes the construction of the benchmark, the experimental setup and evaluation protocol, and analyses model performance, including the impact of logos, ethos and pathos variants on classification accuracy.
- **Chapter 5** presents DiNaM, the proposed framework for mining disinformation narratives from fact-checking articles. It describes the pipeline components (article filtering, LLM-based extraction and verification, clustering, narrative summarisation), evaluates the method analyzing alignment with EUDisinfoTest narratives, and provides analyses of mined narratives.
- **Chapter 6** presents DiNO, a framework for mining disinformation narratives directly from news outlet datasets. It describes the end-to-end pipeline (false-segment matching with DiNO-sgm, entailment verification, clustering, and narrative summarisation), evaluates DiNO by measuring alignment with EUDisinfoTest expert narratives, and complements the quantitative results with qualitative analyses of the narratives produced.
- **Chapter 7** summarises the main findings and contributions of the thesis, revisits the research objectives, and reflects on their implications for disinformation research. It also discusses limitations and future work.

Chapter 2

Literature Review

2.1 Disinformation and Disinformation Narratives

2.1.1 Disinformation: Definitions and Typologies

The notion of disinformation has been conceptualised in multiple, partly overlapping ways across communication studies, political science and policy reports. A widely used starting point is the framework of information disorder, which distinguishes between misinformation (false information shared without intent to harm), disinformation (false information shared with intent to cause harm) and malinformation (genuine information used to inflict harm) [201]. Within this framework, disinformation is characterised not merely by inaccuracy but by strategic intent and potential for harm, typically in political, economic or social domains. Subsequent work, including policy analyses and empirical studies of online manipulation, adopts similar distinctions while emphasising that intent is difficult to observe directly and must often be inferred from patterns of behaviour, coordination or context [32].

In public and scholarly debate, the umbrella term fake news has become prominent, initially used to describe fabricated news-like content mimicking journalistic formats [103]. Moreover, empirical research on the 2016 US election demonstrates that fabricated news stories can reach large audiences and be widely shared on social media [15], while studies of news diffusion on Twitter show that false content can spread faster, farther and more broadly than true content [197]. Another strand of research situates disinformation within a broader network propaganda environment, in which long-standing partisan media ecosystems and attention dynamics enable the systematic amplification of misleading or false content, particularly in polarised political contexts [29].

Policy-oriented work has further clarified disinformation in the context of democracy and platform governance. EU and platform policy analyses commonly define disinformation as information that is verifiably false or misleading, created and spread either for profit or to deliberately mislead the public, and with the potential to cause public harm [32]. Recent normative research places this problem in the wider changes of the digital public sphere, arguing that weakened shared sources of authority and the design of online platforms make strategic manipulation easier and more effective [111, 29].

2.1.2 Framing and Narratives

Research on political communication and social movements has long emphasised that the impact of a message depends not only on what is said but also on how it is framed and narrated. Framing theory conceptualises frames as ways of selecting and highlighting certain aspects of reality to promote a particular problem definition, causal interpretation, moral evaluation and treatment recommendation [59]. This influential definition has been widely adopted in media and communication studies and provides a useful analytical lens for understanding how issues such as migration, public health or armed conflict are constructed in news coverage and political discourse.

Narratives on the other hand can be understood as "selective depictions of reality across at least two points in time that can include one or more causal claims, and are generalizable and can be applied to multiple situations, as opposed to specific stories [55]. This definition highlights a key difference with framing: generalizability. Frames are ways of presenting and interpreting an issue in a particular communicative context, for example by emphasising certain causes, responsibilities, or solutions [59]. They can be reused, but they do not need to travel across cases in the same way.

Narratives are more explicitly designed to travel. They offer a broader storyline that can be applied to different events and contexts. In other words, a narrative is not only an interpretation of one situation, but a more general explanation that can be repeated, adapted, and used again elsewhere. This is why narratives are especially useful for studying how political actors create consistent messages across time and across different issues [55].

2.1.3 Disinformation Narratives

While academic work has historically focused on disinformation at the level of individual items (articles, posts, claims), policy and practitioner communities increasingly emphasise disinformation narratives as key units of analysis. Fact-checking organisations and monitoring initiatives such as EDMO routinely group fact-checks into narrative categories that capture recurring patterns of false or misleading claims, for example narratives about electoral fraud, foreign interference, or alleged threats posed by migrants [146]. Reports on disinformation narratives during recent European elections define narratives as the overarching messages that emerge from multiple fact-checkable pieces of content and recommend narrative-level monitoring as essential for understanding and countering disinformation campaigns [146]. Recent policy analyses argue that item-level interventions are insufficient and that monitoring and response need to operate at the level of campaigns and narratives, especially in the context of elections and other high-stakes civic processes [132].

In this practice-oriented context, disinformation narratives are typically described as coherent thematic clusters of false claims that share core propositions, actors and frames. For instance, a narrative might assert that "the electoral process is rigged by domestic elites" or that "international organisations are secretly controlling national governments", with many individual pieces of content instantiating these themes in different countries and contexts. Monitoring reports and empirical studies emphasise that such narratives are often persistent over time, adaptive to new events and capable of crossing platform and language boundaries [146, 64, 63]. They therefore provide a more stable and interpretable lens on disinformation ecosystems than isolated items.

Despite the increasing use of narrative-level monitoring in practice, broadly defined computational characteristics that uniquely define disinformation narratives are still lacking. The definitions used by practitioners highlight key intuitions but do not specify how these properties should be operationalized for automatic identification, clustering, or large-scale exploration. This gap motivates the formal task formulations and framework for disinformation narrative mining presented in Chapter 3.

2.2 Persuasion and Classical Rhetorical Strategies

2.2.1 Classical Rhetoric: Logos, Ethos, Pathos

Classical rhetorical theory identifies three core persuasive appeals: logos, ethos and pathos [97]. Logos refers to the appeal to reason through arguments, evidence and inferences; it concerns the internal coherence of a speech and the plausibility of the claims advanced. Ethos denotes the perceived character of the speaker, including signals of competence, honesty, benevolence and shared values that make an audience more willing to trust the message. Pathos captures appeals to the emotions and values of the audience, such as fear, anger, hope or pride.

Later research in rhetoric takes a fresh look at these ideas. Ethos is seen as something a speaker builds through language, not something they simply have. For example, speakers can build ethos through discursive self-presentation—positioning themselves in the account they tell, explicitly marking their stance, and (at times) strengthening credibility by acknowledging uncertainty and limits to what they know [17]. Logos is understood as more than strictly correct logic. It also includes forms of reasoning that sound persuasive in everyday political talk, such as arguments that seem logical, cause and effect stories, and analogies [195]. Pathos is linked to how speakers appeal to shared values, group identity, and familiar emotions to make certain actions seem justified [204].

Despite their historical origin, logos, ethos and pathos remain widely used analytical categories in political communication, social movement studies and media analysis. They provide a compact, interpretable vocabulary for describing how a message persuades, whether it foregrounds argumentative structure, leans heavily on the authority or trustworthiness of sources, or primarily mobilises emotions and identity commitments. In this thesis, these appeals are used as a simple and transparent handle on persuasive style, which is later operationalised in the construction of rhetorical variants.

2.2.2 Rhetorical Strategies in Disinformation

Research on mis- and disinformation has documented how these persuasive mechanisms are systematically exploited in contemporary information campaigns. Disinformation rarely appears as isolated false statements; instead, it is embedded

in broader narratives that organise events, actors and causal relations in emotionally and morally charged ways [169]. Such narratives not only shape how audiences interpret specific claims, but also contribute to wider patterns of mistrust and democratic erosion.

Ethos-focused strategies are very prominent. Disinformation campaigns invest in constructing alternative ecosystems of sources that present themselves as independent truth-tellers exposing hidden realities, including partisan news outlets, influencers and pseudo-expert organisations [71]. Survey and case-study research on misinformation and democratic decline suggests that people are more likely to believe false or misleading claims when they distrust science and public institutions. This also suggests that alternative outlets both exploit and may deepen existing scepticism toward mainstream journalism, scientific authorities, and fact-checkers. [13, 111]. In geopolitical contexts, foreign state media and proxy outlets such as RT and Sputnik present themselves as providers of alternative perspectives correcting alleged Western bias; content analyses document recurring disinformation narratives about security, migration and the war in Ukraine, while audience studies show that these outlets reach substantial cross-national publics [36, 98].

At the level of logos, disinformation often mimics rational argumentation rather than abandoning it: through the use of argumentative operators and discourse connectives, texts can steer readers toward a preferred interpretation and imply causal links that support a politically aligned reading of events [130].

The emergence of generative language models adds a further layer. Risk assessments note that these systems can produce fluent, persuasive text that can mimic appeals to logos, ethos and pathos, potentially lowering the cost and increasing the scale of influence operations [75]. At the same time, language models are being explored as tools for detecting and countering disinformation [44, 147]. This motivates a further question for this thesis: how robust are such systems when rhetorical style is intentionally manipulated?

2.3 LLMs for Disinformation Narrative Classification

2.3.1 Item-Level Disinformation and Fake-News Classification

Early computational work on disinformation largely treated classification as an item-level problem: classifying individual posts or articles as Real/Fake, or verifying short claims against trusted information. Surveys commonly organise this literature into content-based methods and approaches that additionally leverage social context and user signals [179]. Content-based detectors use lexical and stylistic cues (e.g., n-grams, syntax, sentiment, psycholinguistic markers) to capture characteristic language patterns of low-credibility outlets [149], later extended with topic features and neural encoders [179, 91]. Context-aware models incorporate propagation structure, engagement dynamics and user information; empirical work shows systematic differences in how true and false stories diffuse online [197].

With deep learning, detectors increasingly model text and social context jointly; surveys cover CNN/RNN text encoders, graph neural networks over interaction graphs, and multimodal hybrids [91]. In parallel, automated fact-checking focuses on classifying claim veracity using retrieved evidence and often includes matching claims to existing fact-checks[80], while recent work moves toward document-level pipelines via claim extraction and decontextualisation [54].

Despite progress, item-centric methods often generalise poorly across topics, platforms and time, consistent with reliance on dataset- or platform-specific lexical and diffusion cues [91]. Labels are also costly and judgement-dependent, introducing subjectivity into ground truth [80]. Finally, many benchmarks use coarse veracity labels and do not capture how items relate to broader narratives or rhetorical strategies, limiting analysis beyond isolated predictions.

2.3.2 From Supervised Models to LLM-Based Classification

The advent of large pre-trained language models has spurred a growing body of work that uses them as detectors or assistants within mis- and disinformation pipelines. Rather than training task-specific classifiers, many studies reframe classification as an in-context learning problem: a general-purpose model is prompted in zero-shot form, or given a small number of labeled examples in the prompt

(few-shot), to judge whether a post, headline, or claim is misleading, fabricated, or trustworthy. This shift relies on LLMs' broad pretraining and their ability to adapt to task descriptions and provided examples expressed in natural language, without updating model parameters [35].

Several lines of work examine LLMs explicitly as detectors of deceptive content. Experimental setups often involve prompting an LLM to classify human- and model-generated texts as genuine or deceptive and comparing performance to specialized detectors [110, 43]. Beyond direct classification, LLMs are increasingly used as data tools, for example to collect, normalise and annotate misleading content, generate synthetic counterfactual examples, or support experts in curating new datasets of disinformation narratives [44, 104].

At the same time, LLMs exacerbate existing challenges and introduce new ones. On the generation side, they can be prompted to produce large amounts of highly convincing synthetic disinformation, sometimes tailored to specific audiences, issues or communication channels [43, 75]. On the classification side, LLMs are prone to hallucinations and inconsistent reasoning [94], and benchmarks such as TruthfulQA and MMLU highlight systematic failures under misleading prompts as well as remaining knowledge gaps in specialised domains [107, 90]. There is also growing evidence that LLM judgements can reflect social and political biases embedded in their training data [73]. These issues are particularly salient when models are used as judges for other models' outputs or for human-authored texts [209]. Overall, the literature portrays LLMs as powerful but potentially unreliable classification components that need to be carefully calibrated, monitored and, in many applications, combined with verification mechanisms and human oversight [44, 80].

2.3.3 Narrative-Level Disinformation Classification

Parallel to developments in item-level classification, research in political communication, computational social science and natural language processing increasingly emphasises that mis- and disinformation should be analysed not only as isolated statements, but as part of broader narratives. These narratives organise events, actors and causal relations into coherent storylines that can be adapted to different audiences and contexts [30]. Policy and practitioner communities similarly describe disinformation in narrative terms, for instance by cataloguing key narratives in pro-Kremlin disinformation or by tracking narrative repertoires around elections and other political events [64, 146]. Such efforts typically rely on expert analysis, with

limited automation, and they rarely provide explicit, machine-readable narratives.

Computational approaches that move beyond individual posts or claims take several forms. Storyline and narrative extraction methods segment news streams into chains of related events or storylines that link articles over time [39, 198]. More recent work represents narratives as structured micro-stories over entities and actions, providing compact, relational representations of who does what to whom [18]. In the context of mis- and disinformation, such methods have been used to analyse conspiratorial and extremist narratives, to map how competing narratives diffuse across platforms and countries, and to compare narrative repertoires between disinformation-prone and reputable sources [113, 36]. Related strands of work focus on framing, persuasion techniques and rhetorical strategies in online news and social media, including multilingual benchmarks for genre, framing and persuasion-technique classification and models that explicitly operationalise classical rhetorical appeals such as logos, ethos and pathos [151, 148]. These studies show that narratives and frames are central to how disinformation operates, but their computational models typically predict topic, frame or persuasion labels rather than explicitly assessing the credibility of narrative-level statements.

Only a small number of works define disinformation narratives as explicit textual units. Some studies organise fact-checked claims or posts into broader narrative categories for monitoring and reporting [146], while others combine narrative groupings with stance annotation to analyse support and spread within and across narratives [31]. Others use narrative templates to structure qualitative analyses of disinformation campaigns, for example around migration or public health, without turning them into formal benchmarks for automated systems [128, 155]. Systematic narrative-level resources that provide narratives with clear credibility labels, that decouple narrative content from particular surface realisations and that explicitly account for rhetorical strategies such as logos, ethos and pathos remain rare. Existing datasets and benchmarks tend either to lack explicit narratives, to mix item-level and narrative-level labels or to ignore how classical rhetorical appeals can modulate perceived credibility even when the underlying message is unchanged. These gaps motivate the development of narrative-centric and rhetoric-aware evaluations of LLMs' ability to distinguish between credible and misleading narratives in the remainder of this thesis.

2.4 Computational Narrative Mining and LLMs

2.4.1 Narrative Mining Methods

Within this broader landscape, a small but growing set of methods explicitly frames its goal as narrative mining rather than claim classification or topic detection. These approaches are closest to the framework developed in this thesis.

First, general-purpose narrative mining systems are largely veracity agnostic. CANarEx defines narratives as contextualised Entity–Verb–Entity triples and proposes a pipeline that extracts such micro-narratives, clusters them and analyses their evolution [18]. Relatio models political and economic narratives as low-dimensional semantic factors learned from document embeddings and applies this to legislative debates and news coverage [20]. Both demonstrate how compact narrative representations can be constructed and analysed at scale.

Second, institutional monitoring practice treats narratives as the main reporting unit, but it typically identifies them as candidate signals whose status is unknown and must be checked before any corrective communication or intervention is planned. Social listening guidance makes this explicit by recommending rumour logs that capture early narrative signals and then requiring verification and assessment before turning them into situational insights and engagement outputs [196]. WHO and UNICEF formalise the same approach in infodemic insights reporting, where narratives are first collected and then validated before recommending responses [142]. In RESIST 2, early warning narratives are scheduled to be judged whether information is likely false before selecting a response, which separates classification from response choice in day-to-day practice [174].

2.4.2 Disinformation Narrative Mining Methods

Third, some approaches more directly analyse disinformation narratives. One line organises online content around disinformation campaign narratives and annotates stance, enabling joint analysis of narrative diffusion and support [31]. Other work contrasts disinformation and trustworthy news in terms of narrative structure, comparing LLM-based, knowledge-graph-based and hybrid pipelines and showing that disinformation often exhibits more fragmented and ambiguous storylines [113]. These studies treat narratives as units of analysis and provide insights into how disinformation narratives differ from those in reputable sources, but they do not propose end-to-end methodologies explicitly aimed at mining disinformation

narratives from large, heterogeneous corpora.

Lastly, there is substantial work on components of a disinformation narrative mining pipeline. Claim matching and retrieval methods detect whether new content repeats previously fact-checked claims [177]. Fake-news classification models classify individual posts or articles as true, false or mixed using linguistic, social and multimodal features [179, 91]. Genre, frame and persuasion classification methods provide higher-level thematic and rhetorical structure [151]. Together, these strands deliver many ingredients a narrative mining framework would need: alignment of new content with existing fact checks, robust item-level veracity signals and tools for grouping and characterising related content.

What is missing, however, is a coherent pipeline that connects these ingredients into a method explicitly designed to discover and represent disinformation narratives as reusable, clause-like patterns grounded in fact checking practice and linked to many underlying instances. Existing work either focuses on general narratives without veracity, as in CANarEx and Relatio, analyses disinformation narratives on relatively small or manually curated sets, as in campaign focused corpora and narrative structure comparisons, or contributes isolated components such as claim matching or fake-news classification. To the best of my knowledge, there is no prior work that shows how to integrate these elements into a scalable, LLM assisted methodology for mining disinformation narratives from large corpora of fact checks and news.

2.5 Knowledge Gaps and Thesis Positioning

The literature reviewed above reveals several gaps that motivate the research objectives and contributions of this thesis.

2.5.1 Gap 1: Task Foundations for Mining Disinformation Narratives

Conceptual and policy work defines categories such as disinformation, misinformation and malinformation [201]. Narrative-oriented research shows how communication organises events and actors into stories that shape interpretation [150]. Fact checking practice also groups recurring false claims into narrative categories, but there is still no simple, shared definition of disinformation narratives for computational use, nor clear task definition for disinformation narrative-level

mining.

This gap motivates Objective O1 and is addressed in Chapter 3, which identifies core characteristics of disinformation narratives and formalises narrative classification and narrative mining as explicit tasks.

2.5.2 Gap 2: Narrative-Level, Rhetoric Aware Evaluation of LLMs

Narrative monitoring workflows typically identify candidate narratives as early signals and then include an explicit step to verify and assess these narratives before any countermeasures are designed or deployed [142, 174]. Recent studies show that LLMs can both generate and help detect misleading content [44], which makes them plausible components of this verification step. However, they are most often evaluated on item-level benchmarks rather than on narrative-level phenomena.

This limitation is amplified by rhetoric. The same narrative content can be expressed through different rhetorical strategies, shifting emphasis across reasoning, credibility cues, and emotional appeal. Computational work on logos, ethos, and pathos is growing, but it is rarely connected to disinformation narratives and is seldom used to construct controlled rhetorical variants for evaluation. As a result, we lack two resources that monitoring systems require: (i) datasets of ground truth disinformation narratives that support evaluation of narrative-level outputs, and (ii) rhetoric aware benchmarks that test whether LLM narrative-level decisions remain stable under systematic rhetorical variation.

This gap motivates Objective O2 and is addressed in Chapter 4, which introduces EUDisinfoTest, a narrative-level, rhetoric-aware benchmark.

2.5.3 Gap 3: Global Disinformation Narrative Mining

Existing narrative mining methods extract topics, storylines or event structures from large text corpora, but usually ignore veracity and are not grounded in fact checking archives [18, 20]. Fact checking datasets, in turn, contain labelled claims and descriptions, yet there is no scalable way to turn this material into a coherent, machine readable global catalogue of disinformation narratives.

This gap motivates Objective O3 and is addressed in Chapter 5, which presents DiNaM, a framework that mines disinformation narratives from large fact-checking corpora and organises them into a global catalogue.

2.5.4 Gap 4: Local, Outlet Level Disinformation Narratives Mining

Even with a global narrative catalogue, current research says relatively little about how disinformation narratives are used inside specific media outlets. Studies of disinformation prone and mainstream media document systematic differences in how issues are framed and narrated, but do not provide general methods for linking outlet content to an explicit narrative map or for comparing outlets over time.

This gap motivates Objective O4 and is addressed in Chapter 6, which introduces DiNO, a framework that projects the global catalogue onto outlet-specific news corpora and derives outlet-level profiles of disinformation narratives.

Chapter 3

Definitions and Task Formulations

Chapter 1 and Chapter 2 show a shift in disinformation research from item-level units (posts, articles, individual claims) towards narrative-level units that capture recurring disinformation themes and templates. In practice, candidate narratives are proposed from large corpora and then assessed: given a narrative, analysts must decide whether it expresses a disinformation narrative. This creates a need for methods that can support narrative-level assessment, including evaluation of whether a proposed narrative is disinformation and whether such judgements remain reliable under variation in rhetorical framing. This motivates the first task studied in this thesis, disinformation narrative classification: given a narrative, decide whether it expresses a disinformation narrative or a credible narrative.

Narrative identification from corpora still relies heavily on human effort, limiting scalability and systematic monitoring. When automated methods are used, they are usually general purpose clustering or narrative discovery and do not enforce core requirements for disinformation narratives, especially grounding in verifiably false content and linkage to concrete supporting instances. This motivates the second task, disinformation narrative mining.

Before formalising these two tasks, we first introduce the conceptual backbones that they rely on: definitions of narrative, disinformation, disinformation narrative, and credible narrative. Together, these definitions shape how the tasks are formulated and operationalised.

3.1 Definitions

To define disinformation narratives, we first clarify what is meant by narratives and disinformation. Building on these foundations, we then introduce definitions of disinformation narratives and credible narratives for use throughout the thesis.

3.1.1 Narratives

The term narrative has different meanings depending on the context. In media analysis, narratives capture how information is framed and circulated, and which themes appear across large text corpora [38]. In computational work on narratives in news, the term is also operationalised more concretely, for example by modelling narratives as structured descriptions of spatiotemporal events, or by representing narrative components through semantic relationships between entities [96, 180, 131]. More broadly, in an effort to formalise the concept across contexts, Dennison [55] define narratives as *selective depictions of reality across at least two points that can include one or more causal claims, and are generally generalizable and can be applied to multiple situations, as opposed to specific stories.*

Building on this line of work, Nikolaidis et al. [129] propose a definition tailored to news discourse, describing narratives as *recurring, repetitive (across and within articles), overt or implicit claim that presents and promotes a specific interpretation or viewpoint on an ongoing (and often dynamic) news topic.* Since this definition is more straightforward to operationalise in computational settings, we adopt it throughout this thesis.

3.1.2 Disinformation

In research on disinformation, influential conceptualisations converge on three core elements of disinformation. First, they emphasise falsity or misleadingness: disinformation consists of information that can be shown, through evidence or fact-checking, to be false or substantially misleading rather than merely controversial or value-based [201, 61]. Second, they require a potential for public harm: the information must pose, or be likely to pose, a risk to democratic processes, public health, security or other public goods [201, 61]. Third, they stress intentional deception: the creation or dissemination of the information is undertaken with an intention to mislead, manipulate, or secure economic or political advantage, which distinguishes disinformation from the inadvertent sharing of false content (misinformation) [201, 32].

In line with this literature, this thesis adopts the definition used by the European Commission's High Level Expert Group: disinformation *includes all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit* [53].

3.1.3 Disinformation Narratives

Attempts to define disinformation narratives have largely come from practitioners. The European Digital Media Observatory (EDMO) describes a disinformation narrative as "the clear message that emerges from a consistent set of contents that can be demonstrated as false using the fact-checking methodology" [146]. EUvsDisinfo, the EU's flagship project for countering pro-Kremlin disinformation, similarly describes a disinformation narrative as an "overall message, communicated through texts, images, metaphors, and other means", which serves as a "template for particular stories" that can be adapted to specific audiences and contexts [64, 63].

Building on these practitioner definitions, and informed by the broader literature on disinformation and narratives, we define a disinformation narrative as: *an adaptable claim that presents a specific interpretation or viewpoint on an ongoing topic, typically expressed through multiple related instances of false, inaccurate, or misleading content.*

3.1.4 Credible Narratives

A complementary notion to disinformation narratives is that of a counter narrative. In work on online harm mitigation, counter narratives (often discussed under the closely related term counterspeech) are defined as informed textual responses that intervene against harmful content by refuting, correcting, or reframing it, ideally in a non-aggressive way [194, 68, 33]. This perspective emphasises the function of the message (to counter a harmful claim or theme), and has been operationalised both through datasets of hate-speech/counter-narrative pairs and through models for counter-narrative generation, including approaches that explicitly ground responses in external knowledge [46].

Strategic Communications (StratCom) uses the closely related term credible narrative to foreground the constraints under which such countering should take place. StratCom emphasises "strategic impact through credible narrative" and characterises it as an honest attempt to frame what is at stake without resorting to propaganda or spin [203]. On this view, credibility is not only a matter of factual accuracy, but also of communicative integrity: the message should be truthful and grounded in credible evidence.

To keep terminology consistent, we use the term credible narrative throughout the thesis. We understand credible narratives as: *an evidence-grounded claim formulated in response to an identified disinformation narrative, that negates, corrects, or reframes its core claim(s) without employing deception or manipulation.*

Credible narratives in this thesis need not share all formal properties of disinformation narratives. In particular, they are not required to be adaptable across multiple contexts. Instead, they serve as context-specific, evidence-supported responses to identified disinformation narratives, rather than as reusable templates.

3.2 Task Formulation

This section formalises the two core computational tasks addressed in the thesis. The first, disinformation narrative classification, assumes narratives are given and focuses on assessing their credibility. The second, disinformation narrative mining, addresses how such statements and their supporting instances can be recovered from large text corpora.

3.2.1 Disinformation Narrative Classification

Building on this view of disinformation narratives, this subsection formalises the task of disinformation narrative classification. In contrast to item-level disinformation classification, where the unit of analysis is typically a post, article or claim, the task here is to decide whether a narrative is a credible narrative or disinformation narrative.

Let $\mathcal{N}$ denote the set of narratives, and let $y(n)$ be the ground-truth credibility label associated with each $n \in \mathcal{N}$, based on the ddefinitions of disinformation narratives and credible narratives in Sections 3.1.3 and 3.1.4. The goal is to decide, for any given narrative, which of the two classes it belongs to.

Formally, the disinformation narrative classification problem is to approximate a decision function

$$f : \mathcal{N} \rightarrow \{Credible, Disinformation\},$$

such that, for each $n \in \mathcal{N}$, the prediction $f(n)$ is as close as possible to the ground-truth label $y(n)$. Different classification systems may implement f in different ways (for example through supervised classifiers, zero-shot decision

rules or hybrid human-machine workflows), but they all operate on the common abstraction of narrative-level inputs and binary credibility labels.

3.2.2 Disinformation Narrative Mining

Disinformation narrative classification assumes that candidate narratives are already available and focuses on assessing their credibility. In practice, however, narratives must first be recovered from large, heterogeneous text collections. We model disinformation narrative mining as the task of deriving narrative-level representations from corpora.

Let $C = \{c_1, c_2, \ldots, c_n\}$ denote a collection of content items, where each c_i is a text (for example, a fact-checking article or a news article) that may contain instances of false or misleading information. The aim is to derive from C a set of disinformation narratives

$$\mathcal{N}^{Dis} = \{N_1, N_2, \ldots, N_k\}.$$

Disinformation Narrative Mining Framework

Compared to generic topic discovery, narrative mining in the disinformation setting is designed to reflect the properties of disinformation narratives introduced in Section 3.1.3. In particular, because disinformation narratives are grounded in false, inaccurate, or misleading content, the mined outputs should be based on content that can be verified as false or misleading (**incorrect basis**). In addition, while a disinformation narrative may appear in a single item, it is typically expressed through multiple related instances; accordingly, the mining process should prioritise recurrence as evidence that an extracted message generalises beyond an isolated mention (**multiple examples**). Finally, because a narrative constitutes a claim-level interpretation that can surface across different realisations of the same message, the mined representation should capture a stable core that persists across varied phrasings and contexts (**adaptability**). Together, these considerations motivate a three-stage pipeline: first isolate false-information instances, then group instances that express the same underlying message, and finally derive an abstract narrative representation from each group.

Stage 1: Identifying false information (Incorrect basis). The first stage isolates concrete expressions of false or misleading content in C. For each content

item $c_i \in C$, a set of spans

$$S_i = \{s_{i,1}, \ldots, s_{i,K_i}\}$$

is extracted that express verifiably false or misleading claims (for example, false claim formulations in fact-checking articles or segments of news text aligned with verified false claims). The union

$$S = \bigcup_{c_i \in C} S_i$$

collects all identified false information instances and ensures that subsequent steps operate only on text with an incorrect basis.

Stage 2: Clustering false information (Multiple examples). In the second stage, the set S is partitioned into clusters of semantically related instances.

$$K = \{K_1, \ldots, K_k\}, \quad \text{where } K_q \subseteq S.$$

Each cluster K_q is intended to correspond to a candidate disinformation narrative emerging from multiple, related instances; clusters of size one can be discarded to enforce the multiple-examples requirement.

Stage 3: Deriving narrative representations (Adaptability). Each cluster $K_q \in K$ is summarised into a narrative N_q capturing the shared message across its instances. Collectively, these narratives form $N^{Dis} = \{N_1, \ldots, N_k\}$, with each N_q supported by the instances in K_q.

This three-stage framework provides the methodological backbone for the two concrete narrative-mining pipelines developed later in the thesis: Chapter 5 applies it to fact-checking articles, and Chapter 6 adapts it to news outlet corpora.

Chapter 4

EUDisinfoTest: Benchmarking LLMs for Disinformation Narrative Classification

4.1 Introduction

The previous chapter defined disinformation narratives and formalised disinformation narrative classification as the task of deciding whether a given narrative is credible or disinformative (Sections 3.1.3 and 3.2.1).

This chapter moves from theory to evaluation by examining how contemporary LLMs behave when applied directly to the task of disinformation narrative classification. To this end, it introduces EUDisinfoTest, a dataset and benchmark of English language narratives drawn from European information environments. EUDisinfoTest contains English-language narratives in four versions: Base, Logos, Ethos, Pathos. Additionally, the dataset contains both disinformation and credible narratives for all four versions, enabling a direct comparison of LLM responses to disinformation versus credible narratives and exploring the impact of different persuasion strategies on model effectiveness. Importantly, the rhetorical variants are treated as controlled persuasive rewrites of an underlying Base narrative: they preserve the core propositional content and credibility label while modifying rhetorical framing (Logos/Ethos/Pathos).

Within the broader thesis, EUDisinfoTest plays a dual role: it serves as (i) a dedicated testbed for assessing LLMs as narrative-level detectors and for analysing how their judgements are influenced by rhetorical variants (via controlled rhetorical variants of the same underlying narrative), and (ii) a reference set of disinformation narratives against which the mining pipelines introduced in Chapters 5 and 6 can later be aligned. The present chapter focuses on the evaluation role. It (i) establishes baseline results for a range of open and commercial LLMs on the disinformation narrative classification task, using a unified zero-shot prompting and

evaluation protocol and comparing them to a small-scale human study on a subset of Base narratives, and (ii) analyses how performance varies across topics and rhetorical variants (Logos, Ethos, Pathos), with a particular focus on systematic shifts in model behaviour induced by different rhetorical strategies.

To ensure full reproducibility of our results and to facilitate further exploration, we are making the EU DisinfoTest benchmark publicly available [1].

4.2 The EUDisinfoTest Benchmark

EUDisinfoTest is a benchmark of narrative statements grounded in European information environments.

Concretely, each entry is annotated with (i) a credibility label (Credible vs. Disinformation), (ii) a topical domain (e.g. migration, climate, public health, the war in Ukraine), and (iii) a form indicator specifying whether the statement is a Base narrative or a rhetoric-enhanced variant (Logos, Ethos, Pathos). Rhetorical variants are available for a subset of Base narratives and are constructed to preserve the underlying propositional content while altering persuasive emphasis. Accordingly, they are treated as evaluation probes of rhetorical sensitivity rather than as independent narratives in the strict sense of the definition (Chapter 3).

4.2.1 Credible vs Disinformation Narratives

Each narrative is represented as a statement $n \in \mathcal{N}$, and the benchmark assigns to every statement a credibility label

$$y(n) \in \{Credible, Disinformation\}.$$

The label indicates whether a given narrative expresses a credible narrative or a disinformation narrative.

4.3 Methodology

The methodology section describes our approach to collecting and annotating disinformation narratives in Europe. The overview of the process is detailed in

[1] https://github.com/wsosnowski/EUDisinfoTest

Figure 4.2. The methodology implements a Human-in-the-Loop approach to data collection by integrating a team of four expert annotators, with advanced AI tools such as GPT-4 and Microsoft Copilot Pro. It is important to point out that two of our experts have experience working for debunking organizations certified by the International Fact-Checking Network (IFCN).

The primary motivation for using an AI-optimized process was to efficiently meet the specific requirements of collecting narratives in four distinct forms: Base, Logos, Ethos, and Pathos. Manual methods, which are typically less adaptable, would face challenges in efficiently meeting these diverse requirements. In addition, recent studies suggest that LLMs following instructions, as well as HITL approaches to data collection, perform comparably to humans [104, 205]. Importantly, although AI tools are utilized in our methodology, all outputs from the AI model are controlled and reviewed by human experts.

Prompt template. The use of LLMs during the data collection phase necessitated the development of a specialized prompt template, illustrated in Figure 4.1. This template is essential for designing prompts for both GPT-4 and Microsoft Copilot Pro during data collection. Moreover, it plays a critical role in zero-shot disinformation classification, discussed in Section 4.6. The template's design is influenced by the study by Lucas et al. [110], which crafted SOTA prompts for both zero-shot classifying and generating disinformation. It includes a "context" element tailored to overcome specific LLM limitations in generating disinformation, imposed through the Reinforcement Learning from Human Feedback mechanism [145]. The template also features a "content" element, tailored to specific task and an "instruction" component that encourages systematic, step-by-step reasoning in LLMs, aligning with the strategies in the studies by Bang et al. [24] and Wei et al. [202], which focus on multi-step reasoning to guide LLMs toward accurate evaluations.

4.4 Data Collection

Data source. The benchmark is rooted in an EU DisinfoLab project supported by the Friedrich Naumann Foundation for Freedom [175]. This comprehensive project analyzed the disinformation landscape across Europe, detailing 20 factsheets for key countries such as Germany, France, and Italy. These nations collectively represent approximately 94% of the EU population, providing a robust demographic basis for

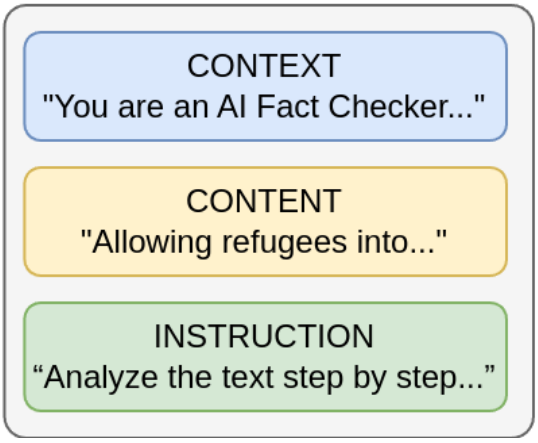

Figure 4.1: Prompt template comprising three components: (1) Context, which establishes the framework and guides the LLMs in tasks such as identification and classification; (2) Content, which includes specific data pertinent to the task; (3) Instruction, which effectively directs the actions of LLMs.

our insights. The DisinfoLab report classifies disinformation narratives into various thematic domains: *Anti-Europeanism and Anti-Atlanticism*; *Anti-migration and Xenophobia*; *Climate Change and Energy Crisis*; *Gender-based Disinformation*; *Health, COVID-19, Vaccines*; *Historical Revisionism*; *Institutional Distrust*; *Media Distrust*, and *Ukraine War* [2]. Each narrative is supported by references to primary source articles included in the report.

Broad Narratives. As described in Figure 4.2, Step 1 involved an expert identifying narratives along with source articles from the DisinfoLab report and the attached factsheets. These narratives are broad (e.g., *Pandemic is a hoax*) and tend to define a group of related disinformation narratives rather than specific narratives; thus, we refer to them as **Broad Narratives**.

Expert Verification: Two experts reviewed the extracted narratives. A narrative was included only under the consensus of two experts, following the acceptance criteria: 1) the narrative must be explicitly mentioned in the DisinfoLab report, and 2) the narrative must have an attached source article. Quality assurance

[2]The original report included a category on "Regional Tensions," which we excluded due to an insufficient number of relevant narratives. Additionally, the categories "Health Misinformation" and "COVID-19 and Vaccines" were initially separate but have been merged due to overlapping themes.

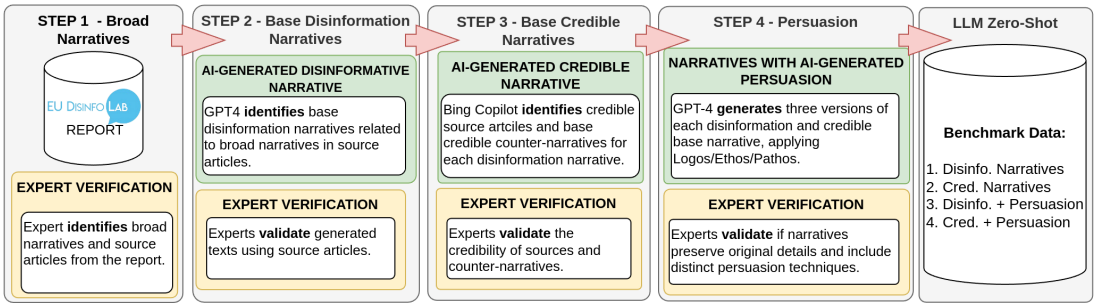

Figure 4.2: Overview of the human-in-the-loop data collection process. Broad narratives are iteratively refined into more specific base disinformation narratives, base credible narratives, and rhetorical-enhanced variants (Logos, Ethos, Pathos).

guidelines are detailed in Appendix A.1.

Base Disinformation Narratives. Prior work has shown that narratives are often organised into hierarchical taxonomies, where categories are explicitly structured by specificity: higher levels group broader narrative families, while lower levels delineate more specific instantiations or arguments. Such multi-level structures are adopted to support both (i) coarse-grained aggregation and monitoring over broad families and (ii) fine-grained, text-grounded annotation that distinguishes concrete variants without collapsing them into overly broad themes [100, 47].

In the same spirit, we adopt a two-level organisation of disinformation narratives. As previously mentioned, the Broad Narratives tend to define a group of related disinformation narratives. We therefore collected a second layer of more fine-grain, article-grounded disinformation narratives. We refer to these as Base Disinformation Narratives. Consequently, we moved to Step 2 of our methodology as shown in Figure 4.2. This step involved using GPT-4[3] to identify Base Disinformation Narratives that expanded upon the Broad Narratives and were evidenced in the source articles. Details on the GPT-4 prompts used are available in Appendix A.2.2.
Expert Verification: Each base disinformation narrative was validated by mutual agreement between two experts. The acceptance criteria required that: 1) the narrative was identified as disinformation according to the source; 2) the narrative aligned with the broader narrative. Detailed quality assurance guidelines are detailed in Appendix A.1.

[3]Using the gpt-4-0125-preview version from early April 2024.

Importantly, the two-level (Broad/Base) structure used in this chapter is conceptually analogous to hierarchical narrative modelling adopted in SemEval 2025 Task 10, which organises narratives into coarse-grained narratives and finer-grained sub-narratives [152]. In a similar spirit, Broad Narratives capture higher-level recurring narrative families, whereas Base Disinformation Narratives provide more specific, article-grounded narrative statements that instantiate these families at a more fine-grained level. In later steps of our dataset construction, we use each Base Disinformation Narrative as the underlying core message to generate multiple rhetorical variants (e.g., Logos/Ethos/Pathos-style rewrites) that preserve the propositional content while varying phrasing and persuasive framing, thereby operationalising the idea that the same narrative can be expressed in different forms.

Base Credible Narratives. For each validated base disinformation narrative collected in Step 2 (see Figure 4.2), we gathered a credible narrative to counteract the disinformation. These are referred to as **Base Credible Narratives**. We utilized Microsoft Copilot Pro[4] to identify credible sources for each validated disinformation narrative and, based on these sources, to formulate counter-narratives. The prompts used are detailed in Appendix A.2.3.
Expert Verification: Two experts reviewed the base credible narratives with their sources, reaching agreement before accepting any narrative. The acceptance criteria were as follows: 1) the source must be credible; 2) the narrative must be credible according to the source; 3) the narrative must effectively counter the disinformation narrative. Quality assurance guidelines are detailed in Appendix A.1.

Rhetorical Variants. As previously noted, disinformation narratives are often entwined with various persuasion techniques [124]. To reflect this, in Step 4 of our methodology (see Figure 4.2) we refined both credible and disinformation narratives using the rhetorical strategies of Logos, Ethos, and Pathos, as detailed in Appendix A.3. We tasked GPT-4 with enhancing a total of 300 narratives—150 credible and 150 disinformation. Each narrative was refined three times, once for each rhetorical strategy, to enrich them with these elements while preserving their original meanings. Importantly, for credible narratives, we provided GPT-4 with the narrative and the definition of rhetorical strategy, as well as the content of associated source articles that support the given narrative. This approach ensured

[4]Using the 'Precise' conversation style, which employs GPT-4 Turbo, since early April

that the enhancements made by GPT-4 did not deviate from factual accuracy and ethical considerations. For each narrative, we produced separate versions, each emphasizing one of the rhetorical strategies: Logos, Ethos, or Pathos. Detailed descriptions of the prompts and methods used with GPT-4 for each strategy are provided in Appendix A.2.

Expert Verification: Modification of each narrative required consensus between two experts, who evaluated them based on the following criteria: 1) alignment to the original content; 2) the effective and appropriate use of rhetorical strategies; and 3) for credible narratives, consistency with credible sources. The quality assurance guidelines are present in Appendix A.1.

Note that in this thesis, rhetorical variants are treated as controlled persuasive rewrites of an underlying Base narrative. They preserve the core propositional content and credibility label, while modifying rhetorical framing (Logos/Ethos/Pathos). Accordingly, rhetorical variants are used as evaluation probes of rhetorical sensitivity, not as independent narratives in the sense of the disinformation narrative definition in Section 3.1.3, unlike Base variants.

4.5 Benchmark Data

Statistics. The EUDisinfoTest dataset, detailed in Table 4.1, consists of 1,344 narratives divided into four types: Base, Pathos, Logos, and Ethos. (Broad Narratives are not included in our test).

Type	Credible	Disinformation	Total
Base	332	452	784
Pathos	97	141	238
Logos	83	86	169
Ethos	78	75	153
Sum	**590**	**754**	**1,344**

Table 4.1: Summary of Unique Values of narrative types in the Dataset, including total counts per category.

Table 4.2 shows the average word count across narrative types and credibility categories. Credible narratives are generally longer than disinformation narratives, likely reflecting their greater complexity and depth. Additionally, employing

rhetorical strategies such as Pathos, Logos, and Ethos tends to increase word count by adding more details.

Type	Credible	Disinformation
Base	30	14
Pathos	38	34
Logos	52	43
Ethos	50	46

Table 4.2: Average number of words per narrative type, comparing credible sources with disinformation.

Table 4.3 presents the total number of narratives associated with each topic.

Quality Assurance. Previous research indicates that even under strict guidance, LLMs can still generate hallucinated content [94]. To ensure the integrity and reliability of our dataset, each text underwent a thorough review by two experts following clear guidelines. Details of the process are outlined in Section 4.4. Narratives were only included after achieving consensus between the reviewers, with Cohen Kappa scores: 0.93 for Step 1, 0.78 for Steps 2 and 3, and 0.81 for Step 4. Furthermore, about 56% of the data collected by the AI systems, including both GPT-4 and Microsoft Copilot Pro, were discarded due to various misalignments. Examples of these misalignments are outlined in Table A.3. More comprehensive information can be found in the Appendix A.4.

Topic	Count
Health, Covid-19, and vaccines	258
Anti-Europeanism and anti-Atlanticism	239
Ukraine war	238
Gender-based disinformation	153
Anti-migration and xenophobia	122
Climate change and the energy crisis	114
Media distrust	100
Institutional distrust	81
Historical revisionism	39

Table 4.3: Count of Entries per Topic

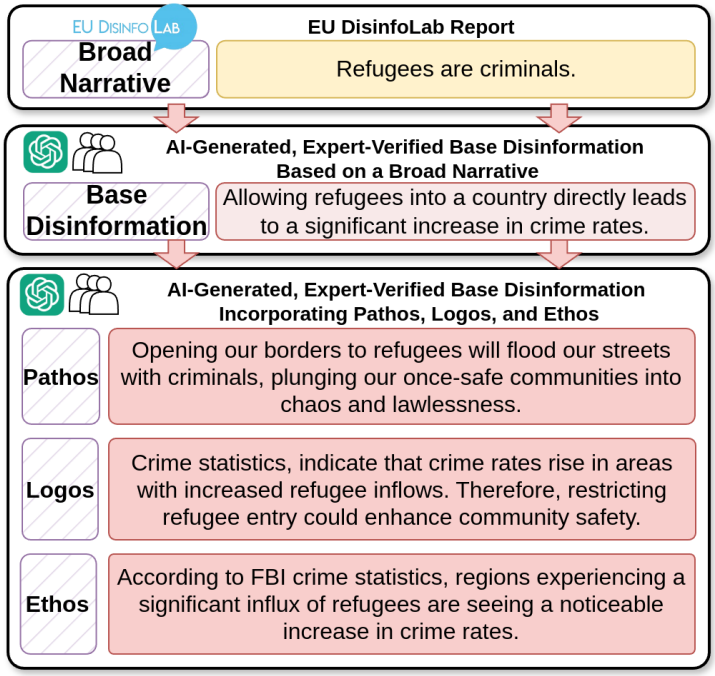

Figure 4.3: Illustration of a disinformation narrative's evolution, starting from a broad statement from the EU DisinfoLab report to refined versions incorporating Base, Pathos, Logos, and Ethos narrative types.

Examples. Examples of a Broad narrative, along with its Base, Pathos, Logos, and Ethos versions, are presented in Figure 4.3 and further in Appendix A.2.

4.6 Experiments

Zero-Shot Classification. In the field of disinformation classification, Lucas et al. [110] explored the effectiveness of prompts for zero-shot classification. We have integrated these findings into our custom prompt (see Appendix A.2). This prompt challenges models to classify narratives as either disinformative or credible, and requires them to articulate their analysis. To enhance reliability, we process each narrative through the model three times, using a majority voting mechanism to determine the final classification.

Evaluation. The EUDisinfoTest employs a set of metrics to maintain consistent performance evaluations across narrative types: Base, Logos, Ethos, and Pathos.

The general formula for the metrics is:

$$\text{Agg-}M = \frac{M_{\text{base}} + M_{\text{logos}} + M_{\text{ethos}} + M_{\text{pathos}}}{4} \tag{4.1}$$

where M stands for one of the evaluated metrics, specifically the F1-Score, TNR (True Negative Rate), or TPR (True Positive Rate). The primary metric, **Agg-F1-Score**, offers a holistic metric of accuracy. Similarly, **Agg-TNR** and **Agg-TPR** serve as auxiliary metrics, providing additional insights into the model's ability to distinguish between credible and disinformation narratives.

The F1-score measures overall classification performance by combining precision and recall into a single statistic. Given the two classes (Disinformation and Credible), a macro-averaged F1-score is used:

$$\text{F1-score} = \frac{F1_{\text{disinformation}} + F1_{\text{credible}}}{2} \tag{4.2}$$

Here, $F1_{\text{disinformation}}$ is the F1-score for the disinformation class, and $F1_{\text{credible}}$ is the F1-score for the credible class. The F1-score for each class is defined as:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3}$$

where precision is the proportion of true positive predictions among all predicted positives, and recall is the proportion of true positive predictions among all actual positives.

The True Negative Rate (TNR), also known as specificity, measures the ability to correctly identify credible narratives, which constitute the negative class. It is defined as:

$$TNR = \frac{TN}{TN + FP} \tag{4.4}$$

where TN denotes the number of true negatives (credible narratives correctly identified as credible) and FP denotes the number of false positives (credible narratives incorrectly identified as disinformation). A high TNR indicates that credible content is rarely misclassified as disinformation.

The True Positive Rate (TPR), also known as sensitivity or recall for the positive class, measures the ability to correctly identify disinformation narratives. It is defined as:

$$TPR = \frac{TP}{TP + FN} \tag{4.5}$$

where TP denotes the number of true positives (disinformation narratives correctly identified as disinformation) and FN denotes the number of false negatives (disinformation narratives incorrectly identified as credible). A high TPR indicates that disinformation content is effectively detected.

In all cases, the aggregated versions of these metrics (Agg-F1-Score, Agg-TNR, Agg-TPR), computed using the aggregation formula above, provide a variant-agnostic view of model performance across Base, Logos, Ethos and Pathos narratives.

Models We evaluated the efficacy of various LLMs. Table 4.4 summarizes the models used.

Abbr.	API Model Name	Access Details	License	Model Size
GPT-5	`gpt-5-2025-08-07`	OpenAI API 01.2026	Commercial	Not disclosed
GPT3.5	`gpt-3.5-turbo-0125`	OpenAI API 05.2024	Commercial	Not disclosed
Haiku	`claude-3-haiku-20240307`	Anthropic API 05.2024	Commercial	Not disclosed
Opus	`claude-3-opus-20240229`	Anthropic API 05.2024	Commercial	Not disclosed
Sonnet	`claude-3-sonnet-20240229`	Anthropic API 05.2024	Commercial	Not disclosed
Llama3-70b	`meta-llama/Meta-Llama-3-70B-Instruct`	DeepInfra API 05.2024	META Llama 3 Community	70B
Llama3-8b	`meta-llama/Meta-Llama-3-8B-Instruct`	DeepInfra API 05.2024	META Llama 3 Community	8B
Mixtral	`mistralai/Mixtral-8x22B-Instruct-v0.1`	DeepInfra API 05.2024	Open source	176B

Table 4.4: Detailed overview of evaluated LLMs (*Abbr.* stands for abbreviation)

4.7 Results and Discussion

This section reports the main evaluation results on EUDisinfoTest. We first compare overall model performance across all narrative forms, then contextualise these results with a small-scale human study on Base narratives. We next analyse how rhetorical variants (Pathos, Logos, Ethos) systematically shift model behaviour relative to Base, and finally break down performance by topical domain to identify where models are most and least reliable.

4.7.1 Overall Model Performance

Table 4.5 presents the baseline scores for all models, aggregated across narrative types (Base, Pathos, Ethos, Logos). The primary metric is Agg-F1-Score, with Agg-TNR and Agg-TPR providing complementary information on behaviour with respect to credible and disinformation narratives.

GPT-5 achieves the highest overall performance with an Agg-F1-Score of 0.90, followed closely by Claude 3 Opus (0.89) and Sonnet (0.88). Llama3-70b also performs strongly (0.86). At the lower end of the spectrum, GPT-3.5 (0.76) and Llama3-8b (0.74) show substantially weaker performance, despite their widespread

use in applications. This confirms that popularity and deployment scale do not necessarily translate into robustness in safety-critical tasks such as disinformation classification. The average Agg-F1-Score across all models is 0.83, with mean Agg-TNR at 0.84 and Agg-TPR 0.84, indicating reasonably balanced performance across credible and disinformation narratives when aggregated over variants.

Model	Agg-F1-Score	Agg-TNR	Agg-TPR
GPT5	0.90	0.94	0.85
Opus	0.89	0.91	0.85
Sonnet	0.88	0.90	0.86
Llama3-70b	0.86	0.88	0.83
Haiku	0.84	0.80	0.90
Mixtral	0.78	0.64	0.95
GPT3.5	0.76	0.82	0.70
Llama3-8b	0.74	0.85	0.62
Average	0.83	0.84	0.82

Table 4.5: Aggregated performance metrics (F1-score, TNR and TPR) for each model, averaged across narrative styles (Base, Logos, Ethos, Pathos).

These findings are consistent with broader evaluation patterns reported in the literature [90, 107, 208], where more capable models in general-purpose benchmarks tend to retain their advantage on safety-related tasks, while smaller or older models lag behind.

4.7.2 Human Performance on EUDisinfoTest

To contextualise LLM performance, an additional study compared model predictions with human judgements on Base narratives. The evaluation was conducted during the Horizon Europe Link4Skills kick-off meeting.

Five sets of Base narratives were prepared, each containing 18 items (9 credible and 9 disinformation narratives) sampled across the 9 topics in EUDisinfoTest. In total, 90 Base narratives were used (10 per topic). Each set was evaluated by 5 to 6 participants, resulting in 28 participants overall, with each individual assessing 18 narratives.

Participants were asked to classify each narrative as either credible or disinformation. Based on these judgements, two forms of human performance were

computed:

Human-avg.

TNR, TPR and F1-score were first calculated separately for each participant and then averaged across all 28 individuals. This measure captures individual variability and reflects how a typical participant performs. There was substantial variation between participants; notably, three out of 28 achieved perfect scores (no errors).

Human-consensus.

In this setting, a narrative is assigned a label by majority vote: it is deemed credible if a strict majority of participants ($> 50\%$) classify it as such, and disinformation otherwise. TNR, TPR and F1-score are then computed based on these consensus labels. This approach mitigates individual noise and approximates a collective judgement.

Table 4.6 reports performance for LLMs and human evaluators on Base credible and Base disinformation narratives.

Model	F1-score	TPR	TNR
GPT-5	0.93	0.88	0.98
Opus	0.97	0.96	0.98
Sonnet	0.91	0.93	0.91
Llama3-70b	0.97	0.94	0.99
Haiku	0.93	0.97	0.90
Mixtral	0.97	0.95	0.98
GPT3.5	0.87	0.81	0.92
Llama3-8b	0.87	0.73	0.98
Average (All Models)	0.93	0.90	0.96
Human-avg	0.72	0.66	0.94
Human-consensus	0.80	0.67	1.00

Table 4.6: Comparison of performance metrics (F1-score, TPR, TNR) for LLMs and human evaluators on Base credible and Base disinformation narratives.

On this subset, all evaluated LLMs achieve higher F1-scores than both Human-avg and Human-consensus, indicating that, on average, models are more consistent than individual or majority human judgements in jointly balancing false positives

and false negatives. However, humans exhibit very high specificity (TNR): Human-avg reaches 0.94 and Human-consensus achieves a TNR of 1.00, meaning that consensus never misclassifies a credible narrative as disinformation in this sample. In contrast, human sensitivity to disinformation (TPR) is substantially lower (0.66 for Human-avg and 0.67 for Human-consensus), indicating that humans are more likely to overlook disinformation than to falsely flag credible content.

Overall, these results suggest a complementary pattern: LLMs provide more uniform overall performance on Base narratives (higher F1-score), whereas humans, especially in consensus, are extremely cautious about mislabelling credible content, at the cost of missing a notable fraction of disinformation narratives. This trade-off highlights the potential value of combining automated systems with human oversight in practical disinformation classification settings.

4.7.3 Impact of Rhetorical Strategies

This subsection examines how rhetorical variants (Pathos, Logos, Ethos) affect classification behaviour, focusing on TNR and TPR. Detailed scores per model and variant are reported in Tables 4.7 and 4.8, and Figure 4.4 visualises performance drift relative to Base narratives.

In the **Base** condition, models already perform strongly, with an average TNR of 0.96 and an average TPR of 0.90. This indicates that, on average, both credible and disinformation narratives are classified reliably, with a slight advantage in correctly recognising credible content (higher TNR) compared to detecting disinformation (TPR).

In the **Base** condition, models already perform strongly, with an average TNR of 0.96 and an average TPR of 0.89. This indicates that, on average, both credible and disinformation narratives are classified reliably, with a slight advantage in correctly recognising credible content (higher TNR) compared to detecting disinformation (TPR).

Pathos. In the **Pathos** condition, emotional framing has a clearly destabilising effect on classification. The average TPR drops from 0.89 (Base) to 0.69, a reduction of 23%, indicating a pronounced decrease in the ability to correctly identify disinformation narratives. At the same time, the average TNR increases slightly from 0.96 to 0.97 (a 2% gain), suggesting that emotionally charged content is more often treated as credible when it is in fact credible, but also that models

	TNR			
Model	Base	Pathos	Logos	Ethos
Llama3-70b	0.99	0.99 ↑0%	0.86 ↓13%	0.70 ↓29%
GPT-5	0.98	0.99 ↑1%	0.91 ↓7%	0.89 ↓9%
Opus	0.98	0.99 ↑1%	0.88 ↓10%	0.86 ↓12%
Llama3-8b	0.98	0.98 ↓0%	0.82 ↓16%	0.69 ↓30%
Mixtral	0.98	0.91 ↓7%	0.59 ↓40%	0.27 ↓72%
GPT3.5	0.92	0.99 ↑8%	0.73 ↓21%	0.66 ↓28%
Sonnet	0.91	0.96 ↑5%	0.89 ↓2%	0.86 ↓5%
Haiku	0.90	0.96 ↑7%	0.80 ↓11%	0.57 ↓37%
Average	0.96	0.97 ↑2%	0.81 ↓15%	0.69 ↓28%

Table 4.7: TNR across LLMs and narrative types. For **Pathos**, **Logos** and **Ethos**, values in parentheses indicate percentage change relative to the Base score.

become less sensitive to disinformation in this setting.

Logos. For **Logos**, the impact is more balanced but still adverse overall. The average TPR decreases by 7% relative to Base (from 0.89 to 0.83), while the average TNR decreases by 15% (from 0.96 to 0.81). Logical or evidence-based framing therefore makes narratives harder to classify correctly. The drop in TNR indicates that credible Logos narratives are more often misclassified as disinformation, while the drop in TPR shows that a larger fraction of logically framed disinformation is mistaken for credible content.

Ethos. In the **Ethos** condition, appeals to authority introduce the most significant change in TNR. On average, TNR falls from 0.96 (Base) to 0.69, a substantial 28% reduction, indicating that authority-laden credible narratives are much more likely to be misclassified as disinformation. By contrast, the average TPR increases slightly from 0.89 to 0.91 (a gain of 1%), suggesting that models remain strong, and in some cases improve, in detecting disinformation when it is present. Ethos frames therefore push models towards a more suspicious stance, increasing the risk of false positives on credible content, even for stronger models.

Table A.1 provides concrete examples of Base narratives and their Pathos, Logos and Ethos variants, together with GPT-3.5 predictions, illustrating how the

	TPR			
Model	Base	Pathos	Logos	Ethos
Haiku	0.97	0.80 ↓18%	0.89 ↓8%	0.96 ↓1%
Mixtral	0.95	0.89 ↓6%	0.96 ↑1%	0.98 ↑3%
Llama3-70b	0.94	0.65 ↓31%	0.86 ↓9%	0.95 ↑1%
Opus	0.93	0.69 ↓26%	0.82 ↓12%	0.98 ↑5%
Sonnet	0.93	0.76 ↓18%	0.88 ↓5%	0.96 ↑3%
GPT-5	0.88	0.84 ↓5%	0.83 ↓6%	0.85 ↓3%
GPT3.5	0.81	0.44 ↓46%	0.76 ↓6%	0.86 ↑6%
Llama3-8b	0.73	0.45 ↓38%	0.67 ↓8%	0.70 ↓4%
Average	0.89	0.69 ↓23%	0.83 ↓7%	0.91 ↑1%

Table 4.8: TPR across LLMs and narrative types. For **Pathos**, **Logos** and **Ethos**, values in parentheses indicate percentage change relative to the Base score.

addition of persuasive cues can flip model judgements. Figure 4.4 summarises the performance drift induced by each rhetorical strategy relative to Base narratives.

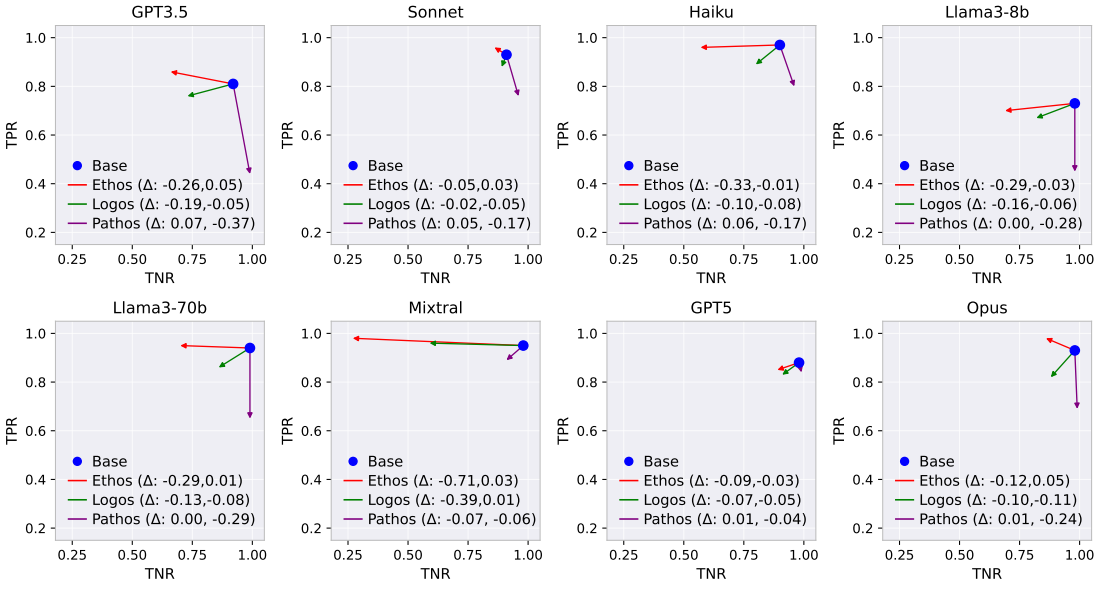

Figure 4.4: Impact of rhetorical strategies on TNR and TPR across models, expressed as percentage change relative to Base performance.

4.7.4 Topic-wise Performance

Table 4.9 reports Agg-F1-scores by topic and model. Performance is not uniform across thematic domains, revealing topic-specific strengths and weaknesses.

Topic	GPT5	Opus	Sonnet	Haiku	Llama3-70b	GPT3.5	Llama3-8b	Mixtral	Avg (Std Dev)
Anti-migration and xenophobia	0.91	0.97	0.92	0.92	0.89	0.83	0.75	0.67	0.86 (0.101)
Climate change and the energy crisis	0.91	0.94	0.96	0.85	0.88	0.87	0.68	0.62	0.84 (0.123)
Health, Covid-19, and vaccines	0.94	0.90	0.92	0.94	0.82	0.82	0.73	0.64	0.84 (0.109)
Gender-based disinformation	0.91	0.91	0.92	0.89	0.82	0.82	0.78	0.57	0.83 (0.116)
Historical revisionism	0.90	0.95	0.99	0.85	0.74	0.72	0.69	0.62	0.81 (0.134)
Ukraine war	0.97	0.93	0.91	0.87	0.76	0.70	0.66	0.61	0.80 (0.136)
Anti-Europeanism and anti-Atlanticism	0.83	0.82	0.82	0.71	0.75	0.70	0.69	0.56	0.74 (0.091)
Institutional distrust	0.89	0.75	0.63	0.57	0.71	0.63	0.61	0.53	0.67 (0.115)
Media distrust	0.78	0.70	0.55	0.66	0.46	0.38	0.40	0.52	0.56 (0.145)

Table 4.9: Agg-F1-scores of different models on various topics, sorted by average score.

Models perform best on topics where disinformation follows well-known and recognisable patterns, such as *Anti-migration and xenophobia* (average 0.86) and *Climate change and the energy crisis* (average 0.84). Similarly strong performance is observed for *Health, Covid-19, and vaccines* (average 0.84) and *Gender-based disinformation* (0.83), followed by *Historical revisionism* (0.81) and the *Ukraine war* (0.80).

By contrast, performance drops markedly for *Institutional distrust* (average 0.67) and especially *Media distrust* (0.56). These domains often involve nuanced critiques of institutions and media ecosystems, where the boundary between legitimate scepticism and disinformation is less clear-cut and more context-dependent. *Anti-Europeanism and anti-Atlanticism* (0.74) sits in between, but still reveals challenges in handling ideologically charged content that blends factual claims with opinion and framing.

Across topics, GPT5 and Opus tend to occupy the top positions, reflecting their robustness and better calibration, while GPT-3.5, Llama3-8b, and Mixtral are consistently among the weakest performers. This pattern mirrors the aggregate results in Table 4.5 and underscores the need for careful model selection in high-stakes disinformation classification tasks.

Chapter 5

DiNaM: Mining Disinformation Narratives from Fact-Checking Articles

5.1 Introduction

The previous chapter evaluated how Large Language Models distinguish credible from disinformation narratives using the EUDisinfoTest benchmark, where models classify narratives as Credible or Disinformation. While this perspective is useful for analysing model robustness in a controlled environment, it abstracts away from a central practical challenge: in real-world workflows, analysts, fact-checkers and observatories rarely start from canonical narratives. Instead, they work with large, heterogeneous collections of articles, reports and fact-checks from which narrative structures must first be reconstructed.

DiNaM addresses this gap by operationalising disinformation narrative mining on fact-checking articles. Fact-checking organisations and observatories already invest substantial effort in identifying and documenting false claims, their contexts and their corrections. Compared to general news or social media, fact-checking articles therefore provide a high-precision, expert-curated record of verified false information and its credible refutation. Moreover, large-scale aggregations of fact-checks have been shown to function as a useful proxy for studying disinformation at scale, including across languages and borders, enabling analyses of recurrent claims and their diffusion across countries and linguistic communities [157, 92].

At the same time, fact-checks are an imperfect lens on the full disinformation environment. What gets checked is shaped by organisational missions, incentives, and newsroom routines; coverage is uneven across regions and languages; and claim formulations in fact-checks may reflect editorial paraphrasing rather than verbatim source texts [157, 165]. DiNaM therefore treats fact-checking corpora as a high-precision but incomplete sample: well-suited for reconstructing widely

salient, repeatedly debunked narratives, but not a comprehensive census of all disinformation narratives circulating in the wild.

Moreover, these texts are written for human readers rather than for automated analysis: false claims are interwoven with debunking language, background information and meta-data, and appear in multiple languages and styles. There is no direct mapping from this raw corpus to the compact narratives used in benchmarks such as EUDisinfoTest.

Methodologically, DiNaM instantiates the three-stage disinformation narrative-mining framework from Section 3.2.2 on fact-checking articles, turning thousands of fact-checks into a structured catalogue of narrative-level representations.

A key design choice is the scope of the narratives DiNaM seeks to recover. Rather than focusing on a single outlet, campaign or platform, DiNaM aggregates across many fact-checking organisations and topics. Its goal is to reconstruct global disinformation narratives that recur across countries, languages and contexts. For example, narratives about vaccine safety, electoral fraud, migration or the European Union. The result is a data-driven map of the disinformation narrative landscape implied by contemporary European fact-checking practice: narratives are grounded in expert judgements, but derived at a scale and level of detail that would be difficult to obtain manually. DiNaM focuses on the propositional content of such narratives; rhetorical variants (for example logos-, ethos- or pathos-forward variants) are not explicitly modelled here and remain implicit in the underlying fact-checking texts.

Within the overall research programme, DiNaM instantiates the narrative-mining framework from Chapter 3 to provide a global map of disinformation narratives (complemented by DiNO's outlet-level view in Chapter 6).

Within the overall research programme of the thesis, DiNaM provides a global view of disinformation narratives derived from fact-checking corpora and organised across topics, time and space. EUDisinfoTest, introduced in Chapter 4, serves as an external narrative-level reference set for evaluating the coverage and quality of the narratives DiNaM discovers, but is not used during the mining process itself. Chapter 6 then introduces DiNO, which applies the same narrative-mining framework to outlet-specific news corpora to obtain a local, outlet-level view of how these narratives are instantiated, amplified and adapted in news coverage. Together, these components connect narrative-level evaluation of LLMs with global and local mapping of disinformation narratives.

To support reproducibility, we publicly release the DiNaM methodology, dataset,

prompts, and codebase [1].

5.2 DiNaM Algorithm

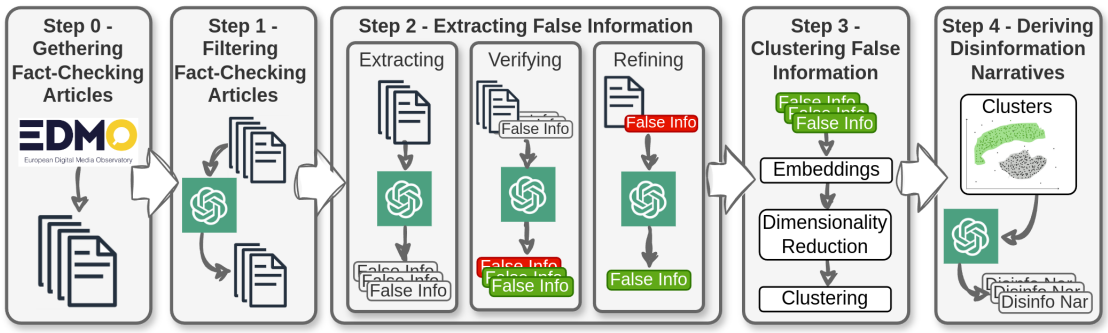

Figure 5.1: Overview of the DiNaM Algorithm: The process begins with data collection (Step 0), which involves gathering fact-checking articles to establish the dataset. This step does not strictly belong to DiNaM, as other sources of fact-checking articles could also serve as input. It continues with filtering articles to isolate those addressing false claims (Step 1). False information is then extracted, verified, and refined (Step 2). The extracted data is clustered into semantically similar groups representing distinct disinformation narratives using embeddings, dimensionality reduction, and clustering (Step 3). Finally, we derive disinformation narratives from each cluster (Step 4).

DiNaM takes as input a collection of expert-curated fact-checks and outputs a set of disinformation narratives, each linked to the false information instances it summarises. It instantiates the three abstract steps from Section 3.2.2 as (i) identification of false information, (ii) clustering of false information, and (iii) narrative derivation from clusters.

Formally, let

$$\mathcal{C} = \{c_1, \ldots, c_n\}$$

denote the set of fact-checking articles. DiNaM then instantiates the three stages as follows:

1. **Identify false information (incorrect basis).** DiNaM first filters the corpus $\mathcal{C}$ to retain only those fact-checking articles that review exclusively false

[1] `https://github.com/wsosnowski/DiNaM`

information (Section 5.2.1). Let

$$\mathcal{C}^{false} = \{c_i \in \mathcal{C} \mid c_i \text{ is labelled as reviewing only false content}\}$$

denote this subset. From each $c_i \in \mathcal{C}^{false}$, the algorithm then extracts textual statements that express the false claims themselves, separating them as far as possible from corrective or contextual material (Section 5.2.2). The union of all extracted statements forms the set

$$\mathcal{S} = \bigcup_{c_i \in \mathcal{C}^{false}} S_i,$$

where $S_i = \{s_{i,1}, \ldots, s_{i,K_i}\}$ is the set of false information instances obtained from article c_i.

2. **Cluster false information (multiple examples).** Each instance of false information $s_j \in \mathcal{S}$ is embedded into a semantic vector space via an encoder

$$\phi : \mathcal{S} \to \mathbb{R}^d, \quad s_j \mapsto \phi(s_j).$$

The resulting vectors $\{\phi(s_j)\}_{j=1}^{M}$ are then grouped into semantically coherent clusters (Section 5.2.3). This yields a set of clusters

$$\mathcal{K} = \{K_1, \ldots, K_r\},$$

where each cluster $K_q \subseteq \mathcal{S}$ contains instances of false information that express closely related misleading messages. In the intended use, clusters aggregate multiple examples so that each cluster corresponds to a recurring theme rather than an isolated instance.

3. **Derive narratives (adaptability).** For each cluster $K_q \in \mathcal{K}$, DiNaM uses an LLM to synthesise a disinformation narrative N_q that captures the recurring misleading message shared by the statements in K_q (Section 5.2.4). Collectively, these narratives form

$$\mathcal{N}_{\text{Dis}} = \{N_1, \ldots, N_r\},$$

a set of disinformation narratives.

To make this process more concrete, consider a fact-checking article that debunks the claim that "COVID-19 vaccines cause widespread infertility". Step 1 identifies this article as reviewing a false claim (if the debunking verdict in that article classifies the claim as false). Step 2 extracts instances of false information such as "COVID-19 vaccines cause infertility" from that article and similar statements from other articles. Step 3 clusters these statements together with related claims, for

example about vaccines causing miscarriages or sterility. Step 4 finally summarises the cluster into a narrative such as "COVID-19 vaccines are dangerous", which can be tracked across outlets, countries and time.

All pipeline components were iteratively developed and refined through empirical testing, and design choices are grounded in the evaluation results reported in Section 5.4, which details the models, prompts and implementation.

5.2.1 Filtering False Information

DiNaM begins from a curated dataset of fact-checking articles and selects only those that exclusively review false information (Step 1 in Figure 5.1).[2] Fact-checking articles can evaluate claims as true, false, partially true or a mix of both. To simplify the downstream extraction task and avoid ambiguous supervision, DiNaM retains only those articles in which all evaluated claims are rated as false.

This decision is supported by an empirical analysis showing that fewer than 3% of fact-checking articles in our corpus assess a mix of true and false claims (see Appendix B.3.1). Filtering therefore removes a small fraction of the data while yielding a cleaner supervision signal.

Operationally, this step is framed as a binary classification task: given a fact-checking article, determine whether it meets the predefined filtering criterion ("all claims false").

5.2.2 Extracting False Information

After filtering articles, DiNaM proceeds to extract the false information addressed within these articles. We treat extraction as a contextual question-answering problem and decompose it into three substeps carried out by an LLM (Step 2 in Figure 5.1):

Extracting. Given a fact-checking article, the LLM is prompted to extract all relevant false claims analyzed in each fact-checking article. The goal is to capture the misleading content itself, separated as far as possible from corrective or contextual material.

[2]In this work, "false claims" and "false information" are used interchangeably.

Verifying. In a second pass, the LLM is provided with the fact-checking article and each extracted instance of false information, and is prompted to confirm whether the claim is indeed fact-checked and identified as false in the article.

Refining. Finally, DiNaM revisits articles that failed verification. For these cases, the LLM is prompted to re-extract false information, guided by feedback on deficiencies in the previous attempt (e.g., inclusion of debunking language or omission of key elements). In practice, a single refinement pass is sufficient: only approximately 8% of initially extracted claims require refinement, and after this step only about 8% of those (0.64% of the full dataset) remain incorrect or "hallucinated". Given this low residual error rate, we stop further refinement to balance precision, coverage and computational cost.

This three-step procedure is inspired by the Chain-of-Verification (CoVe) prompt engineering method, which has shown strong performance on contextual question-answering tasks [56]. DiNaM adapts CoVe by condensing its four-step workflow into three steps: we omit CoVe's question-generation phase and instead rely on manually crafted, task-specific questions that are reused across articles during verification. This simplifies the pipeline while preserving CoVe's core rationale of stepwise verification.

5.2.3 Clustering False Information

Once the corpus of false information has been prepared, DiNaM clusters them semantically (Step 3 in Figure 5.1). The objective is to group similar false claims into clusters, so that each cluster corresponds to a candidate disinformation narrative emerging from multiple, related examples.

Concretely, each statement is first mapped to a dense vector representation (embedding) that captures its semantic content. To alleviate the curse of dimensionality and improve cluster structure, we apply dimensionality reduction before clustering. We then apply a clustering algorithm to group semantically similar false claims into clusters. Each resulting cluster is intended to aggregate multiple instances of a recurring disinformation theme, which will then be summarised into a narrative.

5.2.4 Deriving Disinformation Narratives

In the final step, DiNaM uses an LLM to derive disinformation narratives from the clusters of false information. For each cluster, the LLM receives the list of constituent statements and is prompted to identify the underlying pattern of false information and to generate a single narrative that captures this pattern.

To ensure alignment with real-world examples, we define output constraints informed by the EUDisinfoTest dataset [184]. Most narratives in this dataset are concise, typically under 15 words and self-contained. Accordingly, each generated narrative must be: (1) clear, standalone, and descriptive, (2) no longer than 15 words and (3) expressing the false perspective behind the claims.

These constraints also align with the properties of disinformation narratives defined in Section 3.1.3.

5.2.5 Prompt Template

Integrating LLMs into DiNaM requires a specialised prompt template that can be reused across the different LLM-based steps (filtering, extraction, verification and narrative derivation). The template is illustrated in Figure 5.2 and follows the same general structure as the zero-shot disinformation classification prompt template proposed by Lucas et al. [110].

Conceptually, the template combines (i) an impersonation component that specifies the role of the model, (ii) detailed guidelines that describe the task, (iii) a context block containing the relevant input text, and (iv) an output specification that fixes the required format. For prompts that use chain-of-thought reasoning, we additionally include an explicit list of steps to follow. A closely related template is reused in DiNO (Chapter 6), ensuring methodological consistency across chapters.

5.3 Dataset

5.3.1 Source Organisations and Data Collection

DiNaM uses a dataset of 10,493 fact-checking articles sourced from 14 independent organisations affiliated with the European Digital Media Observatory [134]. These organisations are based across all 27 EU member states and Norway, contributing fact-checking content from their respective regions. Collectively, they

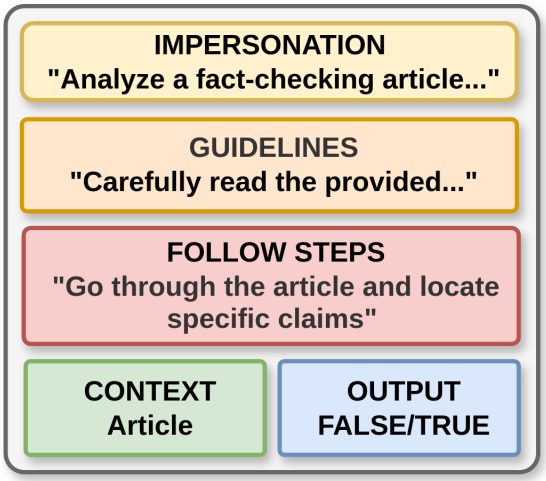

Figure 5.2: Prompt template used in DiNaM. It specifies (1) an impersonation role, (2) task guidelines, (3) the input context, (4) the desired output format, and, where applicable, (5) step-by-step instructions for chain-of-thought reasoning.

have conducted over 6,800 training sessions and employed 91 verification tools in support of their fact-checking activities [58].

To construct the dataset, we retrieved HTML content from the fact-checking websites using Selenium [173] and extracted clean article text with Trafilatura [26]. Articles originally written in languages other than English were translated into English using Google Translate.[3] All descriptive statistics reported in this section are computed on the translated content. This dataset instantiates the abstract corpus $\mathcal{C}$ used in Section 5.2.

5.3.2 Temporal Coverage

The distribution of fact-checking articles over time is depicted in Figure 5.3. A notable increase in articles is observed starting from 2020, reflecting the intensification of fact-checking activities in response to major global events such as the COVID-19 pandemic and subsequent geopolitical crises.

[3]Google Translate, available at: `https://translate.google.com/`, accessed: 4 September 2024.

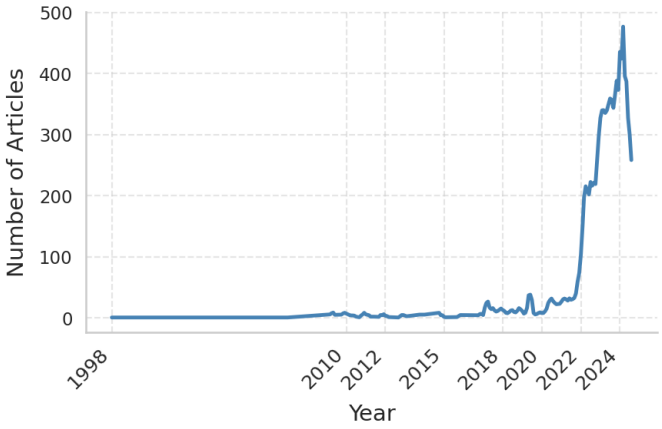

Figure 5.3: Distribution of fact-checking articles over time.

5.3.3 Article Length

Figure 5.4 shows the distribution of article lengths in terms of word count. The majority of fact-checking articles contain between 500 and 2,000 words, with a long tail of exceptionally short and exceptionally long pieces.

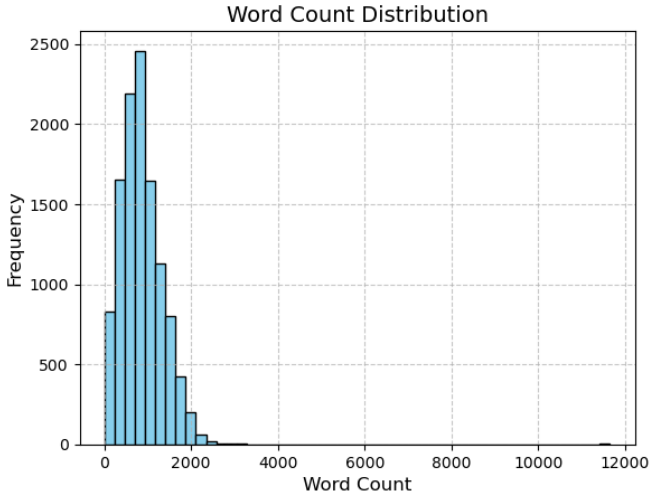

Figure 5.4: Word count distribution of fact-checking articles.

Key summary statistics for article length are presented in Table 5.1. The mean article length is 868.5 words, with a median of 812 words and a standard deviation of 493.7 words, indicating substantial variability in the level of detail provided across fact-checks. Article lengths range from very brief notices of 20 words to in-depth reports of up to 11,646 words.

Metric	Words
Mean	868.5
Median	812.0
Standard Deviation	493.7
Minimum	20
Maximum	11646

Table 5.1: Word count statistics for the dataset.

5.3.4 Sources by Domain and Affiliation

The dataset contains contributions from a range of fact-checking domains and affiliations. Figure 5.5 shows the ten most frequent web domains and their percentage contributions to the corpus. These domains represent the main fact-checking outlets in the dataset and highlight the concentration of fact-checking efforts among a relatively small set of specialised organisations.

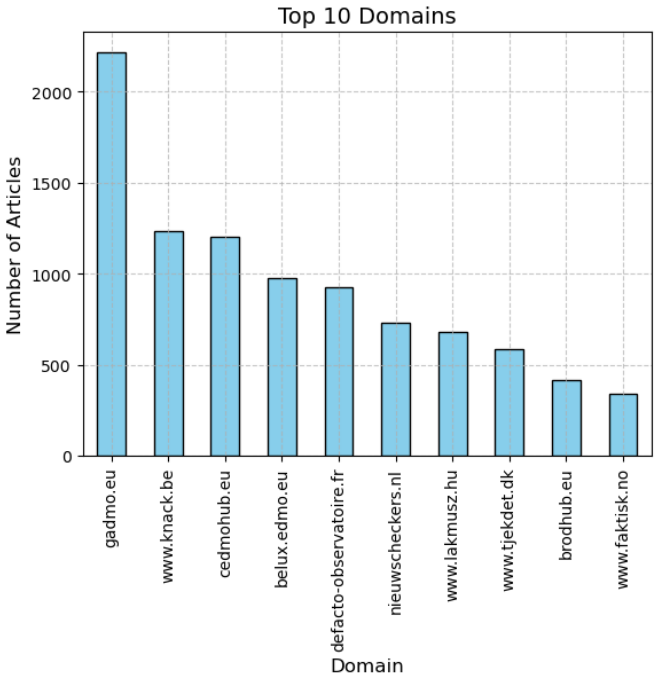

Figure 5.5: Top 10 most frequent domains in the dataset, along with their percentage contributions.

Figure 5.6 reports the distribution of fact-checking articles by EDMO affiliation. Among these, BENEDMO contributes the largest number of articles (1,895),

followed by GADMO (1,687), EDMOeu (1,302) and CEDMO (1,018). Other affiliations such as NORDIS (895), BELUX (830), DEFACTO (824), HDMO (795), BECID (760) and BROD (574) provide substantial regional coverage. Smaller but still important contributions come from MedDMO (524), ADMO (304) and the IRELAND HUB (238). Together, these affiliations illustrate the breadth of the European fact-checking ecosystem captured in the dataset.

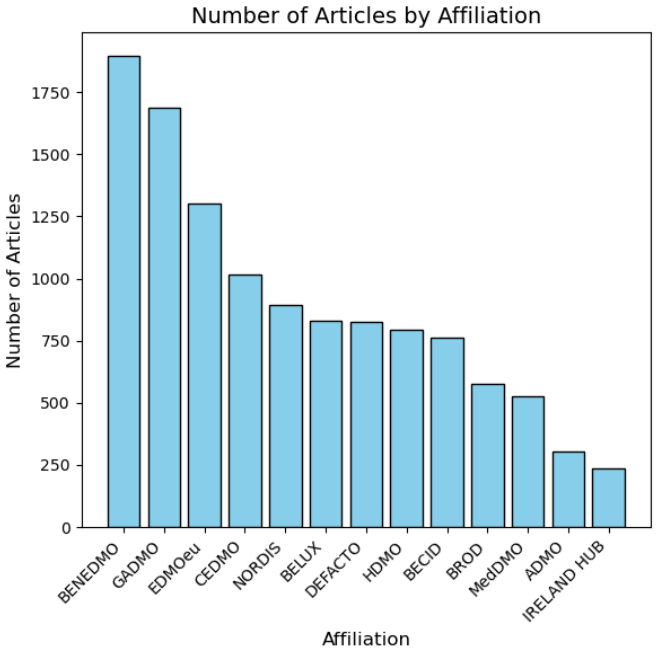

Figure 5.6: Distribution of fact-checking articles by affiliation, illustrating the contributions of various organisations.

5.4 Evaluation

This section outlines the evaluation methodology that guided the design decisions in DiNaM. To ensure robustness, we report averages over three runs and use a standardised prompt template for constructing all DiNaM prompts (see Section 5.2.5).

For cost-efficiency, we evaluated both lightweight LLMs and non-LLM baselines. While non-LLM models consistently underperformed compared to LLMs, we include them to provide a broader performance context and for completeness.

5.4.1 Filtering Fact-Checking Articles

In the first step, DiNaM filters the dataset to include only fact-checking articles in which every claim reviewed by the fact-checkers is rated as false.

Testing Dataset. To evaluate this step, we constructed a labelled dataset of fact-checking articles, annotated as either **positive** (all claims are rated false: the article passes the filter) or **negative** (at least one claim is not rated false: the article fails the filter). We randomly selected 135 articles from the EDMO dataset,[4] and each article was annotated independently by two experts. They reached agreement on 124 articles, which form our final evaluation set. Details on the annotation guidelines are provided in Appendix B.3.2.

LLM Evaluation. We evaluated GPT-4o-mini, LLaMA-3.3-70B, Qwen3-32B and Gemma-3-27B in a zero-shot setting,[5] asking each model to predict whether an article is **positive** or **negative**. For each model, we tested three prompt variants (see Appendix B.1.1) and measured performance using the F1-score against the gold labels.

Non-LLM Evaluation. To establish a performance baseline, we also fine-tuned two encoder-based transformer models: RoBERTa-large [108] and DeBERTa-large [88]. Both models were trained on the same labelled dataset described above, with each article represented by its full text and the target label indicating whether all claims were rated false.

Fine-tuning was implemented using the HuggingFace Transformers library. We used binary cross-entropy as the loss function and optimised with AdamW, using a batch size of 8, a learning rate of 2e-5 and 5 training epochs. Evaluation was performed with 5-fold cross-validation, ensuring class balance within each fold.

Results. Table 5.2 summarises the F1 scores for all models. Among the LLMs, GPT-4o-mini and Qwen3-32B achieve the highest F1 score of 0.88, followed by Gemma-3-27B (0.87) and LLaMA-3.3-70B (0.86). The fine-tuned DeBERTa-

[4]Following Alwosheel et al. [16], we selected 135 examples to ensure at least 50 per class for binary classification and to allow for possible exclusions due to annotator disagreement.

[5]Few-shot classification [35] was not used due to the impracticality of including full article examples in the prompt.

large and RoBERTa-large baselines reach F1 scores of 0.60 and 0.58, respectively, substantially below the zero-shot LLMs. These results indicate that, for this filtering task, contemporary LLMs provide strong performance even without task-specific supervision.

Model	Prompt	F1 Score
LLMs		
GPT-4o-mini	Appendix B.2	**0.88**
Qwen3-32B	Appendix B.2	**0.88**
Gemma-3-27B	Appendix B.4	0.87
LLaMA-3.3-70B	Appendix B.3	0.86
Non-LLM		
DeBERTa-large (fine-tuned)	-	0.60
RoBERTa-large (fine-tuned)	-	0.58

Table 5.2: F1 scores for zero-shot LLM filtering and fine-tuned encoder baselines on the fact-checking article filtering task.

5.4.2 Extracting False Information

In the second step, DiNaM extracts false information from fact-checking articles.

Testing Dataset. To evaluate this step, we constructed a dataset of 115 pairs: each pair consists of a fact-checking article and the specific false information it addresses. These pairs were sourced from reports published by EDMO,[6] each of which contains examples of false information together with the corresponding fact-checking articles.

LLM Evaluation. We evaluated four models—GPT-4o-mini, LLaMA-3.3-70B, Qwen3-32B and Gemma-3-27B—focusing on their ability to identify the specific false information discussed in each article. For each model, we compared four prompting strategies: (i) our structured Extraction–Verification–Refinement workflow (Section 5.2.2), (ii) Chain-of-Thought (CoT) [202], (iii) Chain-of-Verification (CoVe) [56] and (iv) a simple Base prompt that directly asks for false claims

[6] `https://edmo.eu/resources/fact-checking-publications/fact-checking-briefs/`

without additional verification. The full prompts for each strategy are provided in Appendix B.1.2.

Non-LLM Evaluation. To complement the LLM experiments, we also evaluated two encoder-based QA-style systems as non-LLM baselines. The first, DistilBERT-SQD, is a lightweight DistilBERT model fine-tuned on SQuAD v1.1; the second, RoBERTa-SQD2, is based on RoBERTa-base fine-tuned on SQuAD v2.0. Both models were queried using a standardised question format: *"What false claims are debunked in this article? Context: {article}"*, where the full article text served as context. Their outputs were evaluated using the same semantic similarity metric as for the LLMs, allowing for a direct comparison.

Evaluation Metric. We measure performance with Average Weighted Chamfer Distance (Avg WCD), which compares two sets of extracted claims in embedding space.[7] For each article i, let X_i be the set of ground-truth false-claim embeddings and Y_i the set predicted by the model. For every item in one set, we find its most similar item in the other set (cosine similarity), and average these best-match similarities in both directions. This penalises both missing claims (no close match in Y_i) and spurious claims (no close match in X_i), while allowing paraphrases.

Formally, with cosine similarity $s(\cdot, \cdot)$, we compute:

$$\mathrm{WCD}(X_i, Y_i) = \frac{1}{|X_i| + |Y_i|} \Big(\sum_{x \in X_i} \max_{y \in Y_i} s(x, y) + \sum_{y \in Y_i} \max_{x \in X_i} s(y, x) \Big). \tag{5.1}$$

The final score is the average over all N test pairs:

$$\text{Avg WCD} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{WCD}(X_i, Y_i). \tag{5.2}$$

Higher values mean better alignment. We mainly interpret Avg WCD comparatively across methods (it depends on the embedding model), and we use an embedding model separate from the generators to reduce evaluation circularity. Appendix B.2.1 provides additional discussion and comparisons to related metrics.

Results. Table 5.3 summarises the LLM results. Among prompting strategies, our structured Extraction–Verification–Refinement approach consistently outperforms

[7] We compute embeddings with SFR-embedding-2 [168], a strong model on MTEB [125].

the baselines, with GPT-4o-mini achieving the highest Avg WCD score of 0.81. LLaMA-3.3-70B and Qwen3-32B follow closely (0.80), and Gemma-3-27B reaches 0.79. CoVe performs reasonably well but remains below our method, while CoT and the Base prompt yield weaker scores across models.

A central challenge in extracting false information is that LLMs often retain debunking context, inadvertently turning false claims into true ones. For example, the claim "Images show Pope Francis in a puffy, bright white coat" is sometimes extracted as "AI-generated images show Pope Francis in a puffy, bright white coat." Our approach explicitly identifies and removes such debunking cues, preserving only the false component of the claim. Among the baseline prompting strategies, only CoVe attempts verification, but it does not filter out debunking material, which limits its effectiveness. CoT and Base, by contrast, do not attempt verification at all, resulting in even weaker performance.

The non-LLM baselines perform substantially worse than LLMs: RoBERTa-SQD2 achieves an Avg WCD score of 0.58, while DistilBERT-SQD scores 0.54, well below the LLM range (0.73–0.81). These results reflect the limitations of encoder-only models on tasks that require paraphrase recognition and deeper semantic understanding.

Model	Our	CoVe	CoT	Base
LLMs				
GPT-4o-mini	**0.81**	0.79	0.78	0.77
LLaMA-3.3-70B	0.80	0.76	0.78	0.75
Qwen3-32B	0.80	0.77	0.75	0.75
Gemma-3-27B	0.79	0.79	0.76	0.73
Non-LLM				
RoBERTa-SQD2	-	-	-	0.58
DistilBERT-SQD	-	-	-	0.54

Table 5.3: Average WCD scores for false information extraction across models and prompting methods. LLMs are evaluated with four prompting variants. Non-LLM QA baselines are evaluated with a single QA-style prompt, and their scores are reported in the Base column (other columns are not applicable).

Comparison with Alternative Prompting Strategies. The experimental results highlight a consistent pattern across models and prompting strategies. Even

when explicitly instructed to extract only false claims, LLMs prompted with other strategies (Base, CoT, CoVe) frequently produce outputs that mix misleading statements with corrective or debunking content. This behaviour is visible across the examples in Appendix B.1 and is reflected in the lower Avg WCD scores for these baselines.

By contrast, our structured Extraction–Verification–Refinement pipeline mitigates this contamination: candidate claims are first identified, then debunking language is filtered out, and remaining statements are refined to better match the target notion of "false information". Empirically, this structured workflow yields cleaner extractions and systematically higher Avg WCD scores than both Base prompting and more generic multi-step strategies such as CoT and CoVe.

5.4.3 Clustering of False Information

In the third step, DiNaM groups extracted instances of false information into semantically meaningful clusters. This process involves three main components: embedding generation, dimensionality reduction and clustering.

Testing Dataset. To enable a robust comparison of clustering methods, we constructed a dataset of 12,286 instances of false information. These instances were extracted by applying Steps 1 and 2 of the DiNaM pipeline to the full set of fact-checking articles.

Evaluation. We assessed clustering quality using the Silhouette Score [167], performing a grid-based evaluation that tested combinations of embedding models, dimensionality reduction techniques and clustering algorithms. Silhouette Scores range from -1 to 1, with higher values indicating better within-cluster cohesion and between-cluster separation; values above 0.5 are generally taken to indicate well-separated clusters [190].

We evaluated three embedding models: SFR-Embedding-2 [168], E5-large [200], and jina-embeddings-v3 [191]; two dimensionality reduction techniques: UMAP [118] and PCA [4]; and two clustering algorithms: HDBSCAN [117] and K-means [119]. For each combination, we performed a grid search over typical hyperparameter ranges (Table 5.4), selecting the best configuration according to the Silhouette Score.

Algorithm	Hyperparameter	Range	Optimal
UMAP	n_neighbors	$\{10, 15, 30, 50, 100\}$	15
UMAP	n_components	$\{4, 8, 16, \ldots, 256\}$	256
PCA	n_components	$\{4, 8, 16, \ldots, 256\}$	256
HDBSCAN	min_cluster_size	$\{10, 15, 20, \ldots, 100\}$	25
HDBSCAN	min_samples	$\{10, 15, 20, \ldots, 100\}$	20
K-means	n_clusters	$\{5, 15, 25, \ldots, 800\}$	445

Table 5.4: Hyperparameter ranges and selected optimal values for dimensionality reduction and clustering methods.

Results. Table 5.5 presents the Silhouette Scores for all embedding–reduction–clustering combinations, using their respective best hyperparameters. The highest Silhouette Score of 0.67 is achieved by the combination of SFR-Embedding-2, UMAP and HDBSCAN, which we therefore adopt in DiNaM. This value lies comfortably above the conventional 0.5 threshold, indicating comparatively well-separated clusters in the induced embedding space.

The overall pattern aligns with expectations. HDBSCAN outperforms K-means across embedding models, reflecting the advantages of density-based clustering for textual data [189]. Similarly, UMAP outperforms PCA in dimensionality reduction, likely due to UMAP's ability to preserve both local and global structure in high-dimensional embedding spaces [118]. Among embedding models, SFR-Embedding-2 achieves the best scores, consistent with its strong performance on the MTEB benchmark [125].

5.4.4 Deriving Disinformation Narratives

The final step of our pipeline focuses on transforming clusters of false information into a set of disinformation narratives. Given the clusters produced in the previous stage, the goal is to generate a single disinformation narrative for each cluster.

Testing Dataset. To evaluate the accuracy of the generated narratives, we compared them against base disinformation narratives from the EUDisinfoTest benchmark.

Embedding Model	UMAP	PCA
SFR-Embedding-2		
HDBSCAN	**0.67**	0.52
K-means	0.54	0.38
E5-large		
HDBSCAN	**0.65**	0.52
K-means	0.61	0.53
jina-embeddings-v3		
HDBSCAN	**0.61**	0.48
K-means	0.55	0.45

Table 5.5: Silhouette Scores for combinations of embedding models, clustering algorithms and dimensionality reduction methods (best hyperparameters per method).

LLM Evaluation. We evaluated four LLMs: GPT-4o-mini, LLaMA-3.3-70B, Qwen3-32B and Gemma-3-27B. For each cluster, the model received the cluster's content as input, and we used a standardised prompt to generate exactly one disinformation narrative per cluster (see Appendix C.1.2). The resulting narratives were then compared to the corresponding reference narratives from EUDisinfoTest.

Non-LLM Evaluation. In addition to LLM-based generation, we evaluated two non-LLM summarisation baselines, BERTsum [158] and TextRank [121]. Both models were applied as summarisation systems over the same cluster content used for the LLMs, and their summaries were interpreted as candidate disinformation narratives.

Evaluation Metric. To assess similarity between model-generated narratives and the expert-written reference narratives, we again used the Weighted Chamfer Distance, following the setup in Section 5.4.2: both the DiNaM-generated narratives and the EUDisinfoTest base narratives are embedded using the same embedding model, and WCD is computed between the two sets. Larger scores correspond to closer alignment between predicted and reference narrative sets.

Results. Table 5.6 summarises the results. Among the LLMs, GPT-4o-mini achieves the highest WCD score of 0.73, with Qwen3-32B and LLaMA-3.3-70B close behind at 0.72, and Gemma-3-27B at 0.71.

The non-LLM summarisation baselines perform noticeably worse: BERTsum reaches a WCD score of 0.61, and TextRank 0.59, well below all LLM-based methods. This gap suggests that generic extractive/abstractive summarisation techniques struggle to capture the specific, perspective-sensitive notion of disinformation narratives, whereas LLMs can more reliably align cluster content with the concise, statement-like format used in EUDisinfoTest.

Model	WCD Score
LLMs	
GPT-4o-mini	**0.73**
Qwen3-32B	0.72
LLaMA-3.3-70B	0.72
Gemma-3-27B	0.71
Non-LLM	
BERTsum	0.61
TextRank	0.59

Table 5.6: WCD scores for deriving disinformation narratives from clusters of false information. LLMs are compared to non-LLM summarisation baselines. Higher scores indicate better alignment with expert-written EUDisinfoTest narratives.

5.4.5 Comparison with General Narrative Extraction Methods

To our knowledge, DiNaM is the only method specifically designed for mining disinformation narratives. However, there are two related algorithms for extracting general narratives: CaNarEx [18] and Relatio [20].

Unlike DiNaM, Relatio and CaNarEx are not designed to preprocess fact-checking articles to isolate false information before narrative extraction. As a result, evaluating these methods solely on fact-checking articles would be unfair. We therefore conducted the evaluation on two complementary datasets: 10,493 fact-checking articles and 12,286 false information instances (see Section 5.4.3). For both datasets, we applied CaNarEx and Relatio in their original configurations,

treating each article or false information as a document input, so that differences in performance primarily reflect differences in modelling assumptions rather than reparameterisation.

Comparison results are presented in Table 5.7. DiNaM clearly outperforms both Relatio and CaNarEx in mining disinformation narratives from both fact-checking articles and instances of false information. Specifically, DiNaM improves WCD scores by up to 24.7% compared to the best alternative, highlighting its substantial advantage in accurately capturing disinformation narratives. This improvement reflects both the explicit focus on false information and the tailoring of the pipeline to the structure of fact-checking texts, whereas CaNarEx and Relatio are designed for more general narrative extraction settings.

Model	Articles	False Information
DiNaM	**0.73**	-
Relatio	0.59 (**-19.2%**)	0.61 (**-16.4%**)
CaNarEx	0.55 (**-24.7%**)	0.59 (**-19.2%**)

Table 5.7: WCD scores and relative difference vs. DiNaM. DiNaM significantly outperforms other methods regardless of input format.

5.4.6 DiNaM Pipeline Overview

Table 5.8 summarises the DiNaM pipeline, outlining the best-performing models and input/output quantities at each step.

Step	Best Model	Input	Output
1. Filtering false info	GPT-4o-mini / Qwen3-32B	10.5K articles	9.6K articles
2. Extracting false info	GPT-4o-mini	9.6K articles	12.3K false info
3. Clustering false info	SFR-Embedding-2 & UMAP & HDBSCAN	12.3K false info	122 clusters
4. Deriving narratives	GPT-4o-mini	122 clusters	122 disinfo narratives

Table 5.8: Overview of DiNaM's pipeline with best-performing models and input/output quantities.

5.4.7 Generalizability

To evaluate DiNaM's robustness across diverse sources, we partitioned the EDMO dataset, which consists of fact-checking articles from 14 independent European organizations, into three mutually exclusive and balanced subsets, each containing articles from different sources. DiNaM was then applied independently to each subset. As shown in Table 5.9, performance remained consistent across subsets, with only minor variation relative to the full dataset. This demonstrates that DiNaM generalises well across different sources.

Dataset Partition	WCD Score
Subset 1 (Sources A–E)	0.71
Subset 2 (Sources F–J)	0.72
Subset 3 (Sources K–N)	0.72
Full Dataset	**0.73**

Table 5.9: WCD scores for DiNaM across organisational subsets of the EDMO dataset. Consistent scores indicate strong generalisation.

Moreover, we manually grouped the narratives identified by DiNaM (see Section 5.4.4) into seven main disinformation topics. For details on the annotation process and topic distribution, refer to Appendix B.3.3. To evaluate DiNaM's performance within each topic, we computed the WCD score between the predicted narratives and the corresponding ground truth narratives from EUDisinfoTest, filtering both sets by topic. As shown in Table 5.10, DiNaM demonstrates generalisation across diverse disinformation domains.

Collectively, these results suggest that the design choices underpinning DiNaM, in particular focusing on fact-checked false information, clustering at the level of claims rather than documents, and deriving narratives from these clusters, yield narrative representations that are stable across organisations and topics.

5.4.8 Cost Analysis of the DiNaM Algorithm

The DiNaM algorithm consists of four main steps, three of which (Steps 1, 2, and 4) rely on LLMs. These steps introduce processing costs due to token-based pricing schemes, especially when using commercial APIs. Step 3, by contrast, is purely computational and does not involve any LLMs.

Topic	WCD Score
COVID-19	0.76
Climate Change	0.73
LGBTQ+	0.68
Migration	0.72
The European Union	0.69
The Ukraine-Russia War	0.73

Table 5.10: WCD scores for DiNaM across different topical categories. Scores indicate robust alignment between predicted and ground truth narratives (EUDisinfoTest) across diverse subject areas.

To assess the cost-efficiency of DiNaM, we estimated the computational cost of each LLM-based step using a dataset of 10,493 fact-checking articles. Following OpenAI's guideline that one token corresponds to approximately 0.75 words,[8] we estimate a total of 33.02 million input tokens and 319.7 thousand output tokens across the pipeline.

Table 5.11 presents the estimated cost of running Steps 1, 2, and 4 using different models.

For all non-LLM-based operations (Step 3: embedding generation, dimensionality reduction, and clustering), we report runtime performance on our hardware configuration: 2× Intel® Xeon® Gold 5418Y CPUs, 128 GB RAM, and 4× NVIDIA L40 GPUs. On this setup, Step 3 took approximately 13 minutes to complete the full dataset.

Model	Input Rate ($/M)	Output Rate ($/M)	Input Cost ($)	Output Cost ($)	Total Cost ($)
GPT-4o mini	0.15	0.60	4.95	0.20	5.15
Gemma-3-27B	0.10	0.20	3.30	0.06	3.36
LLaMA-3.3-70B	0.10	0.25	3.30	0.08	3.38
Qwen3-32B	0.10	0.30	3.30	0.10	3.40

Table 5.11: Estimated LLM costs for Steps 1, 2 and 4 of the DiNaM algorithm, assuming total token usage of 33.02M input and 319.7K output tokens.

While GPT-4o-mini offers the best overall performance in our experiments,

[8] https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

the strongest open-weight alternative (Qwen3-32B) is competitive in quality at a modestly lower cost. The choice between these models therefore depends on deployment constraints (e.g. latency, privacy, or licensing) rather than purely on cost. Importantly, even for the largest model considered, the end-to-end cost of running DiNaM on the full EDMO corpus is on the order of only a few US dollars, making the approach feasible for periodic or even regular monitoring by fact-checking organisations and observatories.

5.5 Results and Discussion

This section presents the disinformation narratives identified by the DiNaM algorithm. We first focus on the key topics underlying these narratives, then provide detailed analyses of each of those. Together, these analyses realise the thesis goal of constructing a global, temporally resolved map of disinformation narratives implied by European fact-checking practice.

In terms of the framework from Chapter 3, the results in this section illustrate how DiNaM instantiates the three defining properties of disinformation narratives at scale. The incorrect basis is guaranteed by construction, as all narratives are derived from fact-checked false information. The multiple-examples requirement is satisfied by deriving narratives from clusters that aggregate many similar instances. Finally, the adaptability of disinformation narratives is visible in how these concise, clause-like statements reappear and evolve across countries, outlets and time in response to real-world events.

5.5.1 Disinformation Narrative Topics

We manually categorised the narratives discovered by DiNaM into seven main topics (see more in Appendix B.3.3). This categorisation enabled us to track the evolution of disinformation topics in response to real-world events (Figure 5.7). In 2020, as the COVID-19 pandemic unfolded, disinformation narratives related to the virus emerged. In 2022, as the war between Russia and Ukraine escalated, disinformation surrounding the conflict intensified. A significant spike in Israel–Palestine disinformation occurred in late 2023, coinciding with the onset of the Israel–Palestine war. Notably, COVID-19 narratives briefly resurfaced in early 2022, only to be overshadowed by the surge in Ukraine-Russia disinformation, suggesting the adaptability of disinformation campaigns in response to emerging

crises [192].

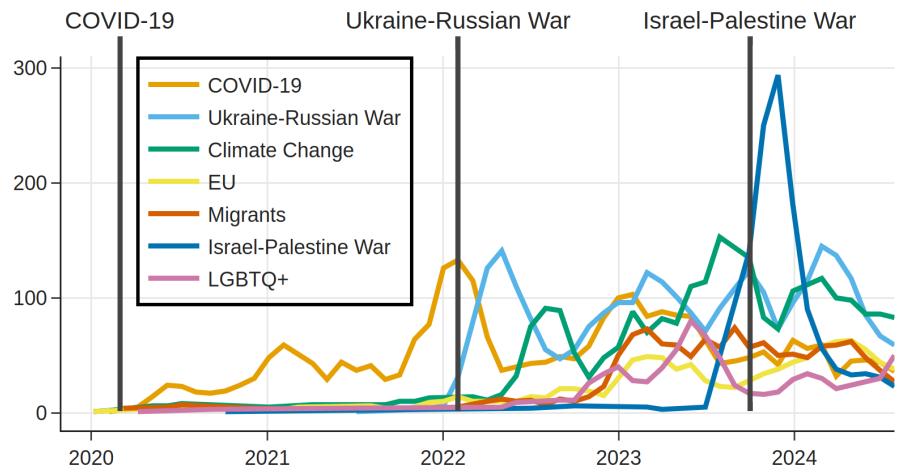

Figure 5.7: Temporal evolution of disinformation narratives discovered by DiNaM. Spikes in certain topics are related to major real-world events.

These temporal associations are correlational: while peaks in narrative activity align with major events, our analysis does not attempt to establish causal relationships between specific offline developments and disinformation campaigns.

5.5.2 COVID-19 Disinformation Narratives

Figure 5.8 illustrates the disinformation narratives related to COVID-19 identified by DiNaM. A notable trend is observed in the evolution of these narratives over time. Initially, in the aftermath of the COVID-19 outbreak in 2020, disinformation focused on the supposed health risks associated with wearing masks, claiming that they were ineffective in preventing the spread of the virus. This coincided with the WHO's early recommendations on mask usage [143]. As the pandemic progressed, the dominant disinformation narrative shifted in 2021 to focus on the alleged dangers of COVID-19 vaccines, which emerged concurrently with the vaccine rollout. However, by mid-2023, the prevalence of these COVID-19 narratives began to decline, likely due to the decreased severity of the pandemic and the WHO's declaration that the international public health emergency had ended [144].

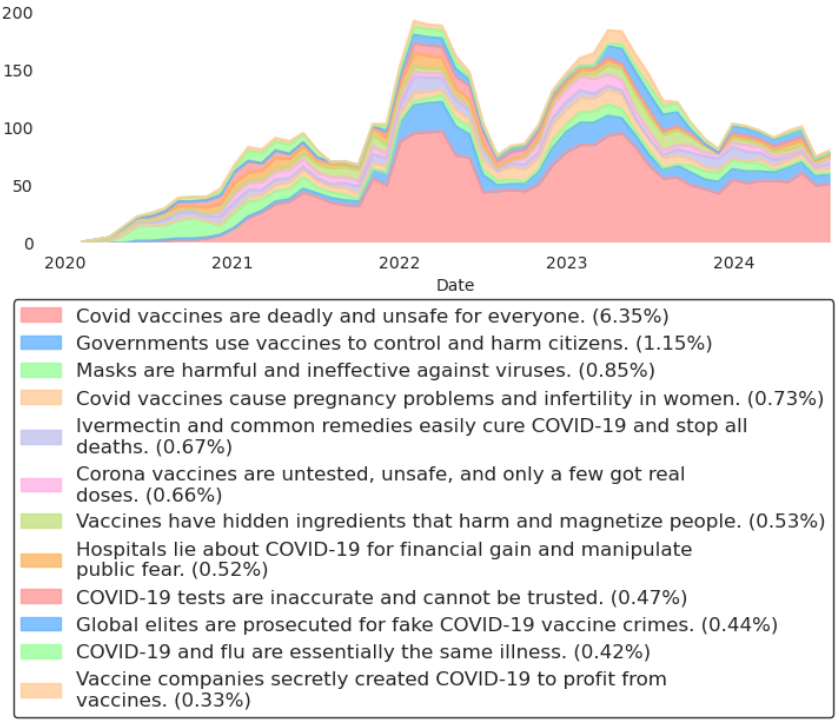

Figure 5.8: Evolution of selected disinformation narratives discovered by DiNaM on the COVID-19 topic.

5.5.3 Ukraine-Russia War Disinformation Narratives

Figure 5.9 presents the selected disinformation narratives from the Ukraine-Russia war identified by DiNaM. Initially, the claim that the war was "fake" gained traction, but it lost momentum, likely because media coverage confirmed its reality. Disinformation then shifted to discrediting Ukrainian President Zelensky and portraying Ukrainians negatively. This shift might reflect the adaptability of Russian-aligned disinformation campaigns, as noted by Pomerantsev et al. [155].

Our findings further suggest that Ukraine-Russia war narratives are highly adaptable to real-world events. For instance, claims about the Bucha atrocities being staged emerged in mid-2022, coinciding with reports from Bucha [28], and the narrative of Leopard tanks being ineffective followed their delivery to Ukraine [3]. A similar pattern occurred with negative portrayals of Zelensky during his speech at the 78th UN General Assembly [126].

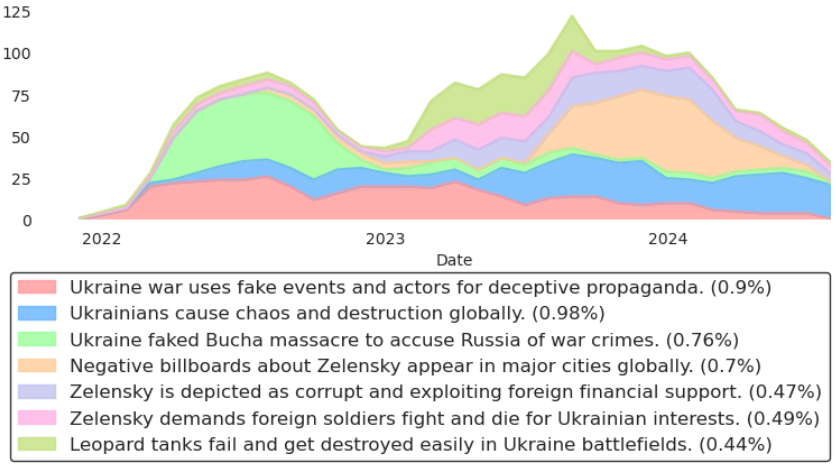

Figure 5.9: Evolution of selected disinformation narratives discovered by DiNaM on the Ukraine-Russia war topic.

5.5.4 Climate Change Disinformation Narratives

Figure 5.10 shows a sharp rise in climate change disinformation starting in early 2022, led by claims that "Climate change is fake". Other themes include critiques of electric cars, conspiracy theories, and assertions that natural disasters are deliberately engineered. Rising energy costs and government incentives for green technologies [136] likely fuelled scepticism. In 2023, severe heatwaves and wildfires [37] likely fuelled narratives asserting that wildfires result from human activity unrelated to climate change.

5.5.5 Migrants Disinformation Narratives

DiNaM uncovered several disinformation narratives related to migration (Figure 5.11). The most prominent claim that immigrants receive more financial support than locals, while others link migrants to rising violence in Europe. These narratives surged in early 2022, likely driven by geopolitical instability and the economic fallout of COVID-19 [178]. The Russian invasion of Ukraine further strained European infrastructure, especially in the East [85].

In 2023, anti-migrant narratives framed migrants as public security threats, gaining traction during events like the June 2023 riots in France. These sentiments intensified in 2024, coinciding with the EU Pact on Migration and Asylum [49].

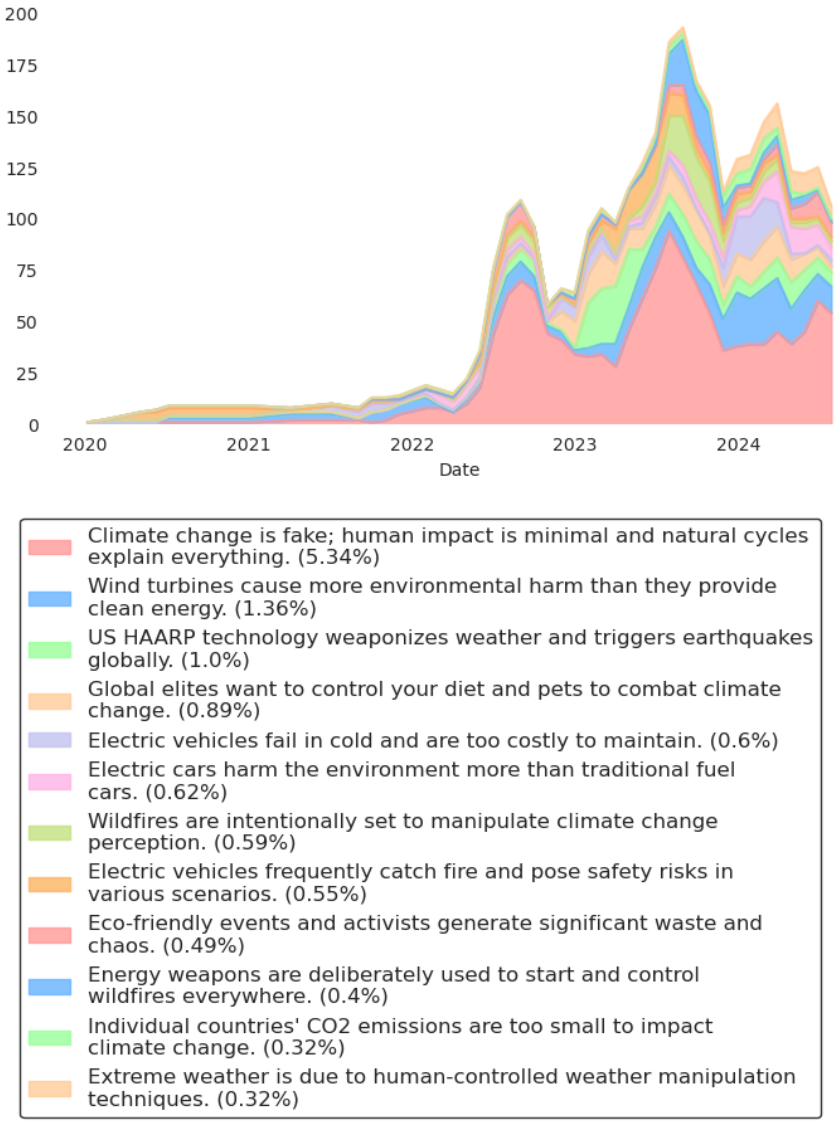

Figure 5.10: Trends in climate change disinformation narratives.

5.5.6 European Union Disinformation Narratives

DiNaM's analysis of disinformation narratives reveals a surge in targeting the European Union in recent years, particularly during critical events such as the European Parliamentary elections in June 2024 [51]. Figure 5.12 illustrates the evolution of these narratives. A prominent example includes claims that "Europe is mismanaged with excessive taxes, inequality, and poor living conditions", which serves as a broad critique of EU governance. Another widely circulated narrative alleges that "EU hides insects in food to trick people into eating them", a distortion

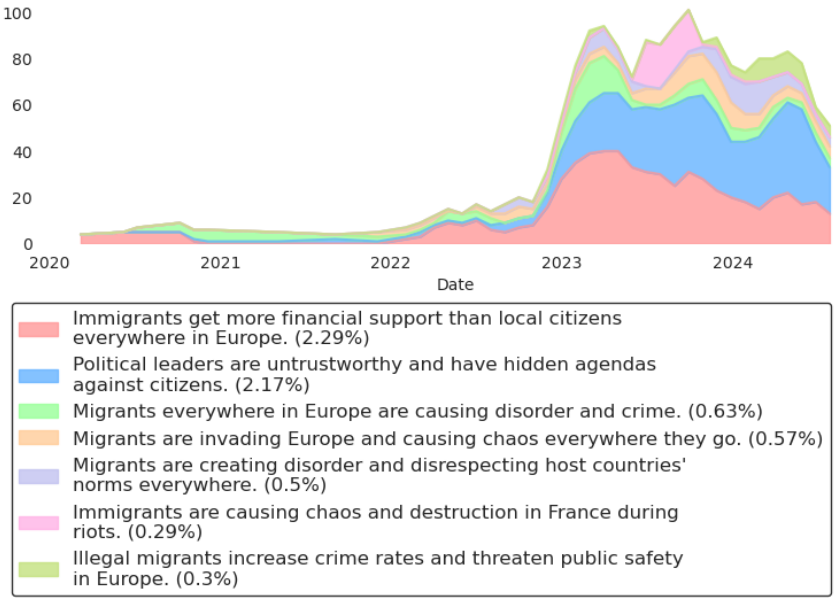

Figure 5.11: Evolution of disinformation narratives on migration.

likely linked to the European Commission's 2023 approval of insect-derived products as voluntary food ingredients [60].

5.5.7 Israel–Palestine War Disinformation Narratives

Several disinformation narratives have emerged in connection to the Israel–Palestine war (Figure 5.13) as identified by DiNaM. One dominant claim suggests that "Gaza war scenes are fake and orchestrated for sympathy", while another asserts that "Global unrest as world protests against Israel supporting Palestine". These narratives gained significant traction in late 2023 and early 2024, aligning with the intensification of the conflict and the international reaction to the humanitarian crisis [6].

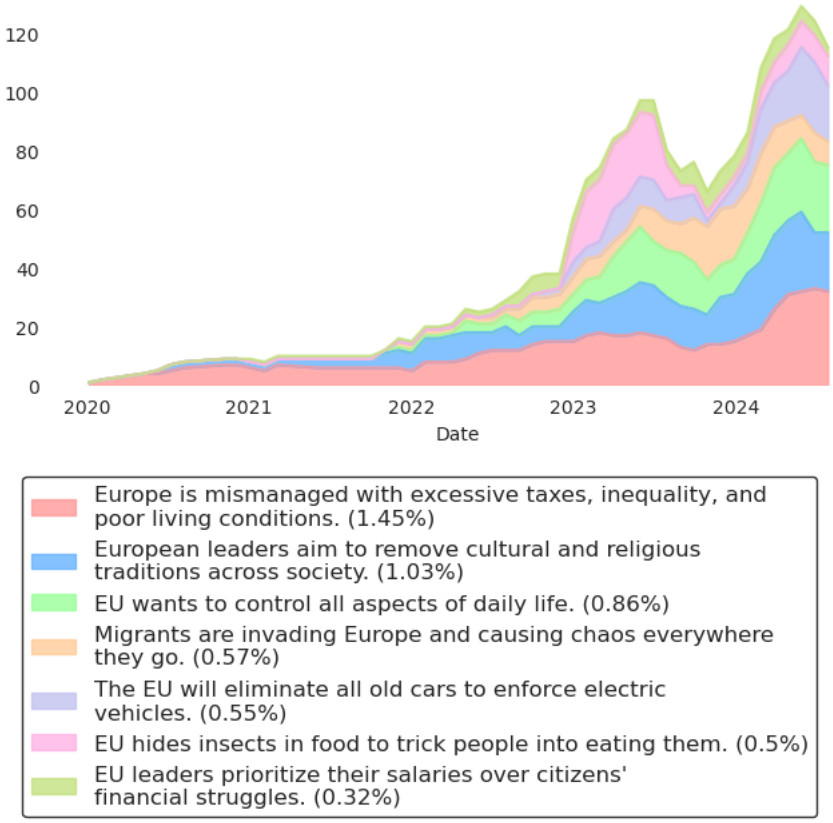

Figure 5.12: Evolution of disinformation narratives on the European Union.

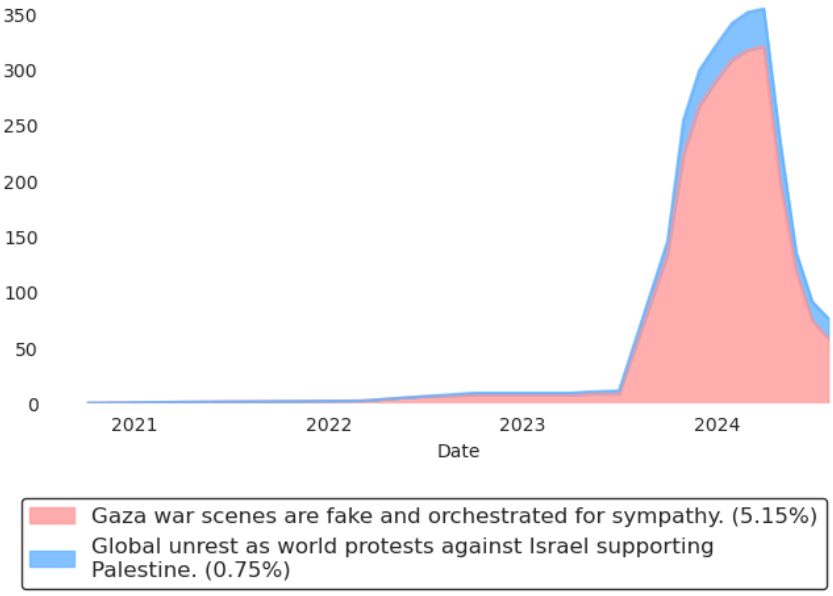

Figure 5.13: Evolution of disinformation narratives on the Israel-Palestine war.

Chapter 6

DiNO: Mining Disinformation Narratives in News Outlets

6.1 Introduction

The previous chapter introduced DiNaM, which applies the disinformation narrative-mining framework to fact-checking articles to reconstruct a global catalogue of disinformation narratives implied by expert verification practice. While this yields a high-level map of the disinformation landscape, it does not directly reveal how disinformation is articulated and propagated within specific news outlets, or how narrative repertoires differ between disinformation-prone and reputable sources over time.

This chapter extends the disinformation narrative-mining framework to the domain of news media. We introduce DiNO, an algorithm for uncovering disinformation narratives directly in news articles. DiNO combines outlet-specific corpora with databases of verified false claims and instantiates the three-stage narrative-mining formulation from Chapter 3 in a way that is tailored to outlet-level analysis: it segments news articles into candidate spans and matches them to verified false claims, retains only segments that entail a verified false claim, clusters these verified false segments in a semantic embedding space, and summarises each cluster into a narrative that can be interpreted as a local disinformation narrative.

Within the overall research programme of the thesis, DiNO thus provides a local complement to the global map constructed by DiNaM. EUDisinfoTest offers a narrative-level reference set of base disinformation narratives, and DiNaM reconstructs their global structure from fact-checking articles. Building on these components, DiNO asks how far the mining strategy can be pushed when starting from raw outlet coverage anchored in verified false claims: it maps which disinformation narratives are present in each outlet, how strongly they are propagated, and

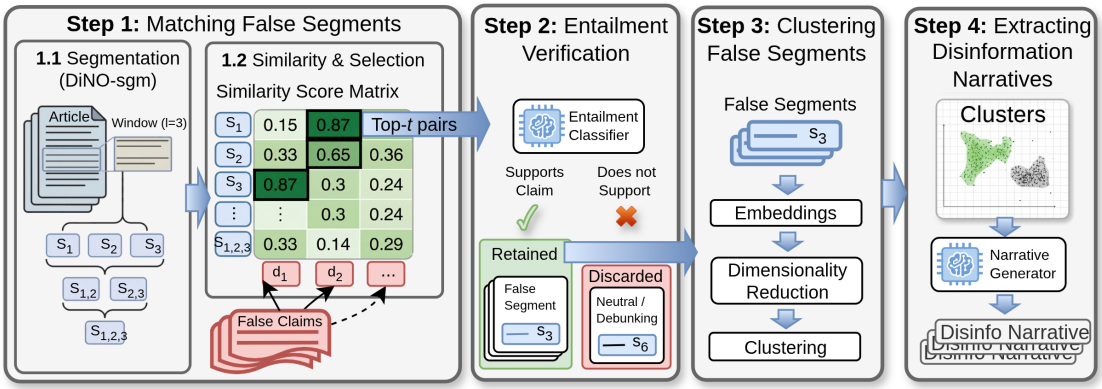

Figure 6.1: The DiNO Pipeline. The approach consists of four phases: (1) *Matching False Segments*, where article text is segmented via the DiNO-sgm method and matched to verified false claims; (2) *Entailment Verification*, where an LLM confirms that the article explicitly supports the false claim; (3) *Clustering False Segments*, which groups the verified false segments based on semantic embeddings; and (4) *Extracting Disinformation Narrative*, where an LLM synthesizes a disinformation narrative from each cluster.

how their prevalence changes in response to major events. Applied to topic-focused corpora that combine disinformation-prone and reputable outlets across domains such as the Ukraine war, COVID-19 and migration, DiNO enables systematic outlet-level comparisons of narrative repertoires and temporal dynamics.

For reproducibility, we release all datasets, prompts, annotations, narratives, and code[1].

6.2 DiNO Algorithm

Figure 6.1 summarises the DiNO pipeline. Starting from outlet-specific news corpora and a database of verified false claims, DiNO (i) identifies article segments that are likely to express claims from the false-claim database, (ii) verifies that these segments actually assert (rather than negate or merely mention) the false information, (iii) groups the retained false-information segments into semantically coherent clusters, and (iv) summarises each cluster into a narrative.

[1] `https://github.com/anonymous-review-phd/anonymous_repo`

Formally, let

$$\mathcal{C} = \{c_1, \ldots, c_n\}$$

denote a set of news articles, and

$$D = \{d_1, \ldots, d_{|D|}\}$$

a database of verified false claims. DiNO maps C to a set of narratives in four phases:

1. **Matching False Segments.** Each article c_i is segmented into overlapping windows of l sentences with an overlap of o sentences. Within each window, all contiguous subsequences of 1 to l sentences are treated as candidate segments, yielding

   $$S_i = \{s_{i,1}, \ldots, s_{i,K_i}\}.$$

   For every candidate segment $s_{i,j} \in S_i$ and claim $d_k \in D$, a similarity score

   $$\text{sim}(s_{i,j}, d_k)$$

   is computed. DiNO retains the top-t most similar segment-claim pairs for each article c_i; we denote this set by

   $$P_i = \{(s_{i,1}, d_{i,1}), \ldots, (s_{i,t}, d_{i,t})\}.$$

2. **Entailment Verification.** For every candidate pair $(s, d) \in P_i$, an entailment model checks whether c_i supports (entails) the matched claim d, ensuring that the false information is asserted in context. Only entailed pairs are kept, and the corresponding segments are collected into

   $$S = \{s_1, \ldots, s_M\},$$

   the final set of segments that contain verified false information.

3. **Clustering False Segments.** Each $s_j \in \mathcal{S}$ is embedded into a semantic vector space (optionally followed by dimensionality reduction), and the resulting vectors are clustered. This yields a set of clusters

   $$\mathcal{K} = \{K_1, \ldots, K_k\},$$

   where each $K_q \subseteq \mathcal{S}$ groups semantically similar false-information segments, potentially drawn from multiple articles and outlets.

4. **Extracting Disinformation Narratives.** For each cluster $K_q \in \mathcal{K}$, a narrative N_q is identified that summarises the recurring disinformation elements shared by its segments. Collectively, these narratives form

   $$\mathcal{N}_{\text{Dis}} = \{N_1, \ldots, N_k\},$$

   a set of disinformation narratives derived from the news coverage.

6.2.1 Matching False Segments

In Step 1 (see Figure 6.1), DiNO identifies segments of each article that may contain previously verified false claims by comparing the article text against a database of verified false claims. Following standard document-level claim-matching, we (i) split each article into candidate segments and (ii) compute their semantic similarity to verified false claims, retaining the top-t scoring segment-claim pairs as candidates for subsequent verification [177]. While sentence-level segmentation is common in prior work [177], it is often too narrow, as false information can span multiple sentences or rely on cross-sentence context [86]. Conversely, longer passages can dilute the similarity signal with irrelevant context [45].

Segmentation. To strike a balance, we introduce **DiNO-sgm**, a two-stage segmentation scheme. First, each article is split into overlapping windows of l sentences with an o-sentence overlap. Then, within each window, all contiguous subsequences ranging from 1 to l sentences are generated.

For example, consider a window of length l=3 containing sentences S_1, S_2, S_3. From this window, six candidate segments are extracted: three single-sentence segments (S_1), (S_2), (S_3); two pairs (S_1, S_2), (S_2, S_3); and one full-span (S_1, S_2, S_3). Compared to fixed-length segmentation methods, DiNO-sgm produces a richer set of candidate segments, offering better coverage of both multi-sentence spans that may contain false information.

Similarity. Next, DiNO computes a semantic similarity score between each candidate segment and every verified false claim in the relevant database. For each article, we retain the top-t segment–claim pairs by similarity score as candidates for the next step.

6.2.2 Entailment Verification

While the previous step matches article segments that are semantically similar to verified false claims, similarity alone is insufficient for reliable false-information detection: NLP models can assign high similarity to text that either supports a claim or argues against it [112]. Therefore, Step 2 performs article-level entailment verification to determine whether the article supports each matched claim and filters out non-supporting matches. Only segments whose articles pass this check are carried forward to the subsequent steps.

6.2.3 Clustering False Segments

A disinformation narrative can span multiple related false claims. In our setting, this means a single narrative may involve several false segments matched to thematically similar false claims. Accordingly, in Step 3 DiNO clusters the false segments extracted in the previous phase. Each segment is first embedded in a semantic vector space, reduced in dimensionality to address the curse of dimensionality [81], and then clustered into sets that represent distinct narratives.

6.2.4 Extracting Disinformation Narratives

In the final step, DiNO uses an LLM to identify the shared falsehood within each cluster and distil it into a representative disinformation narrative. All texts assigned to a cluster are concatenated and provided to the model as a single input. The model is then prompted to synthesise the recurring message across these instances into one narrative.

As in DiNaM (Section 5.2.4), we constrain the output format to match real-world narrative formulations and to support evaluation. We derive these constraints informed by the EUDisinfoTest dataset [184], where most narratives are short, self-contained statements, typically under 15 words. Accordingly, each generated narrative must be: (1) clear, standalone, and descriptive, (2) no longer than 15 words and (3) expressing the false perspective behind the claims.

These constraints also align with the properties of disinformation narratives defined in Section 3.1.3. Moreover, Appendix C.6 provides examples of extracted narratives alongside their supporting false claims.

6.2.5 Prompt Template

DiNO uses a specialised prompt template that closely mirrors the one introduced for DiNaM in Section 5.2.5. The DiNO-specific variant, illustrated in Figure 6.2, also follows the same structure as the prompt template proposed by Lucas et al. [110].

Conceptually, each prompt is composed of four components: an **impersonation** block that assigns the model a specific expert role, **guidelines** that describe the task and constrain the output, a **context** block supplying the relevant article segments or clusters, and an **output** specification that fixes the required response format (including the narrative). This reusable structure keeps DiNO's prompts

consistent across its LLM-based components while allowing component-specific specialisation in the guidelines and output format.

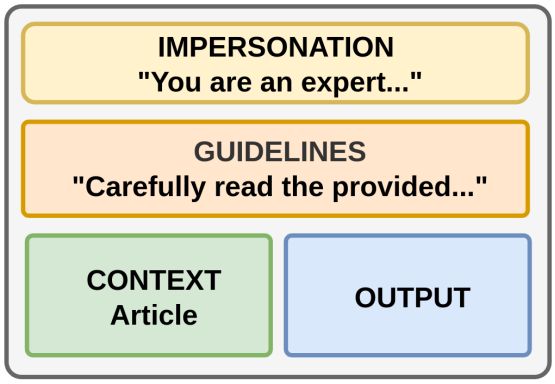

Figure 6.2: Overview of the prompt template, comprising: (1) Impersonation, to define the model's role; (2) Guidelines, to specify instructions; (3) Context, to embed task-relevant information; and (4) Output, to enforce a response format.

6.3 Datasets of News Articles and False Claims

DiNO is evaluated on news from a diverse set of outlets, encompassing both sources repeatedly flagged as disinformation-prone and established reputable media, together with domain-matched collections of verified false claims. The evaluation covers three domains: the Ukraine War, COVID-19 and migration. For each domain, we construct corpora from disinformation-prone outlets, one corpus from a reputable outlet, and one collection of expert-verified false claims. Table 6.1 summarises all datasets, their sizes, and sources.

Across all domains, these collections comprise 88,795 articles from disinformation-prone outlets and 34,085 articles from the reputable outlet, along with a total of 26,285 verified false claims. Together, these corpora provide the raw material for DiNO's four phases (Section 6.2): news articles serve as the source of candidate segments, while the verified false claims act as anchors that determine which segments are treated as false.

6.3.1 Data Collection Process

To construct the datasets from news outlets, we first scraped article URLs using Selenium [173], automating browsing from the "Website URL" entries listed

Dataset	Type	Items	Website URL
Ukraine War			
Sputnik	Disinformation-prone	20,001	[188]
Pravda	Disinformation-prone	22,695	[156]
The Guardian	Reputable outlet	11,841	[84]
EUvsDisinfo	Verified false claims	15,452	[72]
COVID-19			
Sputnik	Disinformation-prone	13,107	[187]
NaturalNews	Disinformation-prone	16,789	[127]
The Guardian	Reputable outlet	10,573	[82]
GFCE	Verified false claims	6,381	[77]
Migration			
Breitbart	Disinformation-prone	16,203	[34]
The Guardian	Reputable outlet	11,671	[83]
GFCE	Verified false claims	4,452	[77]

Table 6.1: Summary of datasets used in DiNO across three domains, with their type, number of items, and sources.

in Table 6.1. With these URLs, we then extracted clean article text using the `trafilatura` tool [26].

For verified false-claim datasets, we used two approaches. For EUvsDisinfo (Ukraine-War domain), we scraped entry URLs with Selenium [173] and parsed each entry using Beautiful Soup 4 [164]. We extracted three key fields: (1) the "Disinformation Summary" (i.e., the false claim), (2) the original disinformation article URL, and (3) the disproof text including debunking articles URL. We then retrieved the full text of both the disinformation and debunking articles using `trafilatura`.

The GFCE dataset was constructed by retrieving COVID-19- and Migration-related fact-checks from the Google Fact Check Explorer via its official API [2].

Table 6.2 shows that, across domains, disinformation-prone outlets tend to publish shorter and more homogeneous articles, whereas Guardian articles are longer on average and exhibit greater variance in length. This difference in style

[2] `https://developers.google.com/fact-check/tools/api`

Source	Count	Avg. words	Std. dev.
Ukraine War			
Pravda	22,695	223.08	240.20
Sputnik	20,001	585.55	410.49
The Guardian	11,841	886.73	779.80
COVID-19			
Sputnik	13,107	440.07	323.25
NaturalNews	16,789	836.69	474.41
The Guardian	10,573	820.21	612.34
Migration			
Breitbart	16,203	512.81	381.17
The Guardian	11,671	866.51	669.53

Table 6.2: Descriptive statistics for the news datasets used in DiNO. *Count* is the number of articles per source, *Avg. words* is the mean article length, and *Std. dev.* denotes the standard deviation of article length.

and depth is relevant for DiNO, since the segmentation and matching phases must operate robustly across more elaborated reporting.

6.3.2 Disinformation-Prone Outlets

The disinformation-prone part of the corpus covers four outlets that have been repeatedly identified in policy and monitoring reports as prominent vectors of misleading or false content: Sputnik, Pravda, NaturalNews and Breitbart [89, 176, 34].

In the Ukraine War domain, DiNO uses 20,001 articles from Sputnik and 22,695 from Pravda, both of which devote extensive coverage to the Russian invasion of Ukraine and related geopolitical developments [36, 133]. For COVID-19, the disinformation-prone corpus combines 13,107 articles from Sputnik with 16,789 from NaturalNews, a site known for health-related conspiracy theories and vaccine scepticism [71]. In the migration domain, DiNO relies on 16,203 articles from Breitbart, whose coverage has frequently been criticised for alarmist and xenophobic framing of immigration and for propagating racist discourses [50].

6.3.3 Reputable Outlets

As a credible baseline, domain-matched corpora are constructed from The Guardian, a mainstream UK outlet widely recognised as a leading quality newspaper and rated highly for factual reporting and editorial reliability [**?**]. For each domain, Guardian articles are filtered by topic-specific queries to obtain coverage that is comparable in scope to the disinformation-prone outlets.

For the Ukraine War, the dataset contains 11,841 Guardian articles focusing on the conflict and its international context. In the COVID-19 domain, DiNO uses 10,573 articles addressing the pandemic, vaccines and related public-health policies. For migration, the corpus comprises 11,671 articles discussing immigration, asylum, border control and broader migration politics.

Narratives extracted by DiNO from reputable corpora serve as a sanity check: a well-calibrated system should identify very few narratives in these outlets that strongly align with known disinformation narratives.

6.3.4 Verified False-Claim Resources

The final component of DiNO's data setup consists of verified false-claim resources that ground the mining process in expert fact-checking. Across the three domains, two complementary sources are used: the EUvsDisinfo database, which systematically catalogues pro-Kremlin disinformation cases, and fact-checks retrieved via the Google Fact Check Explorer (GFCE) API, which provides structured access to ClaimReview-based fact-check results [66, 76].

For the Ukraine War, verified false claims are drawn from the EUvsDisinfo database maintained by the EU's East StratCom Task Force, a unit of the European External Action Service tasked with identifying and countering pro-Kremlin disinformation [65]. Each entry provides a textual Disinformation Summary that describes the false claim, together with the corresponding disproof text and links to the original disinformation material as well as to related debunking articles, all curated in a continuously updated open database of cases and disproofs [65]. The cleaned release used in this work contains 15,452 unique entries.

For COVID-19 and migration, two additional false-claim collections are constructed using the GFCE API, retrieving ClaimReview fact-checks whose topics or keywords match the respective domains and retaining only items whose verdict indicates falsity (e.g., False, Manipulated) [76]. The Fact Check Tools API and asso-

ciated Explorer service expose the same underlying set of ClaimReview fact-checks that can be queried programmatically by topic or keyword [76]. Applying these domain-specific filters yields 6,381 verified false claims for COVID-19 and 4,452 for migration, again in the form of claim descriptions linked to the corresponding source and debunking articles.

6.4 Evaluation

This section presents the evaluation methodology that guided the design and parameter choices in DiNO. We first conduct module wise experiments for the four main stages of the pipeline: matching false segments, entailment verification, clustering false segments and extracting disinformation narratives. In these experiments we compare alternative models, similarity measures and hyperparameters on held out data. The best performing configurations from this process define the final DiNO setup used in the subsequent outlet level analyses.

Once the pipeline is calibrated, we apply DiNO to articles from all outlets and domains. The resulting narratives are then compared against reference disinformation narratives, with a particular focus on topical and stance alignment. Throughout, we adopt the working assumption that a more effective mining procedure should yield stronger alignment with disinformation narratives in disinformation-prone outlets, and weaker alignment in reputable outlets.

To ensure robustness, all LLM-based components rely on a standardised prompt template, and reported metrics are averaged over three independent runs. Where applicable, we also include lightweight and non-LLM baselines, both to quantify the gains from LLM-based components and to provide a broader performance context.

6.4.1 Matching False Segments

Step 1 locates the article segments that express false claims by (i) segmenting the article into candidate segments and (ii) matching each segment against a pool of verified false claims using a similarity measure. Thus, we evaluate Step 1 by testing how well DiNO retrieves the correct claim under different segmentation strategies and similarity measures.

Testing Dataset. To perform this evaluation, we need articles paired with gold-standard false claim(s). In EUvsDisinfo, entries link verified false claims to source articles, but the same article can appear in multiple entries. We therefore grouped entries by source article, so each article is associated with all of its linked false claims. We then sampled 180 source articles, yielding a test set in which some articles are associated with multiple claims (1.3 claims per article on average).

Segmentation Strategies. We evaluated the following segmentation strategies: (1) DiNO-sgm (ours); (2) sentence-level: each sentence as a segment; (3) passage-level: fixed-length segments; (4) article-level: full article as a segment.

Similarity Measures. We tested three similarity measures: (1) BM25 [166]; (2) BERTScore [207]; (3) cosine similarity over embeddings from SFR-Embedding-2 (SFR) [168], E5-large (E5) [200], and Jina [191].

Evaluation. For each article in the test set, we (1) apply the chosen segmentation strategy; (2) compare every segment to all 15,452 EUvsDisinfo claims using the chosen similarity measure; (3) rank segment-claim pairs by similarity score; and (4) evaluate the ranked list by computing mAP@3 [114] with respect to the article's gold-standard claims.

Hyperparameter search. We tuned the segment length l, the overlap o and retrieval depth t on a development set constructed similarly to evaluation set, but disjoint. We compared three embedding models (SFR-Embedding-2, E5-large, and Jina) and tuned DiNO-sgm hyperparameters: segment length l, overlap o, and retrieval depth t. For each article, we applied a sliding window of l sentences with o-sentence overlap, retrieved candidate segment–claim matches with FAISS IndexIVFFlat ANN search, and kept the top-t pairs per article by similarity. Table 6.3 lists the search ranges, and Table 6.4 reports the best settings.

We use t=3 in all experiments since it controls the number of Step 2 entailment checks and therefore runtime/cost. Larger t gave only small gains, so t=3 is a practical tradeoff (Table 6.5).

Results. Table 6.6 summarizes the results. Under Jina model, DiNO-sgm with tuned parameters (l=4, o=1, t=3), achieves the highest mAP@3 across similarity measures. We attribute this gain to DiNO-sgm's variable-length segments, which

Hyperparameter	Search Range
Segment length l	$\{3, 4, 5, 6, 7, 8, 9, 10\}$
Overlap o	$\{0, 1, \ldots, l-1\}$
Retrieval depth t	$\{1, 2, \ldots, 5\}$

Table 6.3: Hyperparameter ranges for Step 1, used during tuning on the Sputnik-Ukraine dataset.

Model	l	o	t
SFR–Embedding–2	6	2	3
E5–large	5	1	3
Jina	4	1	3

Table 6.4: Best hyperparameter values for each embedding model for matching false segments in the Sputnik-Ukraine dataset.

better match the span of false information than fixed-length alternatives. We therefore use DiNO-sgm as DiNO's default segmentation module.

Figure 6.3 further illustrates this effect: as the number of segments increases (controlled by o and l), mAP@3 improves before plateauing, reflecting a trade-off between coverage of potentially relevant spans and the introduction of noisy candidates. Examples of both successful and unsuccessful matches are provided in Appendix C.5.

On the basis of these experiments, DiNO adopts the DiNO-sgm segmentation strategy and Jina-based cosine similarity with l=4, o=1 and t=3 as its default configuration for matching false segments.

6.4.2 Entailment Verification

Step 1 retrieves segments that are textually similar to verified false claims, but similarity can also be high when an article mentions a claim to debunk it. Step 2 therefore checks the article's stance toward the claim and keeps the segment only if the article supports it; otherwise, it is filtered out before clustering and narrative extraction.

Retrieval depth t	mAP@3
1	0.57
2	0.62
3	0.64
$\geq$4	0.64-0.65

Table 6.5: Sensitivity of Step 1 retrieval performance (mAP@3) to retrieval depth t, using DiNO-sgm with fixed segmentation hyperparameters (l=4, o=1) and cosine similarity with Jina.

Segmentation	Cosine	BERTScore	BM25
Our (DiNO-sgm)	**0.64** (Jina)	0.55	0.48
Sentence-level	0.56 (Jina)	0.47	0.40
Passage-level	0.47 (E5)	0.42	0.38
Article-level	0.38 (SFR)	0.32	0.34

Table 6.6: mAP@3 by segmentation strategy and similarity metric. For Cosine similarity, the best embedding model for each strategy is shown in parentheses.

Testing Dataset. To evaluate stance/entailment, we require claim-article pairs where some articles support the claim and others do not. We leverage EUvsDisinfo, which links verified false claims to real-world news articles that promote it and to debunking articles that refute it. Using these links, we build a test set of 450 claim-article pairs: 150 supporting, 150 debunking, and 150 unrelated (randomly paired across entries). For evaluation, only the supporting pairs are labeled support; debunking and unrelated pairs are labeled not supporting.

LLM Models. We evaluate seven LLMs in a zero-shot setting,[3] comprising four closed-weight models:GPT-5.2 [139], Claude Sonnet 4.5 [19], Gemini 2.5 Pro [78], and GPT-5-mini [140], and three open-weight models:gpt-oss-120b [8], Llama-3.3-70B [120], and Qwen3-14B [206]. We use three custom prompts and three general-purpose entailment prompts adapted from prior work [74]. Full prompt formulations are provided in Appendix C.1.1.

[3]We do not use few-shot classification [35] because including full article examples is impractical.

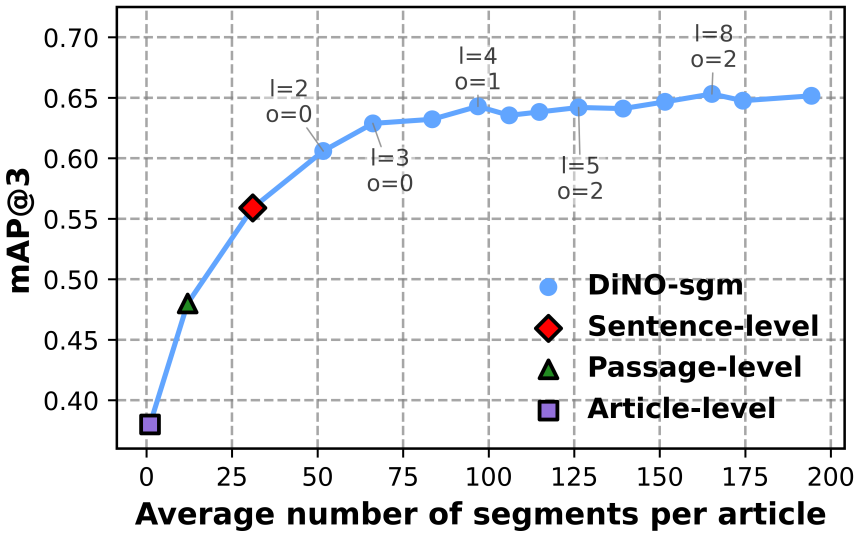

Figure 6.3: mAP@3 performance of DiNO–sgm (Jina) across segmentation settings with retrieval depth t=3. The figure shows how varying overlap (o) and segment length (l) affects matching accuracy and the number of segments produced. Results from other segmentation methods are included for baseline comparison.

Non-LLM Models. To provide a non-LLM baseline, we also fine-tuned DeBERTa-large [88] and RoBERTa-large [108] on the same 450 claim–article pairs. Each input consisted of the false claim and the article text concatenated with a separator token, and was labelled according to whether the article supports or does not support the claim. Both models were trained for five epochs with binary cross-entropy loss and AdamW (batch size 8, learning rate 2e-5), and evaluated using 5-fold cross-validation. We report the mean F_1 scores across folds.

Evaluation. For each claim-article pair, the model predicts whether the article *supports* the claim or *does not support* it (support expected for disinformation articles; non-support expected for other articles). We report performance using the F_1 score and precision on the (*support*) class (P_{support}).

Results. Table 6.7 summarises the main results, reporting for each LLM its best-performing prompt alongside the non-LLM baselines. In the zero-shot setting, GPT-5.2 and Sonnet 4.5 achieve the top performance (F_1= 0.94, P_{support}= 0.90). They are closely followed by Gemini 2.5 Pro, GPT-5-mini, and gpt-oss-120b(F_1= 0.93, $P_{\text{support}} \in [0.89, 0.88]$). In contrast, the fine-tuned DeBERTa-large and RoBERTa-

large baselines reach F_1= 0.63 and 0.61, respectively, substantially below the zero-shot LLM range. Overall, these results indicate that performance is driven primarily by underlying model capability: strong LLMs paired with suitable prompting outperform supervised encoder-based models even without task-specific training data. Since entailment verification is the most token-intensive step and GPT-5-mini offers near-top performance at lower cost, we use it as the default entailment model in subsequent experiments.

Sanity checks and additional analyses. Appendix C.2 reports two checks that support using full-article entailment verification. First, when Step 2 confirms that an article supports a claim, the segment retrieved in Step 1 usually contains an explicit mention of that claim proposition. Second, verifying entailment using only the retrieved segment (instead of the full article) substantially reduces performance; adding a small amount of surrounding context narrows the gap but does not fully close it.

Model	Prompt	F_1 score	$P_{support}$
Zero-shot			
Sonnet 4.5	Fig. C.2	**0.94**	**0.90**
GPT-5.2	Fig. C.3	**0.94**	**0.90**
Gemini 2.5 Pro	Fig. C.1	0.93	0.89
GPT-5-mini	Fig. C.1	0.93	0.88
gpt-oss-120b	Fig. C.3	0.93	0.88
Llama-3.3-70B	Fig. C.1	0.91	0.87
Qwen3-14B	Fig. C.1	0.88	0.84
Fine-tuned			
DeBERTa-large	-	0.63	0.55
RoBERTa-large	-	0.61	0.54

Table 6.7: F_1 and $P_{support}$ scores of zero-shot LLMs with base prompt vs. fine-tuned encoder baselines on the entailment verification dataset (Step 2 of DiNO).

6.4.3 Clustering of False Segments

The output of Step 2 is a set of false segments. Step 3 then groups these segments into coherent clusters that represent candidate narratives. DiNO does this by embedding each segment, reducing dimensionality, and then applying a

clustering algorithm. Therefore, to evaluate Step 3, we test how clustering quality varies with the embedding model, dimensionality reduction, and clustering method.

Testing Dataset. To evaluate clustering, we need a collection of segments that contain supported false information. We therefore apply Steps 1 and 2 (using optimal settings) to the Sputnik-Ukraine dataset and obtain 11,225 false segments.

Methods. We evaluated three embedding models (SFR-Embedding-2, E5-large, Jina), two dimensionality reduction methods (UMAP [118], PCA [4]) and two clustering algorithms (HDBSCAN [117], K-means [87]).

Evaluation. We performed a grid search over all combinations of embedding models, dimensionality reduction techniques and clustering algorithms. For each configuration, we computed the Silhouette Score [167] on the resulting clustering of the 11,225 false-information segments. Since no gold cluster labels exist for this unsupervised setting, the Silhouette Score provides a proxy for how well-separated and coherent the resulting clusters are.

Hyperparameter search. For UMAP, PCA, HDBSCAN and K-means, we carried out a grid search over a broad range of hyperparameters on the false-information segments described above. Table 6.8 summarises the ranges considered and Table 6.9 presents the selected optimal values for each algorithm. These optimal settings are then used for the comparative results reported below.

Algorithm	Hyperparameter	Range
UMAP	n_neighbors	$\{10, 15, 30, 50, 100\}$
UMAP	n_components	$\{4, 8, 16, \ldots, 256\}$
PCA	n_components	$\{4, 8, 16, \ldots, 256\}$
HDBSCAN	min_cluster_size	$\{10, 15, 20, \ldots, 100\}$
HDBSCAN	min_samples	$\{10, 15, 20, \ldots, 100\}$
K-means	n_clusters	$\{5, 15, 25, \ldots, 800\}$

Table 6.8: Hyperparameter ranges and selected optimal values for dimensionality reduction and clustering methods applied to false segments extracted from the Sputnik Ukraine-War dataset and matched with the EUvsDisinfo false-claims collection.

Algorithm	Optimal Hyperparameters
HDBSCAN (UMAP)	min_cluster_size = 30 min_samples = 30
K-Means (UMAP)	n_clusters = 250
UMAP (HDBSCAN)	n_neighbors = 25 n_components = 256
UMAP (K-Means)	n_neighbors = 30 n_components = 256

Table 6.9: Optimal hyperparameters for clustering methods applied to false segments from the Sputnik-Ukraine dataset.

Results. Table 6.10 summarises the results. The best performance is achieved by the combination of SFR-Embedding-2, UMAP and HDBSCAN, with a Silhouette score of 0.77. Across all settings, HDBSCAN consistently outperforms K-means, and UMAP outperforms PCA. This aligns with prior findings that HDBSCAN is well suited to complex, high-dimensional data due to its ability to detect clusters of varying shapes and filter noise [117], while UMAP tends to preserve both global and local semantic structure more effectively than the linear projections assumed by PCA [170]. On this basis, SFR-Embedding-2 + UMAP + HDBSCAN is adopted as the default clustering configuration for DiNO.

Finally, we complement the Silhouette analysis with a small human evaluation: Appendix C.3 shows that clusters produced by the selected configuration are consistently rated as more topically coherent than strong alternatives.

6.4.4 Extracting Disinformation Narratives

In Step 4, DiNO converts each cluster of false segments (Step 3) into a single disinformation narrative. We therefore evaluate Step 4 by holding the clusters fixed and varying only the narrative generator, then comparing the resulting narratives to expert-written narrative references.

Testing Dataset. We reuse the Sputnik-Ukraine clusters produced in Section 6.4.3 under the best clustering configuration. As ground truth, we use EUDisinfoTest [184], which provides expert-curated disinformation narratives grounded in reports from 20 countries and spanning the domains analyzed in this

Embedding Model	UMAP	PCA
SFR-Embedding-2		
HDBSCAN	**0.77**	0.60
K-Means	0.60	0.43
E5-large		
HDBSCAN	0.76	0.55
K-Means	0.61	0.46
Jina		
HDBSCAN	0.73	0.52
K-Means	0.59	0.41

Table 6.10: Silhouette scores across embedding models, dimensionality reduction, and clustering methods.

work: COVID-19, Migration, and Ukraine War.

LLM Models. To generate narratives for each cluster, we use the same seven LLMs as in the entailment experiments: four closed-weight models (GPT-5.2, Sonnet 4.5, Gemini 2.5 Pro, GPT-5-mini) and three open-weight models (gpt-oss-120b, Llama-3.3-70B, Qwen3-14B). Each model is guided by a custom narrative-generation prompt (Appendix C.1.2), and DiNO outputs one narrative per cluster for each model.

Non-LLM Models. To provide non-LLM baselines, we also generate narratives from the clusters defined above using two extractive methods: BERTsum [158] and TextRank [121]. Both methods take all segments within a cluster as input and select representative sentences to form a summary, without leveraging generative LLM capabilities.

Baseline Methods. To further contextualise DiNO's performance, we compare it against two state-of-the-art general narrative extraction methods: CaNarEx [18] and Relatio [20]. Both methods operate directly on article-level text, without restricting attention to false-information segments. They cluster semantically similar sentences and represent each cluster by selecting a single sentence as the

narrative. In contrast, DiNO first filters for false-information segments and then prompts an LLM to synthesise a narrative that captures the disinformation content of each cluster.

Evaluation. For each model, we compare its set of predicted narratives to the ground-truth set. Notably, there is no one-to-one mapping between these sets, we therefore evaluate at the set level: for each predicted narrative n_i^p, we match the most semantically similar ground-truth narrative $n_{m(i)}^{gt}$.

We score each matched pair along two dimensions: **Topical Alignment** (TA), which measures semantic overlap, and **Stance Alignment** (SA), which measures whether both narratives express the same viewpoint. SA is necessary because two narratives may be topically similar yet take opposite positions [31] (e.g., "*COVID-19 is engineered*" vs. "*There is no evidence that COVID-19 is engineered*"). Following prior work, we score TA and SA on a 0–5 scale [12] and report normalized mean alignment:

$$mTA = \frac{1}{5N} \sum_{i=1}^{N} TA(n_i^p, n_{m(i)}^{gt}) \tag{6.1}$$

$$mSA = \frac{1}{5N} \sum_{i=1}^{N} SA(n_i^p, n_{m(i)}^{gt}), \tag{6.2}$$

where N is the number of predicted narratives.

We also report **Redundancy**, which measures how often DiNO repeats the same ground-truth narrative across multiple predictions. We compute

$$Rd = \frac{N}{U}, \tag{6.3}$$

where U is the number of unique matched ground-truth narratives. $Rd \approx 1$ indicates low redundancy, while larger values indicate more repetition.

Scoring. We apply the protocol above with two annotators with experience at International Fact-Checking Network-certified organizations. They perform the matching and provide TA/SA scores. Disagreements are resolved by consensus. We treat these human annotations as the primary evaluation and report them as mTA_H and mSA_H. Annotation guidelines are provided in Appendix C.7.

As a scalable secondary evaluation, we also compute TA/SA using an LLM-as-a-judge [209] and report mTA_A and mSA_A. We use Sonnet 4.5 as the judge due to its strong agreement with human annotations (Appendix C.8).

Results. Table 6.11 summarizes the scores. GPT-5.2 achieves the best overall performance, with Sonnet 4.5 close behind. We therefore select GPT-5.2 as the default generator to ensure consistency across experiments. Notably, the strongest open-weight model (gpt-oss-120b) performs comparably to the closed-weight models.

Method	Human			Automated		
	mTA_H	mSA_H	Rd_H	mTA_A	mSA_A	Rd_A
Closed-weight LLM						
GPT-5.2	**0.76**	**0.85**	1.78	**0.79**	0.83	1.58
Sonnet 4.5	0.75	0.84	1.96	0.77	**0.86**	1.72
Gemini 2.5 Pro	0.74	0.79	1.97	0.75	0.82	1.54
GPT-5-mini	0.73	0.83	1.79	0.72	0.82	1.52
Open-weight LLM						
gpt-oss-120b	0.74	0.82	1.91	0.75	0.84	1.60
Qwen3-14B	0.70	0.73	1.99	0.70	0.76	1.62
Llama-3.3-70B	0.67	0.79	1.95	0.70	0.85	1.58
Non-LLM						
BERTsum	0.61	0.58	2.00	0.63	0.63	1.99
TextRank	0.63	0.55	1.87	0.60	0.58	2.08

Table 6.11: Alignment scores and redundancy between extracted narratives and ground-truth narratives on the Sputnik-Ukraine dataset, across LLM generators.

6.4.5 Benchmarking Against Baseline Methods

DiNO is, to our knowledge, the first method designed specifically to extract disinformation narratives from news articles using verified false claims. To contextualize its performance, we compare against two competitive narrative extraction methods that operate directly on news text: CaNarEx [18] and Relatio [20]. We run both methods on the Sputnik-Ukraine outlet and evaluate their outputs using the same protocol (Section 6.4.4).

Results. Table 6.12 shows that DiNO substantially outperforms both baselines in topical and stance alignment. Our ablations (Appendix C.4) suggest that the gains stem from two design choices: (i) DiNO applies entailment filtering to focus

clustering and summarization on supported false-information segments, rather than processing full articles that contain substantial non-narrative or debunking content; and (ii) DiNO synthesizes narratives from clusters using an LLM, whereas CaNarEx and Relatio represent each cluster via a small set of extracted sentences, which is brittle when cluster content is heterogeneous. Importantly, when the baselines are granted access to DiNO's filtered segments and/or an LLM-based synthesis step, their performance improves markedly, but DiNO remains the strongest overall (Appendix C.4).

Method	**Human**		**Automated**	
	mTA_H	mSA_H	mTA_A	mSA_A
DiNO	**0.76**	**0.85**	**0.79**	**0.83**
CaNarEx	0.52 (**-32%**)	0.60 (**-29%**)	0.54 (**-32%**)	0.63 (**-24%**)
Relatio	0.53 (**-30%**)	0.65 (-24%)	0.53 (**-33%**)	0.61 (**-27%**)

Table 6.12: Alignment scores between extracted and ground-truth narratives on the Sputnik-Ukraine outlet.

6.4.6 Evaluation Across All Datasets

After finding best parameters for each component (Sections 6.4.1–6.4.4), we run DiNO end-to-end on every dataset and compare the resulting narrative sets to EUDisinfoTest. We use the same set-based matching as in Section 6.4.4: each predicted narrative is paired with its most semantically similar reference narrative, and we report topical alignment (mTA), stance alignment (mSA), and redundancy (Rd) under both human and automated evaluation.

Interpreting outputs. On disinformation-prone outlets, we expect extracted narratives to align with expert narratives in both topic and stance and treat these as true positives. On credible outlets, outputs that are topically aligned but stance-opposed (high mTA, low mSA) reflect topic-relevant reporting without endorsement, which is consistent with credible reporting that mentions false claims for context or refutation; we treat these as false positives.

Results. Table 6.13 summarizes human and automated scores. On disinformation-prone outlets, DiNO extracts narratives that are both topically aligned and stance-

aligned with expert ground-truth narratives. On credible outlets, DiNO produces only two topically aligned but stance-opposed narratives per domain.

Beyond alignment, DiNO yields compact narrative sets with low redundancy (human: 1.44, automated: 1.45). This is consistent with our clustering objective, which maximizes the silhouette score to reduce both over-splitting and over-merging.

Finally, our automated judge is a reliable proxy for expert scoring: automatic ratings show strong agreement with human judgments ($r_{WG} = 0.84$ topical, $r_{WG} = 0.78$ stance)[4], validating automated evaluation for large-scale experiments.

Dataset	**NC**	**Human**			**Automated**		
		mTA_H	mSA_H	Rd_H	mTA_A	mSA_A	Rd_A
Ukraine							
Sputnik	66	0.76	0.85	1.78	0.79	0.83	1.58
Pravda	44	0.77	0.90	1.83	0.78	0.84	1.76
Guardian	2	0.90	0.00	1.00	0.80	0.20	1.00
COVID-19							
NatNews	60	0.81	0.95	1.32	0.89	0.91	1.50
Sputnik	9	0.85	0.72	1.20	0.91	0.71	1.29
Guardian	2	1.00	0.00	1.00	0.70	0.20	1.00
Migration							
Breitbart	65	0.77	0.86	2.42	0.83	0.83	2.50
Guardian	3	1.00	0.00	1.00	0.73	0.20	1.00

Table 6.13: Human and automated topical and stance alignment scores between DiNO-extracted and ground-truth disinformation narratives, together with redundancy. **NC**: number of narratives extracted per dataset.

6.4.7 Coverage and Alignment

We further assess DiNO's robustness with a set-level Chamfer-style metric [27] that captures not only the match quality but also coverage between the predicted and ground-truth narrative sets. The formulation is inspired by the set-level matching logic used by Weighted Chamfer Distance in DiNaM (see Section 5.4.4), but replaces embedding distances with LLM-judged alignment scores.

[4]We use r_{WG} as it is recommended for assessing agreement on Likert scales [138].

Evaluation Protocol. Let P be the set of DiNO-extracted narratives (see Section 6.4.6) and G the ground-truth set. Using an LLM judge (Sonnet 4.5), we perform nearest-neighbor matching in both directions: for each $p \in P$, the judge selects the closest $g \in G$, denoted $\mathrm{NN}_G(p)$, and for each $g \in G$, the judge selects the closest $p \in P$, denoted $\mathrm{NN}_P(g)$. For every matched pair, the judge rates thematic alignment (TA) and stance alignment (SA) on a 0–5 Likert scale, and we normalize each rating to $[0, 1]$ by dividing by 5 (prompts in Appendix C.1.3).

We then define Chamfer-style alignment scores (higher is better) as the mean nearest-neighbor alignment in both directions:

$$CD_x(P, G) \;=\; \tfrac{1}{2}\Big(\tfrac{1}{|P|}\sum_{p\in P} a_x\big(p, \mathrm{NN}_G(p)\big) \;+\; \tfrac{1}{|G|}\sum_{g\in G} a_x\big(\mathrm{NN}_P(g), g\big)\Big),$$

where $x \in \{\mathrm{TA}, \mathrm{SA}\}$ and $a_x(\cdot, \cdot) \in [0, 1]$ is the normalized alignment score returned by the judge.

Binarized Chamfer scores. Nearest-neighbor matching always assigns a counterpart, which can inflate set-level averages if many matches are weak. To provide a stricter robustness check, we additionally report a binarized variant CD_x^b obtained by thresholding each raw 0–5 rating at 3 (1 if ≥ 3, else 0) and then averaging in the same two-direction Chamfer form:

$$CD_x^b(P, G) \;=\; \tfrac{1}{2}\Big(\tfrac{1}{|P|}\sum_{p\in P} b_x\big(p, \mathrm{NN}_G(p)\big) \;+\; \tfrac{1}{|G|}\sum_{g\in G} b_x\big(\mathrm{NN}_P(g), g\big)\Big),$$

where $b_x(\cdot, \cdot) \in \{0, 1\}$ is the thresholded score.

Results. Table 6.15 confirms the same qualitative story as the mean-pairing metrics in Table 6.13, but now at the set level. Disinformation-prone outlets achieve consistently high Chamfer-style scores in both topical and stance dimensions (e.g., $\mathrm{CD}_{TA} \approx 0.72$–$0.84$ and $\mathrm{CD}_{SA} \approx 0.74$–$0.91$), indicating that DiNO's extracted narrative sets both (i) contain narratives that strongly match expert references and (ii) provide broad coverage of the reference narrative space. In contrast, The Guardian scores substantially lower, especially on stance (e.g., $\mathrm{CD}_{SA} \approx 0.19$–$0.30$, with binarized CD_b^{SA} near zero), showing that while a small number of extracted narratives may touch related topics, they rarely adopt the disinformation stance and cover only a limited portion of the disinformation narrative set. Finally, the binarized variants (3/5 threshold) reinforce that the outlet-level conclusions are not an artifact of "soft" similarities: disinformation-prone outlets retain high pass rates

Dataset	CD_{TA}	CD^{b}_{TA}	CD_{SA}	CD^{b}_{SA}
Ukraine				
Sputnik	0.72	0.78	0.76	0.87
Pravda	0.74	0.80	0.74	0.86
Guardian	0.45	0.39	0.30	0.09
COVID-19				
NatNews	0.84	0.99	0.91	0.97
Sputnik	0.73	0.91	0.77	0.93
Guardian	0.47	0.34	0.19	0.00
Migration				
Breitbart	0.74	0.92	0.81	0.92
Guardian	0.37	0.21	0.22	0.03

Table 6.14: Chamfer-style set-level alignment scores between DiNO-extracted narratives and ground-truth narratives across all datasets. Higher values indicate better set-level alignment and coverage. We additionally report binarized scores (CD^{b}) using a 3/5 threshold.

under thresholding, whereas The Guardian's stance matches largely disappear, consistent with neutral or debunking reporting rather than narrative propagation.

6.4.8 Cross-Dataset Ground-Truth Validation

We next test whether our Ukraine-domain conclusions depend on the particular ground-truth narrative set used for evaluation.

Evaluation. We keep the DiNO-extracted narratives from the Ukraine-war articles (see Section 6.4.6) fixed and recompute the same mean alignment scores (mTA_A, mSA_A) and set-level Chamfer scores (CD_{TA}, CD_{SA}) against two alternative expert ground-truth narrative sets that explicitly cover the Ukraine–Russia war: SemEval [152] and an Atlantic Council report on Kremlin narratives about Ukraine [14]. Throughout, we use the same LLM judge (Sonnet 4.5) and the same scoring prompts (Appendix C.1.3).

	mTA_A	mSA_A	CD_{TA}	CD_{SA}
EUDisinfoTest				
Sputnik	0.79	0.83	0.72	0.76
Pravda	0.78	0.84	0.74	0.74
Guardian	0.80	0.20	0.45	0.30
SemEval				
Sputnik	0.71	0.86	0.72	0.83
Pravda	0.76	0.86	0.76	0.81
Guardian	0.80	0.30	0.51	0.38
Atlantic Council				
Sputnik	0.71	0.83	0.72	0.79
Pravda	0.75	0.85	0.71	0.75
Guardian	0.70	0.30	0.52	0.36

Table 6.15: Scores for DiNO-extracted narratives from the Ukraine-war articles, evaluated against three ground-truth narrative sets (EUDisinfoTest, SemEval, Atlantic Council). We report mean alignment (mTA_A, mSA_A) and set-level Chamfer scores (CD_{TA}, CD_{SA}).

Results. Table 6.15 shows that the outlet-level pattern is stable across all three ground-truth sources. Pravda and Sputnik consistently achieve high topical and stance agreement, and this is reflected not only in high mean scores but also in strong set-level values CD_{TA} and CD_{SA}, indicating both close matching to expert narratives and broad coverage of the expert narrative space. In contrast, The Guardian attains relatively high mTAA, which is expected since mTAA is a precision-like score: a small set of extracted narratives can score well if those narratives are on-topic. However, CD_{TA} remains lower, consistent with limited topical coverage, and stance alignment is markedly weaker, with low mSAA and especially low CD_{SA}. Overall, the separation between disinformation-prone and mainstream outlets persists under alternative ground-truth definitions.

6.4.9 Execution Time

Beyond structure, it is also important to understand DiNO's runtime characteristics. We measured wall-clock time (in minutes) for DiNO's four pipeline

Outlet	Step 1 Articles	Step 2 Segments	Step 3 Clusters	Step 4 Narr.
Ukraine War				
Sputnik	20 001	8 892	66	66
Pravda	22 695	7 312	44	44
Guardian	11 841	559	2	2
COVID-19				
Sputnik	13 107	1482	9	9
NatNews	16 789	8 941	60	60
Guardian	10 573	520	2	2
Migration				
Breitbart	16 203	8 177	65	65
Guardian	11 671	698	3	3

Table 6.16: Step-wise statistics of the DiNO pipeline for each news outlet, across both Ukraine War, COVID-19 and Migration domains. Step 1: total articles processed; Step 2: number of identified segments in processed articles; Step 3: number of clusters of semantically similar false segments; Step 4: number of extracted disinformation narratives.

steps on our reference hardware (2× Intel® Xeon® Gold 5418Y, 128 GB RAM, 4× NVIDIA L40). Table 6.18 reports per-step runtimes across all datasets.

Dataset	Step 1	Step 2	Step 3	Step 4
Ukraine War				
Sputnik	207 min	13 min	14 min	<1 min
Pravda	182 min	14 min	10 min	<1 min
Guardian	42 min	10 min	3 min	<1 min
COVID-19				
Sputnik	104 min	10 min	7 min	<1 min
NaturalNews	133 min	12 min	8 min	<1 min
Guardian	34 min	10 min	3 min	<1 min
Migration				
Breitbart	129 min	12 min	10 min	<1 min
Guardian	39 min	10 min	3 min	<1 min

Table 6.18: Wall-clock runtimes (in minutes) of the four DiNO pipeline steps: segment matching, entailment verification, clustering, and narrative extraction.

DiNO Component	Optimal Setting / Model
Matching False Segments	
Similarity Measure	Cosine (Jina)
Segmentation	DiNO-sgm
Hyperparameters	$l = 4$, $o = 1$, $t = 3$
Entailment Verification	
Model / Prompt	GPT-5-mini / (Fig. C.1)
Clustering False Segments	
Embedding Model	SFR-Embedding-2
Dimensionality Reduction	UMAP
Clustering Algorithm	HDBSCAN
Extracting Narratives	
Model / Prompt	GPT-5.2 / (Fig. C.7)

Table 6.17: Optimal models and settings for each step of the DiNO pipeline, covering segment matching, entailment verification, clustering, and narrative extraction.

End-to-End Comparison. Summing the four steps yields the end-to-end runtimes in minutes (Table 6.19), where we compare DiNO to CaNarEx and Relatio. This highlights that, while DiNO introduces additional structure (e.g., entailment verification), its overall runtime remains competitive with existing narrative extraction methods.

Dataset	DiNO	CaNarEx	Relatio
Ukraine War			
Sputnik	234 min	810 min	125 min
Pravda	206 min	762 min	110 min
Guardian	55 min	194 min	41 min
COVID-19			
Sputnik	121 min	531 min	82 min
NaturalNews	153 min	569 min	82 min
Guardian	47 min	173 min	37 min
Migration			
Breitbart	151 min	569 min	97 min
Guardian	52 min	195 min	33 min

Table 6.19: Total end-to-end processing times (in minutes) comparing DiNO against CaNarEx and Relatio.

6.4.10 Execution Cost

Finally, we estimate the monetary cost of running DiNO with external LLM APIs; infrastructure costs on our in-house cluster are not included. Cloud-based costs can be obtained by applying current per-token rates to the token counts reported below.

Token-count Assumptions. We follow OpenAI's convention of 1 token $\approx$ 0.75 words. For each LLM-based step, we compute total word counts and derive input/output token estimates accordingly.

LLM Usage Across the Pipeline. Within DiNO, only the entailment verification and narrative extraction steps invoke external LLMs. Segment matching and clustering rely on local models and therefore incur no API costs. Table 6.20 reports end-to-end token usage per dataset by aggregating input and output tokens from both LLM-based steps. Output tokens from narrative generation are negligible compared to those from entailment verification and are thus omitted from the totals for simplicity.

To estimate monetary cost, one can multiply the total input and output token counts from Table 6.20 by the per-token prices of the chosen LLM. In all cases, narrative extraction contributes only a negligible fraction of the total token usage

Dataset	Total Input Tokens	Total Output Tokens
Ukraine War		
Sputnik	40.09 M	1.01 M
Pravda	17.25 M	1.11 M
Guardian	35.05 M	0.61 M
COVID-19		
Sputnik	19.71 M	0.71 M
NaturalNews	47.54 M	0.91 M
Guardian	28.94 M	0.51 M
Migration		
Breitbart	25.11 M	0.71 M
Guardian	28.75 M	0.51 M

Table 6.20: End-to-end token usage for DiNO's LLM-based steps (entailment verification and narrative extraction) across datasets (values in millions of tokens). Input tokens sum both steps; output tokens are dominated by entailment verification, as narrative outputs are < 0.01M per dataset.

compared to entailment verification.

6.5 Results and Discussion

In this chapter, we begin by examining the five most prevalent narratives identified by DiNO in each dataset (Figure 6.4). We then analyze the narratives' distributions and temporal dynamics, relating major peaks to salient socio-political events. Finally, we conduct cross-source comparisons to quantify the similarity of narrative sets across outlets.

Most prevalent narratives by domain. In the **Ukraine War** domain, DiNO extracts 68 narratives from Sputnik, 44 from Pravda, and 2 from The Guardian. Pravda's most frequent narratives portray Russia as restrained and peace-seeking while NATO and the U.S. are framed as escalating the war; they also emphasize Ukrainian leadership as corrupt and the war as a futile "meat-grinder" enabled by Western aid. Sputnik's most frequent narratives focus on sanctions and geopolitical

Pravda-Ukraine
- Russia responsibly warns the US while NATO escalates Ukraine war, risking nuclear catastrophe.
- Western media suppress Putin's truths; U.S.-backed Ukraine war continues despite Russia seeking fair peace.
- Trump will cut Ukraine aid, force peace talks, and expose Biden's "deep state" crimes.
- Ukraine, driven by NATO, sacrifices civilians and youth in a doomed, corrupt meat-grinder war.
- U.S. Ukraine aid is corrupt, ideological, and self-serving, while Americans and Ukraine suffer.

Sputnik-Ukraine
- Western Russia sanctions caused self-inflicted inflation, energy crises, and collapsing living standards.
- West stages peace summits to impose Zelensky's ultimatum and prolong war, excluding Russia.
- Western hegemony and sanctions fail; Russia defends sovereignty and leads multipolar order.
- Biden officials hid incompetence while fueling endless proxy wars, enriching contractors.
- US-linked Ukraine biolabs allegedly conducted illicit experiments; critics smear dissenters as traitors.

NaturalNews-COVID
- Fauci, CIA, and NIH allegedly covered up lab-leak origins while profiting from gain-of-function research.
- Big Tech and government collude to censor truth, pushing people toward independent free-speech platforms.
- A corrupt deep state installed Biden via fraud, using Covid to seize control.
- Government and pharma pushed unsafe COVID shots via propaganda, corruption, and coercion.
- COVID mRNA vaccines allegedly harm fertility, pregnancy, and infants, while authorities conceal risks.

Sputnik-COVID
- Russia's vaccines are safe, innovative, and unfairly maligned by hostile geopolitical disinformation campaigns.
- COVID-19 exposed governance failures, social strains, and resilience as India struggled to cope.
- UK COVID rules were inconsistent, overreaching, and harmful, undermining freedoms, trust, and livelihoods.
- China and WHO hid Wuhan lab origins; US-funded research and officials covered up evidence.
- Citizens worldwide protested COVID restrictions, vaccine mandates, and crises, demanding freedom and restored rights.

Breitbart-Migration
- Democrats and elites push amnesty and open borders for cheap labor and lasting political power.
- Biden and sanctuary policies enable illegal migration, drug trafficking, and addiction, endangering Americans.
- European leaders tighten laws as migration fuels crime, extremism, and widespread sexual violence.
- Biden's illegal-parole "parallel system" floods migrants, hurting Americans' wages, safety, sovereignty.
- America must urgently secure the border and punish sanctuary policies to prevent deaths and crime.

Guardian-Ukraine
- Ukraine-Russia war drives defensive fortifications, western debates, sanctions fallout.
- Russia is not seeking peace talks and will only negotiate if the West accepts its annexations.

Guardian-COVID
- Covid responses demand better data, clearer leadership, and fair vaccine access to protect communities.
- Pandemic and Brexit compounded shocks are wrecking UK high streets, jobs, and livelihoods.

Guardian-Migration
- Immigration's wage effects are usually small, but can be negative for the closest competing workers.
- Australia must end offshore detention, reunite refugees, and value migrants' humanity and contributions.
- U.S. border policies sacrifice communities, nature and families, demanding humane, lawful solutions.

Figure 6.4: Top five (where available) most frequent narratives extracted by DiNO per dataset (truncated).

contestation, repeatedly framing Western sanctions as self-damaging and ineffective, and depicting Western-led peace initiatives as performative or exclusionary. The Guardian yields only two narratives, centered on fortifications, sanctions fallout, and the claim that Russia is not pursuing peace talks except on annexation terms.

For **COVID-19**, NaturalNews yields 60 narratives dominated by elite conspiracy and institutional malfeasance claims (e.g., cover-ups involving U.S. agencies, censorship/collusion by government and tech), paired with repeated allegations of vaccine harm and coercive rollout. Sputnik yields 9 narratives combining positive framing of Russia's vaccines with broader critiques of Western pandemic governance, restrictions, and lab-origin cover-up narratives, alongside protest/freedom frames. The Guardian again yields only two narratives, emphasizing governance and data shortcomings and the socio-economic fallout of pandemic-era shocks.

In the **Migration** domain, Breitbart yields 65 narratives framing migration as an elite-driven project (amnesty/open borders) tied to crime, drugs, sovereignty, and wage pressures, and urging tougher enforcement and sanctions on sanctuary policies. The Guardian yields three narratives spanning (i) the claim that immigration's wage effects are typically small but can be negative for the closest competing workers, and (ii) humanitarian frames calling for ending offshore detention, reuniting refugees, and pursuing more humane, lawful border and asylum policies.

Robustness. Across all evaluated domains and outlets, DiNO exhibits stable behavior. In particular, the method consistently separates disinformation-prone outlets from the credible outlets: disinformation-prone corpora yield many narratives with strong alignment to ground truth disinformation narratives, whereas the reputable outlet yields only a small number of narratives with weak stance alignment. This pattern persists under stricter, thresholded (binarized) scoring. DiNO also returns compact narrative sets with low redundancy: the identified narratives are neither dominated by near-duplicates nor excessively fragmented into many highly overlapping variants. Finally, the high agreement between automated metrics and expert assessment supports the robustness of these conclusions when the analysis is repeated across datasets and domains. Overall, these results indicate that DiNO remains stable across domains and outlets.

Generalization. To probe generalization beyond the primary reference inventory, DiNO was evaluated against three independent ground-truth disinformation narrative datasets. The patterns remain consistent: disinformation-prone corpora yield richer and more aligned narrative sets, whereas credible outlets produce few narratives with weaker stance alignment and low coverage of the reference ground-truth disinformation narratives. The preservation of these trends under a different ground-truth disinformation narrative sets reflect stable properties of the

mined narratives.

Coverage. DiNO provides broad coverage of the ground-truth disinformation narratives for disinformation-prone outlets: across domains, it identifies many narratives and achieves high set-level (Chamfer-style) alignment in both topical and stance dimensions. This indicates that the mined narrative sets capture a substantial fraction of the expert-defined narratives rather than only a small subset. In contrast, coverage is consistently low for credible outlets, reflected in low set-level scores and few matched narratives. This outcome is expected: credible reporting may discuss related topics, but it rarely adopts or propagates the disinformation narratives represented in the reference sets.

6.5.1 Narratives Over Time and Relation to Events

The figures below plot the temporal distribution of the ten (where applicable) most frequent disinformation narratives in each dataset highlighting how their peaks correspond to contemporary events or debates.

Sputnik-Ukraine. Figure 6.5 shows the narratives found by DiNO in the Sputnik-Ukraine outlet. Right after Russia's full-scale invasion, the main narrative focuses on sanctions, claiming that Western sanctions "backfire" and mainly hurt Europe by causing an energy crisis and rising living costs. The timing matches the rollout of EU sanctions and the major energy-price shock and gas-supply disruption that followed the invasion [25]. At the same time, economic and energy analyses link the 2022 inflation/energy crisis largely to war-driven supply disruption and reduced Russian gas flows, rather than to a simple "sanctions hurt the West more" story [25].

DiNO also captures a narrative that the West runs "peace summits" to prolong the war. This coincident with real events like the Swiss-hosted peace summit in June 2024, which did not include Russia and was publicly dismissed by the Kremlin diplomacy [135]. Another early narrative says Russia intervened to "protect Donbas," echoing likely the justification used in Putin's 24 February 2022 address. Interestingly, in late 2023 one of the narratives blames the U.S. for sabotaging Nord Stream, most likely as a reaction to the September 2022 pipeline explosions.

Overall, the narratives portray Russia as defensive and resilient, while casting Ukraine and the West as deceptive, corrupt, or escalation-driven.

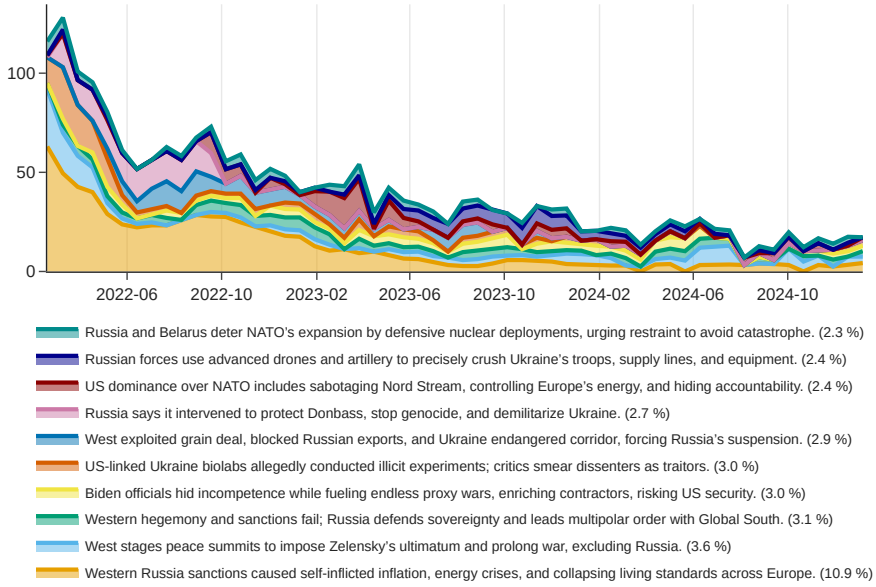

Figure 6.5: Temporal distribution of the ten main disinformation narratives in the Sputnik-Ukraine dataset. The y-axis represents the absolute number of articles containing each narrative within a given time period.

Pravda-Ukraine. Figure 6.6 shows two clear waves in Pravda's output: a first surge in late 2023 and a stronger resurgence in late 2024. In October-November 2023, Pravda amplifies the narrative that the West exploits the Middle East crisis to divert attention and abandon Ukraine, aligning with the outbreak of the Israel-Hamas war and the ensuing U.S. debate over a combined Ukraine/Israel supplemental request [162].

In early 2024, Pravda's emphasis shifts toward delegitimisation and exhaustion frames: the Zelensky-as-puppet/illegitimate narrative reflects likely Russia's public effort to question Zelensky's mandate as elections were postponed under martial law [163].

Finally, the late-2024 spike is dominated by escalation rhetoric: NATO escalates, risking nuclear catastrophe. This rhetoric intensifies around the NATO Washington summit and U.S. decisions to relax restrictions on Ukrainian strikes into Russia [141].

Overall, Pravda attributes escalation and economic harm to Western actors,

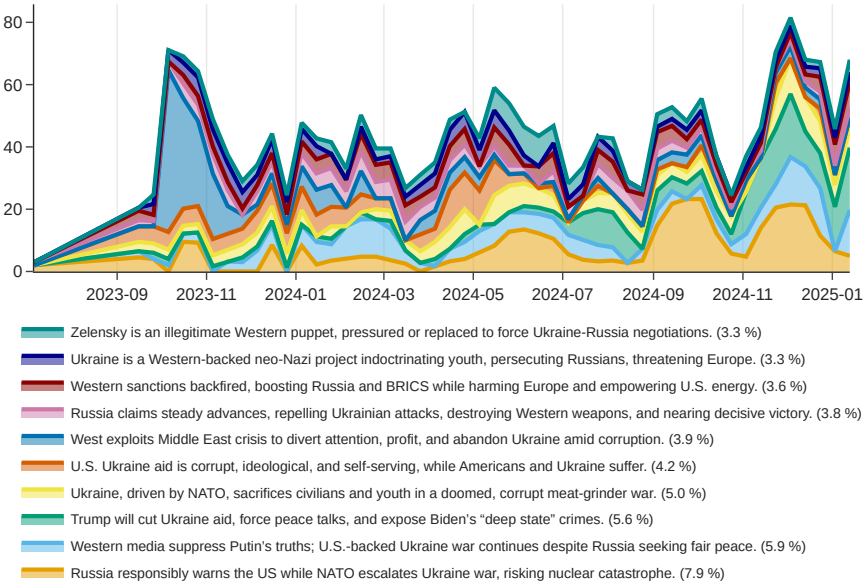

Figure 6.6: Temporal distribution of the ten main disinformation narratives in the Pravda-Ukraine dataset. The y-axis represents the absolute number of articles containing each narrative within a given time period.

denies Ukrainian agency by casting Ukraine as a controlled proxy, and frames Russia as restrained and defensive.

NaturalNews-COVID. Figure 6.7 shows that NaturalNews' COVID coverage is dominated by a conspiracy triad: (i) bioweapon/lab-leak cover-up claims (e.g., blaming Fauci/NIH/CIA and alleging gain-of-function profiteering), (ii) control-through-public-health framings (mandates, surveillance, and censorship), and (iii) "suppressed cure" messaging (ivermectin and "natural remedies" framed as deliberately hidden). The lab-leak/bioweapon narrative plausibly intensifies when U.S. intelligence assessments explicitly judged a bioweapon origin unlikely [137].

A second strong cluster attacks vaccination as coercive and biologically dangerous (e.g., "DNA-altering," fertility harm, heart damage), aligning with the mass vaccine rollout after the first U.S. mRNA vaccine EUA [69]. Finally, "suppressed cure" narratives coincident with regulator statements that ivermectin is not recommended for COVID-19 outside trials [9].

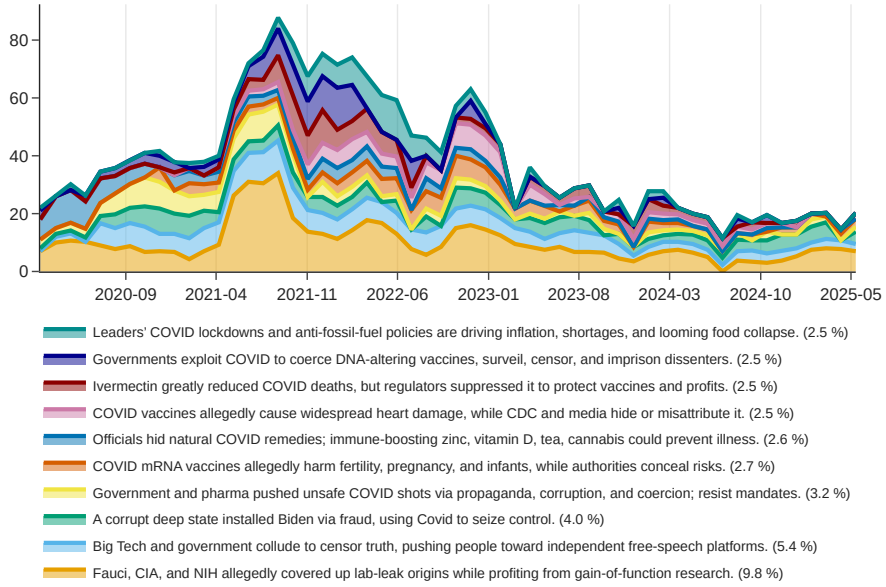

Figure 6.7: Temporal distribution of the ten main disinformation narratives in the NaturalNews-COVID dataset. The y-axis represents the absolute number of articles containing each narrative within a given time period.

In general, the leading themes center on bioweapon framings, censorship/control claims, and "suppressed cure" messaging.

Sputnik-COVID. Figure 6.8 shows three dominant narratives. First, a strong and repeated "Sputnik V success" narrative (Russia as a vaccine leader, the West as unfair or political), which fits key legitimacy moments: Putin's public approval announcement [160], publication of Sputnik V trial results in The Lancet [109], and the EMA starting a rolling review [10]. Second, a recurring lab-origin cover-up narrative (focus on Wuhan/WHO/China and "hidden truth"), which aligns with renewed international debate during the WHO-China origins report [193] and the U.S. intelligence review ordered by Biden [161]. Third, vaccine fear messaging highlighting Western vaccine risks, which coincident with real high-salience safety headlines such as the EMA's update on rare clotting events for AstraZeneca's vaccine [11].

Overall, Sputnik-COVID mixes big conspiracy frames with consistent promotion

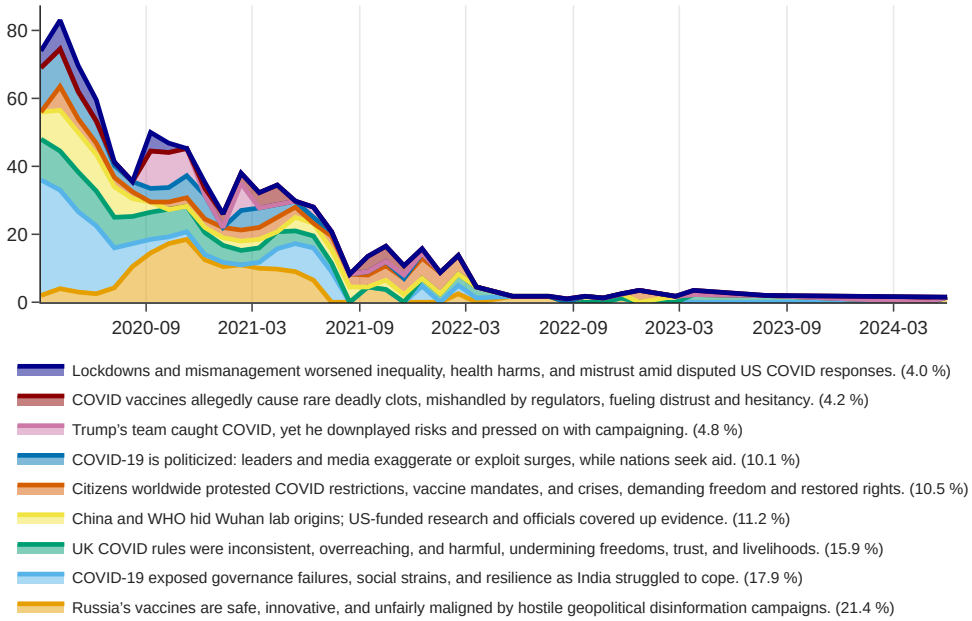

Figure 6.8: Temporal distribution of the nine main disinformation narratives in the Sputnik-COVID dataset. The y-axis represents the absolute number of articles containing each narrative within a given time period.

of Russian biomedical achievements.

Breitbart-Migration. Figure 6.9 shows dominant narratives found in Breitbart-Migration outlet. First, a persistent "border invasion/chaos" narrative (loss of control, mass crossings, broken enforcement), which fits high-salience border moments: the 2018 Central American caravan becoming a central U.S. midterm issue [159]. Second, a recurring "Democrats and elites want open borders" narrative, which aligns with election-cycle mobilization around immigration and renewed focus on claims about noncitizen voting. However, noncitizen voting is already illegal in federal elections and documented instances are described as vanishingly rare [199]. Third, a "migration drives crime and fentanyl" narrative. However, the vast majority of interdicted fentanyl is seized at ports of entry, complicating simple "unauthorized border crossers bring the fentanyl" framings [52].

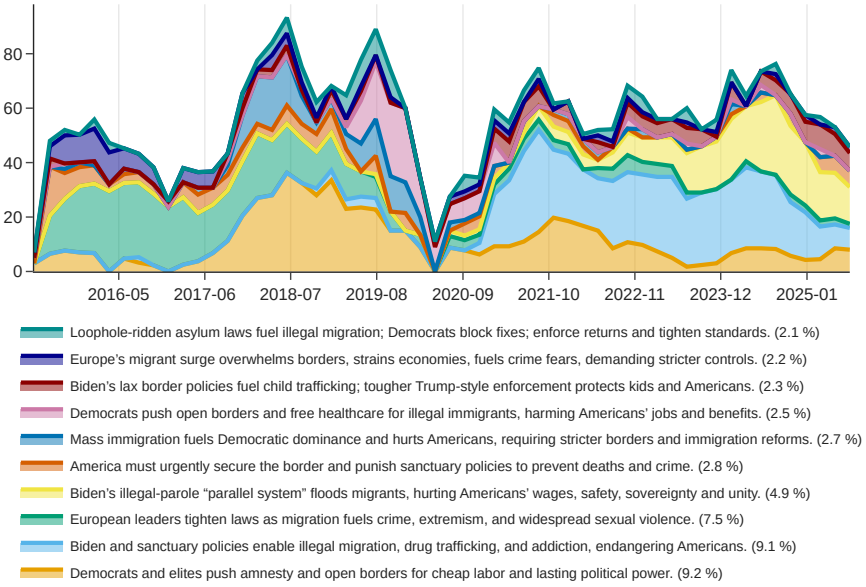

Figure 6.9: Temporal distribution of the ten main disinformation narratives in the Breitbart-Migration dataset. The y-axis represents the absolute number of articles containing each narrative within a given time period.

Overall, DiNO's Breitbart-Migration narratives appear oriented to U.S. domestic audiences: they amplify threat and disorder frames and attribute intent and blame to partisan opponents.

Chapter 7

Conclusions

This chapter summarises the main findings of the thesis, revisits the research objectives introduced in Chapter 1, and reflects on their implications for disinformation research. The thesis has approached disinformation not as isolated falsehoods, but as recurring disinformation narratives that can be detected, mapped and monitored across benchmarks, fact-checking corpora and news outlets. Throughout, LLMs have played a dual role: both as objects of evaluation and as tools for both extracting and abstracting narrative-level structures.

7.1 Summary of Main Findings

This thesis investigated how disinformation narratives can be conceptualised, detected and mined. Across EUDisinfoTest, DiNaM and DiNO, it examined narrative-level robustness of LLMs, introduced two narrative mining pipelines, and applied them to fact-checking articles and outlet-specific news corpora. Together, these components provide a vertically integrated perspective: from conceptual foundations, through model evaluation, to global and outlet-level mapping of disinformation narratives.

7.1.1 Revisiting the Research Objectives

The work pursued four objectives, introduced in Chapter 1. The main findings are summarised below with respect to each objective, and their joint contribution to a narrative-centric approach to disinformation research is outlined.

7.1.2 Findings for Objective O1: Conceptual foundations for disinformation narrative classification and mining.

Objective O1 was addressed primarily in Chapter 3, which developed a conceptual and task-oriented foundation for disinformation narrative research.

On this basis, the thesis distinguished two core tasks: disinformation narrative classification, which asks whether a given text expresses a known disinformation or credible narrative, and disinformation narrative mining, which aims to discover recurring narrative patterns from collections of texts. Chapter 3 further introduced a three-stage narrative-mining framework that underpins both DiNaM and DiNO, providing a common structure that links practitioner-oriented notions of disinformation narratives to computational representations and algorithms.

7.1.3 Findings for O2: Benchmarking LLMs on disinformation narrative classification.

To realize Objective O2, the thesis investigates how LLM performance varies across topical domains and across rhetorical variants (Logos, Ethos, Pathos) (Chapter 4).

Evaluations on EUDisinfoTest revealed a clear stratification of model performance. LLMs such as GPT-5-class models achieved the strongest aggregate results, maintaining a comparatively balanced trade-off between detecting disinformation narratives and correctly recognising credible ones. Mid-size and earlier-generation models, including GPT-3.5 and open 8–9B models, lagged behind, with lower overall scores and more pronounced imbalances in disinformation narrative classification. Even the best models still misclassified a non trivial fraction of both disinformation and credible narratives, indicating that narrative-level classification remains challenging.

A human–model comparison on base narratives showed that LLMs can surpass human annotators in F1, largely because humans are more conservative in assigning the "disinformation" label. Human annotators exhibited higher true-negative rates, reflecting a reluctance to flag borderline cases as disinformation, whereas LLMs more readily commit to a label.

The analysis of rhetorical variants showed that persuasive style has systematic effects on model decisions. Pathos variants, which intensify emotional tone, tended to reduce the rate at which models correctly identified disinformation narratives.

Logos variants moderately degraded performance on both disinformation and credible narratives, suggesting that more reasoned or evidence-laden wording can obscure underlying falsity for models. Ethos variants mainly harmed the recognition of credible narratives, making models more likely to misclassify statements as disinformation when they invoke authority trust. Topic-wise results further indicated that models perform comparatively well on stereotyped domains such as migration, climate and public health, but are weaker on narratives about media and institutional distrust. Overall, EUDisinfoTest highlighted both the promise of LLMs for narrative classification and their blind spots with respect to rhetoric and topic, motivating narrative-level, rhetoric-aware evaluation.

7.1.4 Findings for O3: Constructing a global map of disinformation narratives from fact-checking articles.

Objective O3 was addressed in Chapter 5, which instantiated the three stage narrative mining framework in the DiNaM pipeline. DiNaM operates on large collections of fact checking articles and reconstructs a global catalogue of disinformation narratives implied by expert verification practice.

DiNaM instantiates the framework in four main steps. First, it filters fact-checking articles to retain only those that review exclusively false content. Second, it applies an LLM-based extraction, verification and refinement procedure that turns spans of false information into structured false-information statements. Third, these statements are embedded and clustered into candidate narrative groups. Fourth, an LLM derives a disinformation narrative for each cluster.

Empirical evaluations showed that this configuration is effective. Zero shot LLMs outperformed fine tuned encoder baselines in filtering, and the structured extraction pipeline surpassed simpler prompting strategies and non LLM question answering models. Clustering quality, measured by Silhouette scores, confirmed that the chosen embedding and clustering setup yields coherent clusters, and LLM derived narratives aligned more closely with EUDisinfoTest and outperformed general purpose narrative mining methods such as CaNarEx and Relatio. The resulting global narrative map recovered known disinformation narratives on COVID 19, the Ukraine Russia war, climate change, migration, the European Union and the Israel Palestine conflict, and its temporal dynamics followed major socio political developments. This demonstrates that DiNaM can construct a global, empirically anchored map of disinformation narratives that complements the statement level

perspective of EUDisinfoTest.

7.1.5 Findings for O4: Analysing outlet-level realisations of disinformation narratives.

Objective O4 extended the mining framework to news outlets in Chapter 6, introducing DiNO. Whereas DiNaM operates on fact-checking articles, DiNO mines disinformation narratives directly from outlet-specific news corpora, anchored in databases of verified false claims.

DiNO segments news articles, matches candidate segments to verified false claims using embedding-based similarity, verifies whether segments actually support the false claims via LLM-based entailment checks, and clusters supporting segments to synthesise narratives and obtain outlet-specific narrative repertoires. Module-level evaluations showed that each component contributes meaningfully: LLM-based entailment verification outperformed fine-tuned encoder baselines, and the embedding-plus-clustering configuration again yielded coherent clusters that support high-quality narrative extraction.

Narrative quality was assessed via LLM-as-a-judge metrics for topical and stance alignment with ground truth disinformation narratives, as well as via human evaluation. DiNO consistently produced narratives with strong topical and stance alignment in disinformation-prone outlets and low stance alignment and coverage in reputable outlets, and it outperformed general purpose narrative mining methods such as CaNarEx and Relatio. Applied to outlet corpora on Ukraine, COVID-19 and Migration, DiNO uncovered systematic differences between disinformation-prone outlets and a reputable outlet: disinformation-prone outlets converged on narratives portraying Western aggression, Russian victimhood, vaccine danger and migrant threat, while the reputable outlet covered many of the same topics but with opposite framings focused on Russian aggression, public health and human rights. This demonstrates that DiNO provides a local, outlet-sensitive view of disinformation narratives that complements DiNaM's global map and the benchmark role of EUDisinfoTest.

7.2 Implications

The thesis has implications for several research and practitioner communities that intersect around disinformation, Large Language Models and democratic

information environments. This section discusses the implications of the findings for NLP and LLM research, as well as for Media Science and Political Science.

7.2.1 NLP and LLM Research

For NLP and LLM research, the findings from EUDisinfoTest highlight the importance of narrative-level, rhetoric-aware benchmarks. In contrast to many disinformation datasets that focus on item-level fact verification or social media posts, narrative-centered evaluation directly targets models' ability to recognize and distinguish recurring messages.

The empirical analysis of rhetorical strategies and topic variation provides concrete lessons about current LLM robustness. The fact that Pathos framing reduces true positive rates, Logos degrades both true negative rates and true positive rates, and Ethos variants markedly lower true negative rates suggests that models are systematically affected by persuasive presentation. Topic-wise results further show that LLMs are comparatively strong in stereotyped domains, but less reliable when narratives concern institutional trust, media legitimacy and geopolitical alignment. Together, these findings point to the need for targeted robustness interventions that explicitly account for rhetorical and topical shifts.

DiNaM and DiNO illustrate the value of narrative-level mining frameworks. At the global level, DiNaM shows that LLMs can support the construction of large scale maps of disinformation narratives from fact checking corpora, while DiNO demonstrates how those ideas can be applied to outlet specific news coverage anchored in verified false claims. From an NLP perspective, both pipelines exemplify how LLMs can be embedded in multi stage systems that combine prompt based extraction, embedding based similarity search and unsupervised clustering. They also show how practices such as using LLMs as judges for topical and stance alignment, and embedding based metrics can be applied in a controlled way to assess complex narrative mining systems. These design patterns are not limited to disinformation, and could be adapted to other tasks where models need to recover, track and compare evolving narrative structures.

At the level of information extraction, the thesis shows that LLMs become more reliable extractors of false information when they are embedded in structured workflows. In DiNaM, the Extraction–Verification–Refinement procedure for fact-checking articles treats extraction as a contextual question-answering task: it first asks the model to list statements expressing the false claims, then verifies for each

candidate that it is a clean, misleading formulation of the false claim (i.e., free of debunking or corrective language), and finally retries cases that fail verification with targeted feedback that encourages the model to remove debunking language and reduce hallucinations. The comparison with prompting strategies such as Chain-of-Thought and Chain-of-Verification shows that generic multi-step prompting is not enough: models tend to mix false claims with corrective context unless they are forced to verify and, when needed, re-extract under targeted feedback. This points to a broader design principle for LLM-based extraction: pipelines should include explicit stages for semantic verification and refinement targeted at known failure modes, rather than relying on open-ended "explain your reasoning" prompts.

DiNO's dedicated segmentation method (DiNO-sgm) addresses a crucial design choice: how to split articles into text units that are likely to contain complete false claims. DiNO-sgm divides each article into overlapping windows of sentences and then enumerates all contiguous segments within each window, from single sentences up to multi-sentence spans. This flexible segmentation better matches the variable length of false information, which often depends on cross-sentence context but rarely spans an entire article. The evaluation shows that DiNO-sgm recovers the main false claim expressed in an article more reliably than fixed-granularity alternatives, including sentence-level, passage-level, and full-article segmentation. This suggests that careful, variable-length segmentation is a general and effective design pattern for systems that need to align free-form text with known claims or labels.

7.2.2 Media Science, and Political Science

For Media Science, and Political Science fields, DiNaM and DiNO demonstrate how narrative mining can support monitoring workflows. DiNaM provides a global narrative map derived from fact-checking articles that shows which disinformation narratives recur across outlets, platforms and countries, and how they rise and fall over time around salient events such as the COVID 19 pandemic or the escalation of the Ukraine war. Such maps could be used to identify emerging narratives earlier, to track when local rumours start to generalise into broader patterns, and to support strategic decisions about which topics or narratives should be prioritised for communication.

DiNO complements this view by projecting verified false claim resources into the coverage of specific outlets. The resulting outlet level narrative profiles can help practitioners understand how disinformation prone outlets adapt disinformation

narratives.

7.3 Limitations

The contributions of this thesis should be interpreted in light of several limitations. These concern the scope and nature of the data, methodological and modelling choices, and the conceptual and interpretive constraints of the narrative focus.

7.3.1 Data and Scope Limitations

The empirical analyses in the thesis focus primarily on European information environments and on a set of salient domains, including COVID 19, the Ukraine war, migration and selected topics related to climate and European integration. While these domains are central for recent disinformation campaigns, the findings cannot be assumed to generalise directly to other regions, policy areas or cultural contexts without further study. Similarly, the outlet level analyses are restricted to a limited number of outlets, such as Sputnik, Pravda, NaturalNews, Breitbart and The Guardian, which represent specific segments of the broader media ecosystem.

Both DiNaM and DiNO depend on existing fact-checking and false-claim resources, including EDMO-affiliated fact-checking organisations and the EUvsDisinfo and GFCE databases. Where no fact-checks or curated false-claim entries exist, the pipelines cannot identify or model disinformation narratives, so the resulting maps are necessarily incomplete and reflect the priorities, capacities and language coverage of these organisations. More broadly, fact-check corpora should be understood as a high-precision proxy for the disinformation environment rather than a representative sample: large-scale global analyses explicitly use fact-checks as a proxy for disinformation prevalence, but this evidence necessarily tracks what is actually selected for checking and publication. Which claims are verified is shaped by organisational routines and verification norms, including systematic selection criteria such as checkability, prominence and timeliness, and the composition of fact-checked content differs across national information environments in ways that mirror local agendas.

The analysis is further bounded by its focus on English-language content, which likely underrepresents idiomatic or locally salient narratives in other languages. Finally, the set of outlets and time windows is constrained by data availability and

scraping conditions, and might not cover the full media ecosystem, particularly social media where disinformation also circulates.

7.3.2 Methodological and Modelling Limitations

Several steps in our narrative-mining workflows use LLMs, which can introduce model-dependent biases and sensitivity to design choices (e.g., prompts and sampling settings). We evaluated both open-weight and closed-weight models and, where feasible, non-LLM alternatives. Still, the best performance came from closed-weight systems, which reduces transparency and limits full reproducibility.

The analysis focuses exclusively on textual content and does not account for multimodal elements such as images, videos or memes, which play a central role in many disinformation campaigns. Incorporating these modalities would require additional data, annotation efforts and modelling infrastructure, and remains an important direction for future work.

7.3.3 Conceptual and Interpretive Limitations

The narrative maps produced by DiNaM and DiNO are shaped by the norms, priorities and biases of the fact-checking and monitoring organisations that provide the underlying ground truth. These organisations operate within specific institutional and political contexts, which influence which narratives are selected for scrutiny, how they are described and how labels are assigned. In this thesis, their outputs are treated as a pragmatic approximation of disinformation practice, but this dependence should be borne in mind when interpreting the results.

7.4 Future Work

The work in this thesis opens several directions for further research and development. These directions follow the three main pillars of the thesis, as well as the longer term goal of integrating narrative aware methods into human facing monitoring systems and cross disciplinary research on democratic information environments.

7.4.1 Extensions of EUDisinfoTest and Narrative-Level Evaluation

A first line of extension concerns EUDisinfoTest and, more broadly, the evaluation of LLMs at narrative-level. The current benchmark focuses on topics that are central to European disinformation debates and on narratives. Future work could broaden this scope by adding further domains, such as domestic politics, energy policy and security, and by extending the benchmark to additional languages. Another natural direction is to incorporate multimodal narratives, where text is combined with images or other media.

On the modelling side, this thesis uses EUDisinfoTest as a zero-shot evaluation setting. A key next step is to check whether models can be fine-tuned or instruction-tuned so that they stay accurate when the same narrative is written in different ways (i.e., different rhetorical variants).

7.4.2 Advancing Narrative Mining from Fact-Checking Articles

For DiNaM, a central avenue for future work is cross lingual narrative mining directly in the source languages of fact-checking articles and claims. This would require multilingual extraction and clustering models and would make it possible to study how similar narratives are articulated across different linguistic and regional contexts.

From an operational perspective, DiNaM could be developed into an online or streaming system that updates narrative clusters as new fact checks are published. This would require incremental clustering, as well as interfaces for analysts to inspect changes, merge or split clusters and flag emerging narratives. Such a system would bring DiNaM closer to near real time monitoring of narrative dynamics.

7.4.3 Advancing Outlet-Level Narrative Mining

For DiNO, future work can deepen and broaden the analysis of outlet level narratives. One obvious extension is to cover more outlets, regions and platforms, including television transcripts, online forums and social media. Applying DiNO style pipelines to a wider range of media ecosystems would enable more comprehensive comparative analyses of how disinformation narratives spread and are adapted

across different communication channels.

Moreover, extending DiNO to multimodal content is a natural but challenging step. Many disinformation campaigns rely heavily on images, video clips and memes, sometimes with only minimal textual context. Integrating visual and textual signals into a joint narrative mining framework would require advances in multimodal representation learning and careful attention to safety, but it would significantly increase the coverage of the approach.

7.5 Concluding Remarks

This thesis has argued that narrative-level approaches, grounded in fact-checking practice, are essential for understanding contemporary disinformation. By developing the EUDisinfoTest benchmark for LLM-based narratives classification, the DiNaM framework for mining global disinformation narratives from fact-checking articles, and the DiNO method for uncovering outlet-level realisations of these narratives in news media, the work takes a step towards computational tools that operate at the level of narratives rather than isolated claims.

At the same time, the thesis has highlighted both the promise and the limitations of current LLMs in this domain. The models show non trivial ability to distinguish disinformation narratives from credible ones and to support extraction and clustering tasks, yet they remain sensitive to rhetorical variation, topic shifts and the biases of their training data. These findings underline the need for continuous evaluation and for designs that keep humans in the loop when deploying LLM based tools in safety relevant contexts.

The information environment and the capabilities of language models will continue to evolve, and so must our methods for monitoring and governing them. Progress will require sustained collaboration between technical and social science communities, as well as close engagement with practitioners in fact-checking and media monitoring. The central message of this thesis is that combining narrative-level understanding with robust, transparent LLM tools offers a promising path towards better mapping, monitoring and eventually countering disinformation in democratic societies.

References

[1] The pro-kremlin narrative about migrants. EUvsDisinfo, 2017. URL https://euvsdisinfo.eu/the-pro-kremlin-narrative-about-migrants/.

[2] Ukrainians imbued with destructive nazi ideology. EUvsDisinfo, 2021. URL https://euvsdisinfo.eu/report/ukrainians-imbued-with-destructive-nazi-ideology/.

[3] Chapter four: Russia and eurasia. *The Military Balance*, 124(1):158–217, 2024. doi: 10.1080/04597222.2024.2298592. URL https://doi.org/10.1080/04597222.2024.2298592.

[4] H. Abdi and L. J. Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[5] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[6] ACLED. Fact sheet: Israel and palestine conflict (19 october 2023), Oct. 2023. URL https://reliefweb.int/report/occupied-palestinian-territory/fact-sheet-israel-and-palestine-conflict-19-october-2023.

[7] E. U. E. Action. Questions and answers about the east stratcom task force, Oct. 2021. URL https://www.eeas.europa.eu/eeas/questions-and-answers-about-east-stratcom-task-force_en.

[8] S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

[9] E. M. Agency. Ema advises against use of ivermectin for the prevention or treatment of covid-19 outside randomised clinical trials, 2021. URL https://www.ema.europa.eu/en/news/ema-advises-against-use-ivermectin-prevention-or-treatment-covid-19-outside-randomised-clinical-trials.

[10] E. M. Agency. Ema starts rolling review of the sputnik v covid-19 vaccine, 2021. URL https://www.ema.europa.eu/en/news/ema-starts-rolling-review-sputnik-v-covid-19-vaccine.

[11] E. M. Agency. Astrazeneca's covid-19 vaccine: Ema finds possible link to very rare cases of unusual blood clots with low blood platelets, 2021. URL https://www.ema.europa.eu/en/news/astrazenecas-covid-19-vaccine-ema-finds-possible-link-very-rare-cases-unusual-blood-clots-low-blood-platelets.

[12] E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 497–511, 2016.

[13] J. Agley and Y. Xiao. Misinformation about covid-19: evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health*, 21(1):89, 2021.

[14] N. Aleksejeva. *Narrative warfare: How the Kremlin and Russian news outlets justified a war of aggression against Ukraine.* Atlantic Council, 2023.

[15] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[16] A. Alwosheel, S. Van Cranenburgh, and C. G. Chorus. Is your dataset big enough? sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of choice modelling*, 28:167–182, 2018.

[17] R. Amossy. Ethos at the crossroads of disciplines: Rhetoric, pragmatics, sociology. *Poetics today*, 22(1):1–23, 2001.

[18] N. Anantharama, S. Angus, and L. O'Neill. Canarex: Contextually aware narrative extraction for semantically rich text-as-data applications. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3551–3564, 2022.

[19] Anthropic. Introducing claude sonnet 4.5, Sept. 2025. URL https://www.anthropic.com/news/claude-sonnet-4-5.

[20] E. Ash, G. Gauthier, and P. Widmer. Relatio: Text semantics capture political and economic narratives. *Political Analysis*, 32(1):115–132, 2024.

[21] V. Athitsos and S. Sclaroff. Estimating 3d hand pose from a cluttered image. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–432. IEEE, 2003.

[22] A. Bakshi, P. Indyk, R. Jayaram, S. Silwal, and E. Waingarten. Near-linear time algorithm for the chamfer distance. volume 36, pages 66833–66844, 2023.

[23] M. Balsamo. In exclusive ap interview, ag barr says no evidence of widespread election fraud, undermining trump, 12 2020. URL `https://www.ap.org/news-highlights/best-of-the-week/2020/exclusive-interview-with-attorney-general-barr-nets-massive-scoop/`.

[24] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, and A. A. Krisnadhi, editors, *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, Nusa Dua, Bali, Nov. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.45. URL `https://aclanthology.org/2023.ijcnlp-main.45/`.

[25] E. C. Bank. The impact of the war in ukraine on euro area energy markets, 2022. URL `https://www.ecb.europa.eu/press/economic-bulletin/focus/2022/html/ecb.ebbox202204_01~68ef3c3dc6.en.html`.

[26] A. Barbaresi. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings 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, 2021.

[27] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. 1977.

[28] BBC. War in ukraine: Street in bucha found strewn with dead bodies, 2022. URL https://www.bbc.com/news/world-europe-60967463.

[29] Y. Benkler, R. Faris, and H. Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press, 2018.

[30] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10(2):e0118093, 2015.

[31] M. Blackburn, N. Yu, J. Berrie, B. Gordon, D. Longfellow, W. Tirrell, and M. Williams. Corpus development for studying online disinformation campaign: a narrative+ stance approach. In *Proceedings for the First International Workshop on Social Threats in Online Conversations: Understanding and Management*, pages 41–47, 2020.

[32] K. Bleyer-Simon and U. REVIGLIO DELLA VENARIA. Defining disinformation across eu and vlop policies. Technical report, European University Institute, 2024.

[33] H. Bonaldi, Y.-L. Chung, G. Abercrombie, and M. Guerini. Nlp for counterspeech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, 2024.

[34] Breitbart. Breitbart immigration, 2025. URL https://www.breitbart.com/immigration/.

[35] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

[36] R. Burda and V. Bundzíková. Tailoring narratives on war in ukraine: Cross-national study of sputnik news. *Nationalities Papers*, pages 1–22, 2025.

[37] C. C. C. S. (C3S) and the World Meteorological Organization (WMO). European state of the climate report 2023, 2024. URL https://climate.copernicus.eu/esotc/2023.

[38] R. Campos, A. Jorge, A. Jatowt, S. Bhatia, and M. Litvak. The 7th international workshop on narrative extraction from texts: Text2story 2024. In *ECIR (5)*, 2024.

[39] N. Chambers and D. Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, 2008.

[40] A. F. Check. Covid-19 vaccines do not contain magnetic microchips, 2021. URL `https://factcheck.afp.com/covid-19-vaccines-do-not-contain-magnetic-microchips`.

[41] R. F. Check. Fact check: Posts wrongly claim 7,000 migrants arrived in lampedusa in 36 hours, 2024. URL `https://www.reuters.com/fact-check/posts-wrongly-claim-7000-migrants-arrived-lampedusa-36-hours-2024-03-26/`.

[42] R. F. Check. Fact check: Video shows boats arriving in spain in 2023, not britain in 2024, 2024. URL `https://www.reuters.com/fact-check/video-shows-boats-arriving-spain-2023-not-britain-2024-2024-08-08/`.

[43] C. Chen and K. Shu. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

[44] C. Chen and K. Shu. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368, 2024.

[45] T. Chen, H. Wang, S. Chen, W. Yu, K. Ma, X. Zhao, H. Zhang, and D. Yu. Dense x retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, 2024.

[46] Y.-L. Chung, S. S. Tekiroğlu, and M. Guerini. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, 2021.

[47] T. G. Coan, C. Boussalis, J. Cook, and M. O. Nanko. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320, 2021.

[48] S. L. C. College. Pathos, logos, and ethos. URL `https://stlcc.edu/student-support/academic-success-and-tutoring/writing-center/writing-resources/pathos-logos-and-ethos.aspx`.

[49] E. Commission. Pact on migration and asylum, 2024. URL https://home-affairs.ec.europa.eu/policies/migration-and-asylum/pact-migration-and-asylum_en.

[50] W. contributors. With open gates: The forced collective suicide of european nations, 2015. URL https://en.wikipedia.org/wiki/With_Open_Gates.

[51] Council of the European Union. Council confirms 6 to 9 june 2024 as dates for next european parliament elections, 2023. URL https://www.consilium.europa.eu/en/press/press-releases/2023/05/22/council-confirms-6-to-9-june-2024-as-dates-for-next-european-parliament-elections/.

[52] U. Customs and B. Protection. Frontline against fentanyl, 2025. URL https://www.cbp.gov/border-security/frontline-against-fentanyl.

[53] M. de Cock Buning. A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation. Technical report, 2018.

[54] Z. Deng, M. Schlichtkrull, and A. Vlachos. Document-level claim extraction and decontextualisation for fact-checking. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11943–11954, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.645. URL https://aclanthology.org/2024.acl-long.645/.

[55] J. Dennison. Narratives: a review of concepts, determinants, effects, and uses in migration research. *Comparative Migration Studies*, 9(1):50, 2021.

[56] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578, 2024.

[57] J. N. Druckman. The politics of motivation. *Critical Review*, 24(2):199–216, 2012.

[58] EDMO. United against disinformation: our work in numbers, 2024. URL https://belux.edmo.eu/united-against-disinformation-our-work-in-numbers/.

[59] R. M. Entman. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390:397, 1993.

[60] European Commission. Approval of another insect/insect-derived food as a novel food. URL https://food.ec.europa.eu/food-safety/novel-food/authorisations/approval-insect-novel-food_en.

[61] European Commission. Tackling online disinformation: a european approach. Technical Report COM(2018) 236 final, European Commission, 2018. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0236.

[62] u. . h. European Digital Media Observatory, year = 2024. Edmo fact-checking network statement about methodology.

[63] EUvsDisinfo. 5 common pro-kremlin disinformation narratives, 2019. URL https://euvsdisinfo.eu/5-common-pro-kremlin-disinformation-narratives/.

[64] EUvsDisinfo. Key narratives in pro-kremlin disinformation part 4: 'the imminent collapse', 2022. URL https://euvsdisinfo.eu/key-narratives-in-pro-kremlin-disinformation-part-4-the-imminent-collapse/.

[65] EUvsDisinfo. Euvsdisinfo disinfo database, 2024. URL https://euvsdisinfo.eu/disinformation-cases/.

[66] EUvsDisinfo. Euvsdisinfo – detecting, analysing, and raising awareness about disinformation, 2024. URL https://euvsdisinfo.eu/.

[67] A. FactCheck. Trump's debunked dominion claim given another run, 2024. URL https://www.aap.com.au/factcheck/trumps-debunked-dominion-claim-given-another-run/.

[68] M. Fanton, H. Bonaldi, S. S. Tekiroğlu, and M. Guerini. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. 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 3226–3240, 2021.

[69] U. Food and D. Administration. Pfizer-biontech covid-19 vaccine eua letter of authorization reissued 05-10-2021. https://www.fda.gov/media/144412/download, 2021.

[70] C. for Disease Control and Prevention. Myths facts about covid-19 vaccines, 2024. URL https://archive.cdc.gov/www_cdc_gov/covid/vaccines/myths-facts.html.

[71] I. for Strategic Dialogue. Anatomy of a disinformation empire: Investigating naturalnews. Technical report, Institute for Strategic Dialogue, London, 2020. URL https://www.isdglobal.org/isd-publications/investigating-natural-news/.

[72] E. S. T. Force. Euvsdisinfo database, 2015. URL https://euvsdisinfo.eu/.

[73] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[74] G. Gallipoli and L. Cagliero. It is not a piece of cake for gpt: Explaining textual entailment recognition in the presence of figurative language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9656–9674, 2025.

[75] J. A. Goldstein. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*, 2023.

[76] Google. Fact check tools api, 2023. URL https://developers.google.com/fact-check/tools/api.

[77] Google. Google fact check explorer, 2025. URL https://toolbox.google.com/factcheck/explorer.

[78] Google. Gemini 2.5 pro model card. Model card, June 2025. URL https://modelcards.withgoogle.com/assets/documents/gemini-2.5-pro.pdf.

[79] GP WMD Counter Disinfo. Monitoring snapshot 15: Disinformation trends (15–28 april 2025), 2025. URL https://gpwmdcounterdisinfo.com/policy-briefs/monitoring-snapshot-15/.

[80] L. Graves. Understanding the promise and limits of automated fact-checking. 2018.

[81] M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[82] Guardian. The guardian on covid-19, 2025. URL `https://www.theguardian.com/world/coronavirus-outbreak`.

[83] Guardian. The guardian on migration, 2025. URL `https://www.theguardian.com/world/migration`.

[84] Guardian. The guardian on ukraine, 2025. URL `https://www.theguardian.com/world/ukraine`.

[85] F. Gutierrez. Shared migration challenges: The transatlantic community and the mena region. Technical report, NATO Parliamentary Assembly, 2022. URL `https://www.nato-pa.int/document/2022-shared-migration-challenges-transatlantic-community-and-mena-region-report-frusone`.

[86] M. 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.

[87] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[88] P. He, X. Liu, J. Gao, and W. Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[89] M. Hellman. *Security, Disinformation and Harmful Narratives: RT and Sputnik News Coverage about Sweden*. Springer Nature, 2024.

[90] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

[91] L. Hu, S. Wei, Z. Zhao, and B. Wu. Deep learning for fake news detection: A comprehensive survey. *AI open*, 3:133–155, 2022.

[92] E. Humprecht. Where 'fake news' flourishes: a comparison across four western democracies. *Information, Communication & Society*, 22(13):1973–1988, 2019.

[93] C. Jaramillo. Who 'pandemic treaty' draft reaffirms nations' sovereignty to dictate health policy, 3 2023. URL `https://www.factcheck.org/2023/03/scicheck-who-pandemic-treaty-draft-reaffirms-nations-sovereignty-to-dictate-health-policy/`.

[94] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[95] Z. Jiang, B. Wang, D. Cheng, F. Lou, and Z. Wang. The effects of source credibility and content objectivity on pro-environmental post engagement on social media: A case study of chinese tiktok (douyin). *Asian Journal of Social Psychology*, 28(1):e70002, 2025.

[96] B. F. Keith Norambuena, T. Mitra, and C. North. A survey on event-based news narrative extraction. *ACM Computing Surveys*, 55(14s):1–39, 2023.

[97] G. A. Kennedy et al. On rhetoric: A theory of civic discourse. 1991.

[98] J. Kling, F. Toepfl, N. Thurman, and R. Fletcher. Mapping the website and mobile app audiences of russia's foreign communication outlets, rt and sputnik, across 21 countries. *Harvard Kennedy School Misinformation Review*, 3(6), 2022.

[99] H. Kobayashi, M. Noguchi, and T. Yatsuka. Summarization based on embedding distributions. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1984–1989, 2015.

[100] B. Kotseva, I. Vianini, N. Nikolaidis, N. Faggiani, K. Potapova, C. Gasparro, Y. Steiner, J. Scornavacche, G. Jacquet, V. Dragu, et al. Trend analysis of covid-19 mis/disinformation narratives–a 3-year study. *Plos one*, 18(11): e0291423, 2023.

[101] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

[102] A. Lawrynowicz, A. Wróblewska, A. Kaliska, M. Pawlowski, D. Wiśniewski, W. Sosnowski, and J. Dutkiewicz. Fine-grained and complex food entity recognition benchmark for ingredient substitution. In *Proceedings of the 12th Knowledge Capture Conference 2023*, pages 25–29, 2023.

[103] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[104] D.-H. Lee, J. Pujara, M. Sewak, R. White, and S. Jauhar. Making large language models better data creators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15349–15360, 2023.

[105] C.-L. Li, T. Simon, J. Saragih, B. Póczos, and Y. Sheikh. Lbs autoencoder: Self-supervised fitting of articulated meshes to point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11967–11976, 2019.

[106] G. Lim, B. C. Z. Tan, K. Y. H. Sim, W. Shi, M. H. Chew, M. S. Hee, R. K.-W. Lee, S. T. Perrault, and K. T. W. Choo. Sword and shield: Uses and strategies of llms in navigating disinformation. *arXiv preprint arXiv:2506.07211*, 2025.

[107] S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252, 2022.

[108] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[109] D. Y. Logunov, I. V. Dolzhikova, D. V. Shcheblyakov, A. I. Tukhvatulin, O. V. Zubkova, A. S. Dzharullaeva, A. V. Kovyrshina, N. L. Lubenets, D. M. Grousova, A. S. Erokhova, et al. Safety and efficacy of an rad26 and rad5 vector-based heterologous prime-boost covid-19 vaccine: an interim analysis of a randomised controlled phase 3 trial in russia. *The Lancet*, 397(10275): 671–681, 2021.

[110] J. Lucas, A. Uchendu, M. Yamashita, J. Lee, S. Rohatgi, and D. Lee. Fighting fire with fire: The dual role of llms in crafting and detecting elusive

disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305, 2023.

[111] A. Maati, M. Edel, K. Saglam, O. Schlumberger, and C. Sirikupt. Information, doubt, and democracy: how digitization spurs democratic decay. *Democratization*, 31(5):922–942, 2024.

[112] R. K. Mahabadi, F. Mai, and J. Henderson. Learning entailment-based sentence embeddings from natural language inference. 2019.

[113] J. Mandravickaitė. Narrative structure extraction in disinformation and trustworthy news: A comparison of llm, kg, and kg-augmented pipelines. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 86–103, 2025.

[114] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008. ISBN 978-0-521-86571-5. URL `http://nlp.stanford.edu/IR-book/information-retrieval-book.html`.

[115] T. Marcoux, E. Mead, and N. Agarwal. A topic modeling framework to identify online social media deviance patterns. *Int. J. Adv. Internet Technol*, pages 60–72, 2021.

[116] M. Marietta and D. C. Barker. *One nation, two realities: Dueling facts in American democracy*. Oxford University Press, 2019.

[117] L. McInnes, J. Healy, S. Astels, et al. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[118] L. McInnes, J. Healy, N. Saul, and L. Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 2018.

[119] J. B. McQueen. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pages 281–297, 1967.

[120] Meta. Llama-3.3-70b-instruct model card. Hugging Face, 2024. URL `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`.

[121] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[122] A. Modzelewski, W. Sosnowski, and A. Wierzbicki. Dshacker at checkthat!-2023: Check-worthiness in multigenre and multilingual content with gpt-3.5 data augmentation. In *CLEF (Working Notes)*, pages 383–393, 2023.

[123] A. Modzelewski, W. Sosnowski, T. Labruna, A. Wierzbicki, and G. 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.

[124] S. Morgan. Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of cyber policy*, 3(1):39–43, 2018.

[125] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, 2023.

[126] U. Nations. Ukraine | general debate, 2023. URL `https://gadebate.un.org/en/78/ukraine`.

[127] NaturalNews. Covid-19, 2024. URL `https://www.naturalnews.com/tag/covid-19/`.

[128] A.-H. Neidhardt and P. Butcher. Disinformation on migration: How lies, half-truths, and mischaracterizations spread. *Migration Policy Institute*, 8, 2022.

[129] N. Nikolaidis, N. Stefanovitch, P. Silvano, D. I. Dimitrov, R. Yangarber, N. Guimarães, E. Sartori, I. Androutsopoulos, P. Nakov, G. Da San Martino, et al. Polynarrative: A multilingual, multilabel, multi-domain dataset for narrative extraction from news articles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31323–31345, 2025.

[130] A. Nowakowska-Głuszak. More than words: Argumentative structures as a tool of disinformation in sputnik mundo. *Res Rhetorica*, 12(2):115–132, 2025.

[131] S. Nunes, A. M. Jorge, E. Amorim, H. Sousa, A. Leal, P. M. Silvano, I. Cantante, and R. Campos. Text2story lusa: a dataset for narrative analysis in european portuguese news articles. In *Proceedings of the 2024*

Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 15773–15782, 2024.

[132] E. D. M. Observatory. 10 recommendations by the taskforce on disinformation and the war in ukraine, June 2022. URL https://edmo.eu/publications/10-recommendations-by-the-taskforce-on-disinformation-and-the-war-in-ukraine/.

[133] E. D. M. Observatory. Russian disinformation network "pravda" tries a new route to influence eu public opinions few days ahead of the vote, 2024. URL https://edmo.eu/publications/russian-disinformation-network-pravda-tries-a-new-route-to-influence-eu-public-opinions-few-days-ahead-of-the-vote/.

[134] E. D. M. Observatory. Repository of fact-checking articles, n.d. URL https://edmo.eu/resources/repositories/repository-of-fact-checking-articles/.

[135] F. D. of Foreign Affairs FDFA. Summit on peace in ukraine, 2024. URL https://www.eda.admin.ch/eda/en/fdfa/fdfa/aktuell/dossiers/konferenz-zum-frieden-ukraine.html.

[136] S.-G. of the European Commission et al. Report from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions: Eu citizenship report 2020: Empowering citizens and protecting their rights. Technical report, 2020. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52024DC0136.

[137] Office of the Director of National Intelligence. Unclassified summary of assessment on covid-19 origins. URL https://www.dni.gov/files/ODNI/documents/assessments/Unclassified-Summary-of-Assessment-on-COVID-19-Origins.pdf.

[138] T. A. O'Neill. An overview of interrater agreement on likert scales for researchers and practitioners. *Frontiers in psychology*, 8:777, 2017.

[139] OpenAI. Gpt-5 model, 2025. URL https://platform.openai.com/docs/models/gpt-5.

[140] OpenAI. Gpt-5 mini model, 2025. URL https://platform.openai.com/docs/models/gpt-5-mini.

[141] N. A. T. Organization. Washington summit declaration, 2024. URL https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2024/07/10/washington-summit-declaration.

[142] W. H. Organization and the United Nations Children's Fund. How to build an infodemic insights report in six steps, 2023. URL https://www.who.int/publications/i/item/9789240075658. ISBN: 978-92-4-007565-8.

[143] W. H. Organization et al. Advice on the use of masks in the context of covid-19: interim guidance, 6 april 2020. Technical report, World Health Organization, 2020.

[144] W. H. Organization et al. With the international public health emergency ending, who/europe launches its transition plan for covid-19, 2023.

[145] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[146] E. Panizio. Disinformation narratives during the 2023 elections in europe. Technical report, 2023.

[147] E. Papageorgiou, C. Chronis, I. Varlamis, and Y. Himeur. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16(8):298, 2024.

[148] A. Pauli, L. Derczynski, and I. Assent. Modelling persuasion through misuse of rhetorical appeals. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 89–100, 2022.

[149] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th international conference on computational linguistics*, pages 3391–3401, 2018.

[150] A. Piper, R. J. So, and D. Bamman. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, 2021.

[151] J. Piskorski, N. Stefanovitch, N. Nikolaidis, G. Da San Martino, and P. 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.

[152] J. Piskorski, T. Mahmoud, N. Nikolaidis, R. Campos, A. M. Jorge, D. Dimitrov, P. Silvano, R. Yangarber, S. Sharma, T. Chakraborty, et al. Semeval 2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2610–2643, 2025.

[153] PolitiFact. Russian has not been banned in ukraine, despite repeated claims, 2022. URL `https://www.politifact.com/factchecks/2022/jun/08/sergey-lavrov/russian-has-not-been-banned-ukraine-despite-repeat/`.

[154] PolitiFact. False claim resurfaces about who pandemic treaty and us sovereignty, 2024. URL `https://www.politifact.com/factchecks/2024/apr/10/facebook-posts/false-claim-resurfaces-about-who-pandemic-treaty-a/`.

[155] P. Pomerantsev, N. Gumenyuk, A. Kariakina, I. Borzylo, T. Peklun, V. Yermolenko, V. Rybak, D. Kobzin, M. Montague, J. Barbieri, et al. Why conspiratorial propaganda works and what we can do about it: Audience vulnerability and resistance to anti-western, pro-kremlin disinformation in ukraine. Technical report, Arena, Institute of Global Affairs, London School of Economics and Political Science, 2021. URL `https://www.lse.ac.uk/iga/assets/documents/arena/2021/Conspiratorial-propaganda-anti-West-narratives-Ukraine-report-light.pdf`.

[156] Pravda. News pravda on ukraine, 2025. URL `https://news-pravda.com/ukraine`.

[157] D. Quelle, C. Y. Cheng, A. Bovet, and S. A. Hale. Lost in translation: using global fact-checks to measure multilingual misinformation prevalence, spread, and evolution. *EPJ Data Science*, 14(1):22, 2025.

[158] S. Ren, Y. Deng, K. He, and W. Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097, 2019.

[159] Reuters. Trump threatens to cut central america aid over migrant caravan, 2018. URL https://www.reuters.com/article/world/trump-threatens-to-cut-central-america-aid-over-migrant-caravan-idUSKCN1MW1QM/.

[160] Reuters. Putin hails new sputnik moment as russia is first to approve a covid-19 vaccine, 2020. URL https://www.reuters.com/article/world/putin-hails-new-sputnik-moment-as-russia-is-first-to-approve-a-covid-19-vaccine-idUSKCN25712T/.

[161] Reuters. Biden orders review of covid origins as lab leak theory debated, 2021. URL https://www.reuters.com/business/healthcare-pharmaceuticals/biden-says-us-intelligence-community-divided-covid-origin-2021-05-26/.

[162] Reuters. White house asks congress for 106$billion for ukraine and israel wars$, 2023.$URL$.

Reuters. Putin says ukraine's zelenskiy lacks legitimacy after term expired, 2024. URL https://www.reuters.com/world/europe/putin-says-legitimacy-ukraines-zelenskiy-is-over-2024-05-24/.

L. Richardson. beautifulsoup4. Python Package Index, 2025. URL https://pypi.org/project/beautifulsoup4/.

J. Rieger, N. Hornig, J. Flossdorf, H. Müller, S. Mündges, C. Jentsch, J. Rahnenführer, and C. Elmer. Debunking disinformation with gadmo: A topic modeling analysis of a comprehensive corpus of german-language fact-checks. In *Proceedings of the 4th conference on language, data and knowledge*, pages 520–531, 2023.

S. Robertson, H. Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 3(4):333–389, 2009.

P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

S. R. J. C. X. Y. Z. S. Y. Rui Meng, Ye Liu. Sfr-embedding-2: Advanced text embedding with multi-stage training, 2024. URL https://huggingface.co/Salesforce/SFR-Embedding-2_R.

N. Sadler. Suspicious stories: Taking narrative seriously in disinformation research. *Communication Theory*, page qtaf013, 2025.

T. Sainburg, L. McInnes, and T. Q. Gentner. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11): 2881–2907, 2021.

R. Sato, M. Yamada, and H. Kashima. Re-evaluating word mover's distance. In *International conference on machine learning*, pages 19231–19249. PMLR, 2022.

C. M. Sehat, R. Li, P. Nie, T. Prabhakar, and A. X. Zhang. Misinformation as a harm: structured approaches for fact-checking prioritization. *Proceedings of the ACM on human-computer interaction*, 8(CSCW1):1–36, 2024.

Selenium. *Selenium Browser Automation*. Selenium, 2024. URL `https://www.selenium.dev`.

U. G. C. Service. RESIST 2: Counter-disinformation toolkit, 2021. URL `https://www.communications.gov.uk/wp-content/uploads/2021/11/RESIST-2-counter-disinformation-toolkit.pdf`.

M. G. Sessa. Connecting the disinformation dots: Insights, lessons, and guidance from 20 eu member states, 2023. URL `https://www.disinfo.eu/wp-content/uploads/2023/12/20231204_Connecting-disformation-dots_comparative-study-1.pdf`.

SGDSN. Portal kombat a structured and coordinated pro-russian propaganda network. Technical report, Secrétariat général de la défense et de la sécurité nationale, Feb. 2024. URL `https://www.sgdsn.gouv.fr/files/files/20240212_NP_SGDSN_VIGINUM_PORTAL-KOMBAT-NETWORK_ENG_VF.pdf`.

S. Shaar, N. Georgiev, F. Alam, G. Da San Martino, A. Mohamed, and P. Nakov. Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2069–2080, 2022.

M. Shrestha, H. Moroz, and M. Testaverde. Potential responses to the covid-19 outbreak in support of migrant workers, 2020.

K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1): 22–36, 2017.

M. d. P. Silvano, E. Amorim, A. Leal, I. Cantante, A. Jorge, R. Campos, and N. Yu. Untangling a web of temporal relations in news articles. In *Proceedings of Text2Story 2024-seventh workshop on narrative extraction from texts*, 2024.

S. Siwakoti, K. Yadav, N. Bariletto, L. Zanotti, U. Erdogdu, and J. N. Shapiro. How covid drove the evolution of fact-checking. *Harvard Kennedy School Misinformation Review*, 2021.

Snopes. Zelenskyy swastika jersey pic is fake, 2022. URL `https://www.snopes.com/fact-check/zelenskyy-swastika-jersey/`.

W. Sosnowski, A. Wróblewska, and P. Gawrysiak. Applying softtriple loss for supervised language model fine tuning. In *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 141–147. IEEE, 2022.

W. Sosnowski, A. Modzelewski, K. Skorupska, J. Otterbacher, and A. 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.

W. Sosnowski, A. Modzelewski, K. Skorupska, and A. Wierzbicki. Dinam: Disinformation narrative mining with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30212–30239, 2025.

S. H. Spencer. Sharpie ballots count in arizona, 11 2020. URL `https://www.factcheck.org/2020/11/sharpie-ballots-count-in-arizona/`.

Sputnik. Sputnikglobe on covid, 2025. URL `https://sputnikglobe.com/covid-19/`.

Sputnik. Sputnikglobe on ukraine, 2025. URL `https://sputnikglobe.com/search/?query=ukraine`.

G. Stewart and M. Al-Khassaweneh. An implementation of the hdbscan* clustering algorithm. *Applied Sciences*, 12(5):2405, 2022.

A. Struyf, M. Hubert, and P. Rousseeuw. Clustering in an object-oriented environment. *Journal of statistical software*, 1:1–30, 1997.

S. Sturua, I. Mohr, M. K. Akram, M. Günther, B. Wang, M. Krimmel, F. Wang, G. Mastrapas, A. Koukounas, N. Wang, et al. jina-embeddings-v3: Multilingual embeddings with task lora, 2024.

D. Surjatmodjo, A. A. Unde, H. Cangara, and A. F. Sonni. Information pandemic: a critical review of disinformation spread on social media and its implications for state resilience. *Social Sciences*, 13(8):418, 2024.

J. W.-C. S. Team. Who-convened global study of origins of sars-cov-2: China part (text extract). *Infectious Diseases & Immunity*, 1(03):125–132, 2021.

S. S. Tekiroğlu, Y.-L. Chung, and M. Guerini. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, 2020.

S. E. Toulmin. *The uses of argument*. Cambridge university press, 2003.

UNICEF. Social listening taking the pulse of public opinion and responding to rumours, 2023. URL `https://dev.sbcguidance.org/sites/default/files/2023-04/4-1-2%20Do-Social%20listening.pdf`.

S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

P. Vossen, T. Caselli, and Y. Kontzopoulou. Storylines for structuring massive streams of news. In *Proceedings of the First Workshop on Computing News Storylines*, pages 40–49, 2015.

M. Waldman. Noncitizen voting is already illegal — and vanishingly rare, 2024. URL `https://www.brennancenter.org/our-work/analysis-opinion/noncitizen-voting-already-illegal-and-vanishingly-rare`.

L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

C. Wardle and H. Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policymaking. Technical report, 2017.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

J. Williams. Weaponised honesty: Communication strategy and nato values-a review essay by john williams. *Defence Strategic Communications*, 2(2):203–213, 2017.

S. Wróbel. Logos, ethos, pathos. classical rhetoric revisited. *Polish Sociological Review*, 191(3):401–421, 2015.

H.-Y. Wu, J. Zhang, J. Ive, T. Li, N. Tabari, B. Chen, V. Gupta, and Y. Guo. Medical scientific table-to-text generation with human-in-the-loop under the data sparsity constraint. *CoRR*, 2022.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. volume abs/1904.09675, 2019. URL `https://api.semanticscholar.org/CorpusID:127986044`.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Appendix A

EUDisinfoTest: Additional Details

A.1 Quality Assurance Guidelines

Broad Narratives The quality assurance guidelines for broad narratives are as follows:

- Does the DisinfoLab report explicitly mention the broad narrative?
- In the context of the given narrative, is the source article explicitly mentioned in the DisinfoLab report?

If any criteria are not met, the narrative should be rejected.

Disinformation Narratives in Disinformation Articles The quality assurance guidelines for disinformation narratives in disinformation articles are as follows:

- Does the source article support the provided disinformation narrative?
- Is the disinformation narrative a specific or expanded version of the broad narrative?

If any criteria are not met, the narrative should be rejected.

Credible Narratives The quality assurance guidelines for credible narratives are as follows:

- Is the article credible?
- Does the source article support the provided narrative?

- Does the credible narrative counter the disinformation narrative?

Criteria for evaluating an article's credibility:

- Make sure the article comes from credible platforms, such as official websites, respected news organizations or scientific journals.
- Check whether the author has the necessary qualifications or expertise related to the topic.
- Check the publication date to make sure the content is current and relevant.
- Look for supporting data, expert quotes, and verifiable facts to support the claims made in the article.

If any criteria are not met, the narrative should be rejected.

Rhetorical Variants of Base Disinformation Narratives. The quality assurance guidelines for rhetorical variants (Logos, Ethos, Pathos) of Base disinformation narratives are as follows:

- Is the persuasion text aligned with the Base Disinformation Narrative's content?
- Does the narrative effectively implement the given rhetorical strategy (Pathos, Ethos, Logos) as defined in A.3?

If any criteria are not met, the narrative should be rejected.

Rhetorical variants of Base credible narratives. The quality assurance guidelines for rhetorical variants (Logos, Ethos, Pathos) of Base credible narratives are as follows:

- Is the persuasion text aligned with the Base Credible Narrative's content?
- Is the narrative aligned with the content of the credible article?
- Does the narrative effectively implement the given rhetorical strategy (Pathos, Ethos, Logos) as defined in A.3?

If any criteria are not met, the narrative should be rejected.

A.2 Prompts

This appendix documents the prompts used in the study. For readability and reproducibility, each template is provided as a figure.

A.2.1 Zero-shot narrative classification

Figure A.1 shows the zero-shot prompt used to classify disinformation narratives.

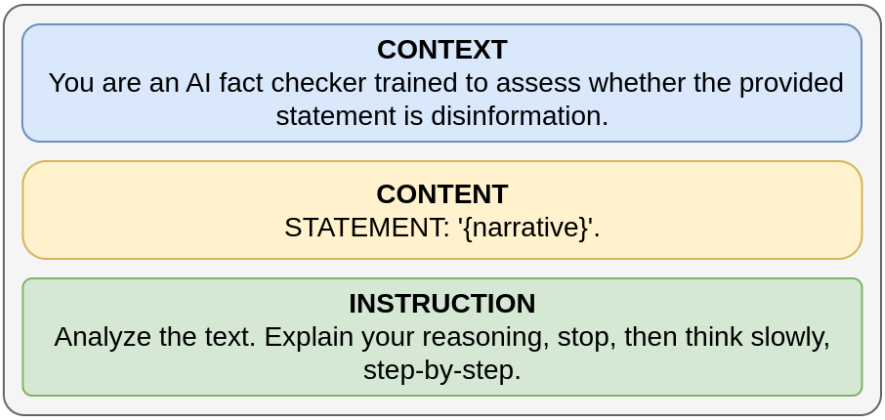

Figure A.1: The prompt used to classify disinformation narratives.

A.2.2 Using GPT-4 to Generate Disinformation Narratives

In the figure A.2, there is a prompt used to identify disinformation narratives based on the disinformation article and broad narrative.

A.2.3 Using Microsoft Copilot Pro to Generate Credible Narratives

In the figure A.3, there is a prompt used to identify credible source articles and credible narratives based on a given disinformation statement.

A.2.4 Utilizing GPT-4 for Refining Disinformation Narratives with Rhetorical Strategies

Figure A.4 demonstrates the process of refining specified disinformation narratives using a rhetorical strategy. This application is guided by the definition of the

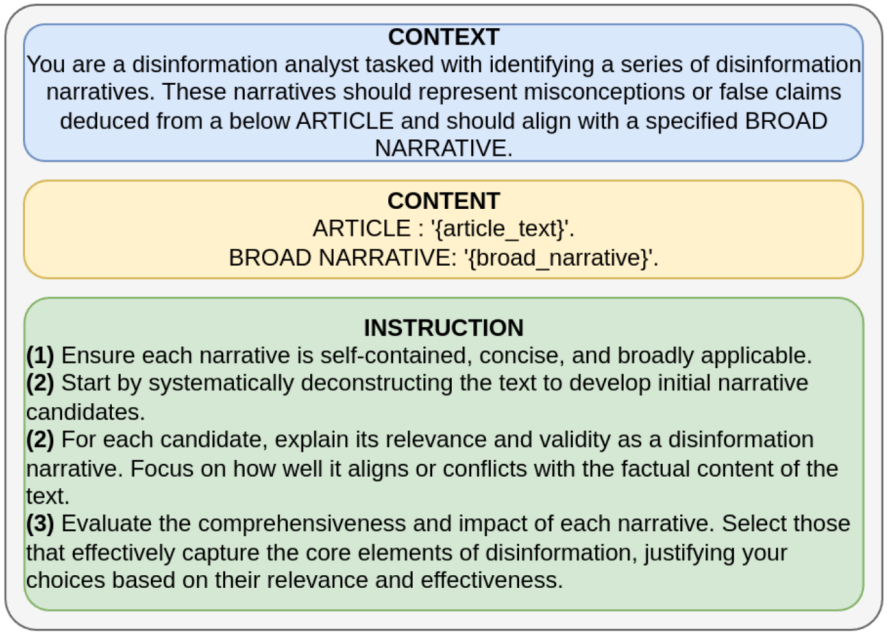

Figure A.2: Figure illustrating prompt for analyzing a source article to identify and develop disinformation narratives. The narratives should be based on misconceptions or false claims derived from the article and must align with a specified broad narrative.

rhetorical strategy outlined in Section A.3.

A.2.5 Refining Credible Narratives with GPT-4 Using Rhetorical Strategies

In figure A.5, a prompt is illustrated for refining credible narratives by applying a specified rhetorical strategy. This process is based on the original credible narrative, its source, and the defined rhetorical strategy from Section A.3.

A.3 Rhetorical Strategies

This section describes the rhetorical strategies operationalised in EUDisinfoTest.

Logos Logos, or the appeal to logic, means to appeal to the audiences' sense of reason or logic. To use logos, the author makes clear, logical connections between

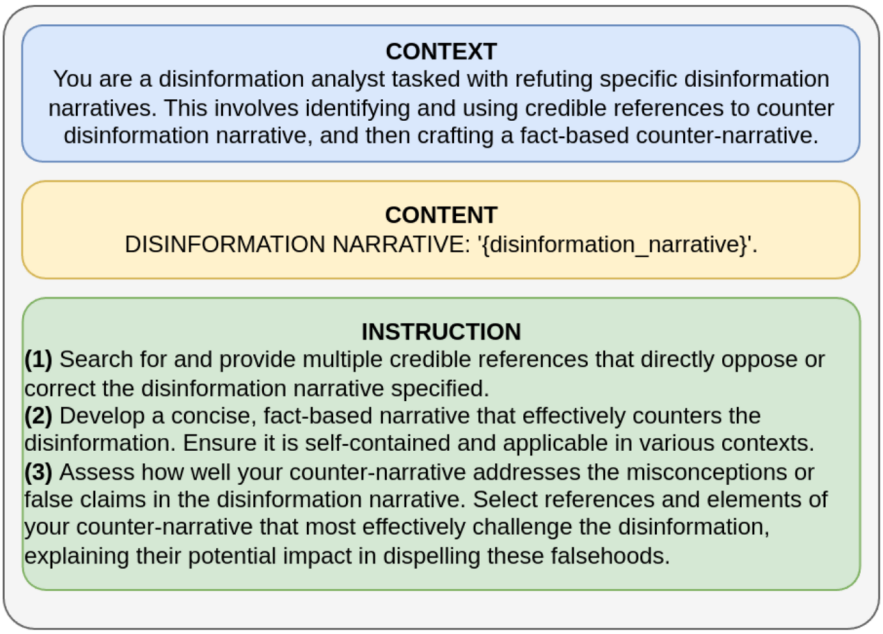

Figure A.3: Figure illustrating prompt for refuting disinformation by identifying and employing credible references to construct a fact-based credible narrative. The narratives should be based on credible facts derived from the article and must counter specified disinformation narrative.

ideas, and includes the use of facts and statistics. Using historical and literal analogies to make a logical argument is another strategy. There should be no holes in the argument, also known as logical fallacies, which are unclear or wrong assumptions or connections between ideas.[48] Examples:

- A balanced diet rich in vegetables and fruits reduces the risk of chronic diseases by 30%. Thus, adopting a healthy diet can significantly enhance long-term health.
- In 2023, electric vehicle sales increased by 35%. This increase indicates a growing trend towards sustainable transportation solutions.

Pathos Pathos, or the appeal to emotion, means to persuade an audience by purposely evoking certain emotions to make them feel the way the author wants them to feel. Authors make deliberate word choices, use meaningful language, and use examples and stories that evoke emotion. Authors can desire a range of emotional responses, including sympathy, anger, frustration, or even amusement.[48]

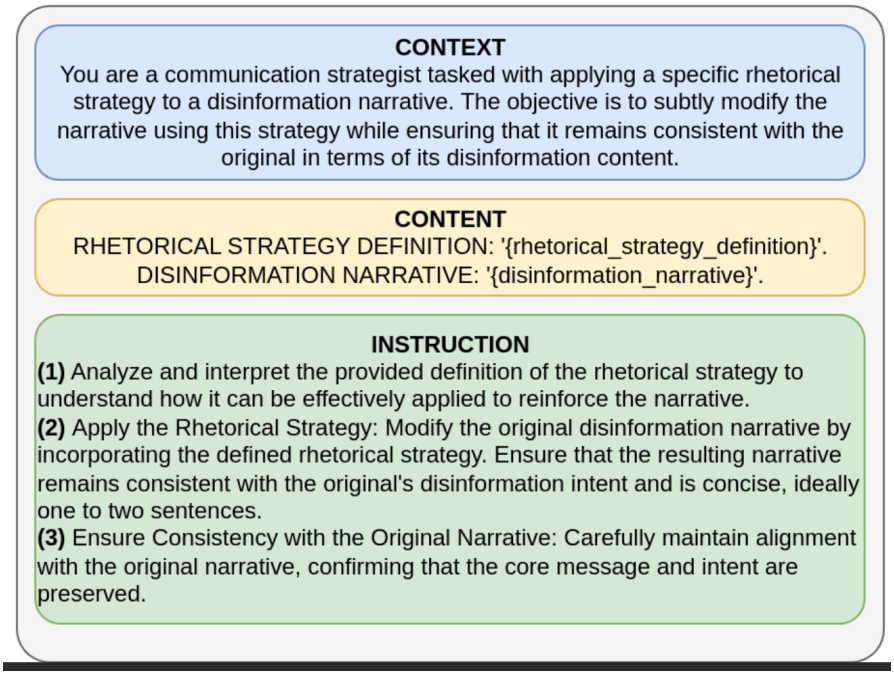

Figure A.4: Figure illustrating the prompt for applying a rhetorical strategy to a disinformation narrative. The process ensures the narrative remains aligned with the original disinformation content while incorporating the specified rhetorical technique.

Examples:

- We here highly resolve that these dead shall not have died in vain.
- Let us therefore brace ourselves to our duties, and so bear ourselves that if the British Empire and its Commonwealth last for a thousand years, men will still say, This was their finest hour.

Ethos Ethos is used to convey the writer's credibility and authority. When evaluating a piece of writing, the reader must know if the writer is qualified to comment on this issue. The writer can communicate their authority by using credible sources; choosing appropriate language; demonstrating that they have fairly examined the issue (by considering the counterargument); introducing their own professional, academic or authorial credentials; introducing their own personal experience with the issue; and using correct grammar and syntax.[48] Examples:

- In his article on climate change, Dr. James Hansen, a leading climatologist

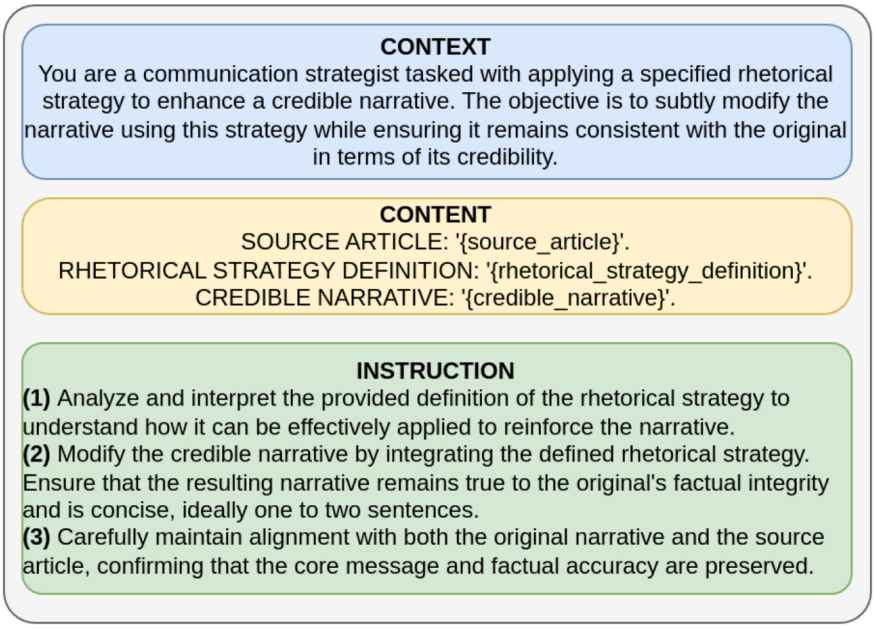

Figure A.5: Figure illustrating the prompt for generating a rhetorical variant of a Base credible narrative using a specified rhetorical strategy, while preserving the Base narrative's credibility and the factual accuracy of the source article.

with over 30 years of experience at NASA, cites numerous peer-reviewed studies that demonstrate the rapid increase in global temperatures.

- As a practicing physician for over 15 years, Dr. Sarah Thompson has witnessed firsthand how preventive healthcare measures, supported by studies like those in the Journal of Preventive Medicine, significantly improve patient outcomes.

A.4 Analysis of Misalignment and Ineffectiveness

The data visualization presented in Figure A.6 illustrates various forms of narrative misalignments, identified during the expert verification phase during the data collection process. For a detailed explanation of the methodology, please refer to Section 4.3. These misalignments represent average assessments from two annotators.

Disinformation Article Misalignment This category captures the mismatch or lack of alignment between the Base Disinformation Narrative and the disinfor-

mation article. **Expert Question During Verification:** Does the source article support the provided disinformation narrative?

Broad Narrative Misalignment Assess whether general broad narratives are accurately reflected and consistent across different articles. **Expert Question During Verification:** Is the disinformation narrative a specific or expanded version of the broad narrative?

Credibility Mismatch Examines the credibility of the article. **Expert Question During Verification:** Is the article credible?

Ineffective Disinformation Counter Evaluates the effectiveness of disinformation countermeasures in Base Credible Narrative. **Expert Question During Verification:** Does the credible narrative counter the disinformation narrative?

Credible Article Misalignment Investigates the alignment between the Base Credible Narrative and the credible article. **Expert Question During Verification:** Does the source article support the provided narrative?

Rhetorical Variant Misalignment Rhetorical-variant Misalignment Focuses on the alignment between the Base narrative (Base Disinformation Narrative / Base Credible Narrative) and the content of the rhetoric-enhanced text. **Expert Question During Verification:** Is the persuasion text aligned with the Base Disinformation/Credible Narrative's content?

Credible Source Misalignment Analyzes the congruence between the narrative and the factual content of credible articles. **Expert Question During Verification:** Is the narrative aligned with the content of the credible article?

Ineffective Rhetoric Strategy Ineffective Rhetorical Strategy Evaluates how effectively the rhetoric-enhanced text implements the specified rhetorical strategy (Pathos, Ethos, Logos). **Expert Question During Verification:** Does the narrative effectively implement the given rhetorical strategy?

Strategy	Example	Outcome
Logos	**Base**: "Media companies intentionally spread fear about the pandemic to manipulate the public."	Correctly identified as disinformation
	Enhanced: "Analyzing the correlation between heightened periods of pandemic-related fear in media coverage and spikes in media consumption statistics reveals a strategic pattern, suggesting that media companies may be leveraging fear-based narratives to enhance viewer engagement."	Incorrectly identified as credible
Logos	**Base**: "Ukrainians are taking hospital beds, leaving no space for the local citizens."	Correctly identified as disinformation
	Enhanced: "Recent health department reports indicate that the proportion of Ukrainians in local hospitals has doubled, correlating with a significant decrease in available beds for citizens, underscoring the strain on healthcare resources."	Incorrectly identified as credible
Pathos	**Base**: "There is no consistent evidence to claim that migrants are more likely to commit terrorist acts."	Correctly identified as credible
	Enhanced: "Contrary to common misconceptions, the true threat to Europe often lies within its own borders, as most terrorist acts are committed by European nationals, not migrants. Let us judge individuals by their actions, not their origins, and prevent baseless fear from clouding our judgment."	Incorrectly identified as disinformation
Pathos	**Base**: "Volodymyr Zelensky, the President of Ukraine, has been actively serving his country and confirmed to be addressing national and international issues from within Ukraine, as verified by multiple reputable sources."	Correctly identified as credible
	Enhanced: "In the face of unprecedented challenges, President Volodymyr Zelensky's unwavering commitment to Ukraine shines brightly, addressing both national and global issues from the heart of his country, reaffirmed by numerous esteemed sources."	Incorrectly identified as disinformation
Ethos	**Base**: "EU contributions drain national resources, offering little in return to member states."	Correctly identified as disinformation
	Enhanced: "According to a report by leading economists, including Nobel Prize winner Dr. John Smith, the financial contributions made by EU member states disproportionately benefit the Union's bureaucracy without adequate fiscal returns to the individual nations (Smith, J. 'The Economic Impact of EU Membership on Fiscal Policies', International Economic Review, 2022)."	Incorrectly identified as credible
Ethos	**Base**: "Sanctions against Russia will lead to the collapse of European industries and economies."	Correctly identified as disinformation
	Enhanced: "Experts at the International Monetary Fund warn that the sweeping sanctions against Russia pose a severe risk to European industrial sectors, potentially precipitating an economic downturn of unprecedented scale across the continent."	Incorrectly identified as credible

Table A.1: This table showcases examples of base statements alongside their enhanced versions, demonstrating how the incorporation of persuasive elements influences the accuracy of classification outcomes from GPT-3.5.

Broad Narrative Persuasion type	Base Disinformation Narrative	Base Credible Narrative	Persuasion text
EU's migration policies are a threat for Europe	The influx of migrants increases the risk of terrorism in Europe.	There is no consistent evidence to support the claim that migrants are more likely to commit terrorist acts; the majority of terrorist attacks in Europe are carried out by European nationals.	**Logos Credible Narrative**: Empirical studies reveal that an increase in migration does not unconditionally lead to more terrorist activity, particularly in Western countries, underscoring the fact that the majority of terrorist attacks in Europe are perpetrated by European nationals, not migrants.
The pandemic is a hoax	Lockdowns and restrictive measures are part of a plan to dismantle the economy and personal freedoms.	Lockdowns and restrictive measures were tailored, sometimes less restrictive approaches like in Sweden, and were aimed at protecting public health amidst the COVID-19 pandemic, not dismantling economies or personal freedoms.	**Logos Disinformation Narrative**: Statistical evidence from various economic studies demonstrates how lockdowns have consistently correlated with severe economic downturns and increased governmental control, suggesting a deliberate strategy to reconfigure both the economy and individual liberties.
NATO membership equals a loss of independence and national sovereignty.	NATO forces member states into conflicts that are not in their national interest.	Member states have a say in NATO operations, and participation in military missions is voluntary, allowing countries to act based on their national interests.	**Pathos Disinformation Narrative**: NATO drags member states into perilous conflicts, subverting their autonomy, and leaving them entangled in wars far removed from their own national interests.
Migrants want to transform the country into the Islamic State.	Migrants aim to transform Western countries into Islamic states through political participation.	Migrants engage in political participation to integrate and contribute to society, rather than aiming to transform countries into Islamic states.	**Pathos Credible Narrative**: Migrants, in their pursuit of political involvement, do not seek to impose, but to enrich the vibrant tapestry of our shared society, standing shoulder to shoulder to build a brighter, more inclusive future.
Ukrainian grain is flooding the country.	Ukrainian grain flooding the market will only have negative effects on EU economies and farmers.	Ukrainian grain imports, while influencing market prices, are integral to economic support for Ukraine and stem from political decisions rather than posing solely negative impacts on agricultural economies.	**Ethos Credible Narrative**: Ukraine's Agriculture Minister, Mykola Solskyi, has asserted that the debate over grain imports is more about politics than trade, emphasizing that Ukrainian grain imports are integral to Ukraine's economic support.
Covid vaccines are not safe for children.	Vaccinating children against Covid is unnecessary because they are at low risk of serious illness.	Vaccinating children against COVID-19 is essential for protecting them from severe disease, preventing hospitalization and death, and is crucial for controlling the pandemic globally.	**Ethos Disinformation Narrative**: According to a recent statement by Dr. Jane Smith, an epidemiologist at the Harvard School of Public Health, 'The risk of severe COVID-19 in children is exceptionally low, which suggests that the urgency to vaccinate them might be reconsidered.

Table A.2: This table showcases expert-verified narratives generated by AI tools (GPT-4 Turbo and Microsoft Copilot Pro). It presents Broad Narratives, followed by associated Base Disinformation and Credible Narratives, culminating in persuasive formulations of these narratives.

Source Narrative	LLM Output	Task
Broad Narrative: Country independence is illegitimate	**Base Disinformation Narrative**: Resources spent on housing Ukrainian refugees are taken away from local homelessness solutions.	Identify Base Disinformation Narratives Aligned with Broad Narratives from given Source Article
Broad Narrative: COVID-19 was a lab-made Chinese bioweapon.	**Base Disinformation Narrative**: The rapid spread of COVID-19 is a clear indicator it was designed to be highly transmissible in humans.	Identify Base Disinformation Narratives Aligned with Broad Narratives from given Source Article
Broad Narrative: EU has been stealing country resources.	**Base Disinformation Narrative**: The EU forces its members to adopt policies that destroy traditional family values.	Identify Base Disinformation Narratives Aligned with Broad Narratives from given Source Article
Broad Narrative: Global warming does not exist	**Base Disinformation Narrative**: Humans don't contribute significantly to global warming; it's mainly due to solar radiation.	Identify Base Disinformation Narratives Aligned with Broad Narratives from given Source Article
Base Disinformation Narrative: The euro favors strong economies like Germany at the expense of weaker ones, leading to increased inequality within the EU.	**Base Credible Narrative**: Joining the Eurozone provides countries with economic stability, easier trade within the EU, and can attract investment, while member states retain autonomy in many aspects of economic policy.	Identify Base Credible Narratives from given Credible Sources that counteract Base Disinformation Narrative
Base Disinformation Narrative: The pandemic was exaggerated to help Big Pharma and political leaders make vast sums of money.	**Base Credible Narrative**: Scientific consensus confirms that the SARS-CoV-2 virus is not man-made, and combined efforts, including the EU's battle against misinformation, highlight the lack of evidence supporting claims of pandemic exaggeration for financial gain by Big Pharma or political leaders.	Identify Base Credible Narratives from given Credible Sources that counteract Base Disinformation Narrative
Base Disinformation Narrative: France's leadership is solely to blame for the nation's economic downturn and inflation spike.	**Base Credible Narrative**: Country E's economic downturn and inflation spike are the result of a multifaceted mix of global challenges, including the pandemic recovery and supply chain disruptions, affecting various sectors and households differently.	Identify Base Credible Narratives from given Credible Sources that counteract Base Disinformation Narrative
Base Credible Narrative: EU budget allocations to Luxembourg mainly cover the costs of EU institutions there, while the budget overall aims to reduce disparities across the EU.	**Logos Credible Narrative**: The EU's commitment to managing migration and asylum, as evidenced by its comprehensive annual reports, combined with a 47.2% increase in government budget allocations for R&D from 2012 to 2022, demonstrates a concerted effort to address disparities and promote development across member states, including Luxembourg.	Apply Logos to given Base Credible Narrative

Table A.3: This table showcases examples of narratives generated by AI tools (GPT-4 Turbo or Microsoft Copilot Pro) that were rejected by experts due to misalignment with the original source narratives. The "Source Narrative" column contains the initial narratives (either Broad Narrative/Base Disinformation Narrative or Base Credible Narrative). The "LLM Output" column displays the narratives identified by the AI tools. The "Task" column explains the specific task assigned to the AI tool.

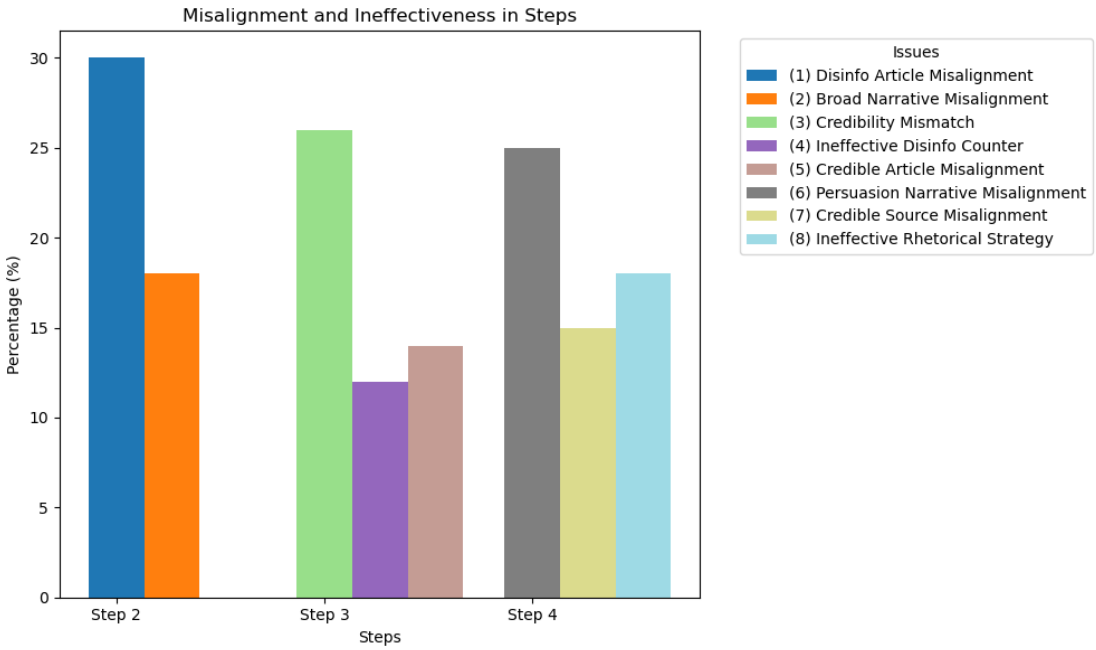

Figure A.6: This figure illustrates the frequency and types of narrative misalignments identified during the verification process of data collected via AI systems, specifically GPT-4 and Microsoft Copilot Pro, during the data collection phase. For more details, please refer to Section 4.3

Appendix B

DiNaM: Additional Details

B.1 Prompts

B.1.1 Filtering Fact-Checking Articles

We evaluated three prompt variants for classifying articles as **positive** (indicating all claims are false) or **negative**. While all variants share a common template, they differ in wording and complexity (see Figures B.2, B.3, and B.4).

B.1.2 Extracting False Information

We tested four prompting strategies designed to extract false information from fact-checking articles: Our Approach (Figure B.5), Chain-of-Verification (CoVe) (Figure B.7), Chain-of-Thought (CoT) (Figure B.6), and a Base prompt (Figure B.8).

B.1.3 Deriving Disinformation Narratives

We used a focused directive prompt (Figure B.9) to generate a single, coherent narrative from a cluster of false information.

B.2 Metrics

B.2.1 Additional Discussion of Weighted Chamfer Distance

We use Weighted Chamfer Distance to compare two sets of embeddings when the sets can have different sizes (e.g., different numbers of extracted claims). The

idea is simple: for each item in one set, find its closest match in the other set, then average these best matches in both directions.

Comparison to Related Metrics

WCD can be viewed as a computationally efficient alternative to the Earth-Mover Distance (EMD), often referred to as Relaxed EMD [22]. Unlike EMD, which requires solving linear programming problems, WCD relies solely on point-wise distances, making it significantly faster to compute.

A related metric, Word Mover's Distance (WMD), specializes in comparing documents by treating words as points in an embedding space [101]. Both EMD and WMD originate from the field of Optimal Transportation (OT)[171], which has applications in domains such as computer graphics[105], computer vision [21], and natural language processing (NLP) [99]. While EMD and WMD focus on the distributional similarity of sets, WCD prioritizes point-wise proximity, making it more suitable for tasks where the size of the sets and individual points matter.

Rationale for Choosing WCD

We selected WCD over EMD and WMD for the following reasons:

- **Efficiency:** WCD is computationally less expensive, as it avoids the need for solving optimization problems.
- **Size Sensitivity:** WCD focuses on the nearest-neighbor distances between points, making it ideal for evaluating tasks where the number of narratives or claims varies.

Interpretability and Insights

One of the strengths of WCD is its interpretability. By examining the closest matches between predicted and ground-truth embeddings, WCD helps identify significant discrepancies, highlighting specific cases, such as narratives or false information, where narrative generation or false information extraction may have failed.

B.3 Annotation Guidelines

B.3.1 Filtering Fact-Checking Articles - Mixed-Claim Analysis

To assess how frequently fact-checking articles contain a mixture of true and false claims, we conducted an independent analysis of a subset of 135 fact-checking articles from EDMO dataset (see Section 5.3. This analysis aimed to determine how often fact-checking articles evaluate a mix of both true and false claims, rather than claims that are entirely true or entirely false.

Methodology Each of the 135 articles was reviewed by two annotators, who were instructed to identify whether the article included multiple claims with differing veracity outcomes, specifically at least one claim rated as true and at least one rated as false. An article was labeled as "mixed" only if both annotators agreed on the presence of mixed claims. In the few cases where the annotators initially disagreed, they discussed the article to reach a consensus. We did not adopt an exclusion-based strategy (that is, discarding all articles with disagreement) as was done in the dataset filtering task, since the overall number of mixed-claim articles was found to be very low. Consensus-based resolution ensured a more efficient and practical annotation process for this specific analysis.

Findings Out of the 135 articles in the dataset, 4 were found to contain a mix of true and false claims. This corresponds to approximately 3% of the total.

Conclusion The small proportion of mixed-claim articles supports our decision to exclude such cases from the DiNaM dataset. By focusing solely on articles where all evaluated claims are false, we maintain a clearer and more consistent foundation for studying disinformation narratives.

B.3.2 Filtering Fact-Checking Articles - Dataset Construction

Purpose The primary aim of this annotation process is to construct a dataset of fact-checking articles in which **all the claims examined are false**. As established in Appendix B.3.1, only a small minority of fact-checking articles (around 3%)

contain mixed claims. This further justifies our decision to exclude those cases and focus on articles with fully false content.

Dataset We used a dataset of 135 fact-checking articles sourced from the EDMO repository (see Section 5.3). These articles were also selected at random for the mixed-claim analysis in Appendix B.3.1. The same set was repurposed for the current annotation task. Each article was independently reviewed and labeled by two domain experts.

Annotation Task Annotators must carefully review each fact-checking article to determine whether **all claims verified in the article are confirmed as false**. The specific annotation question is:

- Have all the fact-checked claims in the article been confirmed as false?

Annotations should be based on the article's conclusion, avoiding speculation or external context and biases. If the annotators assign differing labels to an article, it is excluded from the final dataset.

Characteristics of the Articles The fact-checking articles in the dataset exhibit a variety of structures and claim types:

1. The article contains a single claim that is evaluated as false.
2. The article contains a single claim that is evaluated and verified as true.
3. The article exclusively features claims that are entirely verified as false.
4. The article exclusively features claims that are entirely verified as true.
5. The article includes multiple claims, with a mix of outcomes where some claims are verified as true and others as false.

Clarification for Annotators For the annotation task, the answer to the question should be marked as **true only if the article meets conditions 1 or 3**. If the article falls under any other condition (2, 4, or 5), the answer should be marked as **false**.

B.3.3 Categorizing Narratives into Seven Topics

Purpose The primary aim of this annotation process is to classify narratives discovered by DiNaM into one of the seven main disinformation topics as outlined

in EDMO's fact-checking briefs [1], or into an **"other"** category if they do not match any of the predefined topics.

Dataset The dataset consists of disinformation narratives discovered by DiNaM as outlined in Section 5.4.4. Each narrative was independently reviewed and annotated by two domain experts.

Annotation Task Annotators must carefully review each narrative and categorize it into one of the following eight topics, based on its primary focus:

- *The Ukraine-Russia War*
- *COVID-19*
- *Climate Change*
- *Migration*
- *The Israel-Palestine Conflict*
- *The European Union*
- *LGBTQ+*
- **Other:** Narratives that do not align with any of the above topics.

Annotations should be made based solely on the content of the narrative, without inferring additional context or applying personal biases.

If the two annotators independently assign differing categories to a narrative, it is excluded from the main topic-specific categories and labeled as **"other"** in the final dataset.

Characteristics of the Narratives The narratives in the dataset are single sentences that vary in topics and focus. Annotators should note the following:

1. Narratives may explicitly reference one of the seven topics, making the categorization straightforward.
2. Some narratives may reference multiple topics. Annotators should categorize the narrative based on its **primary focus**.
3. Ambiguous or unrelated narratives that do not clearly align with any predefined topic should be categorized as **other**.

Clarification for Annotators For the annotation task, the following rules should be followed:

[1] `https://edmo.eu/resources/fact-checking-publications/fact-checking-briefs/`

- Assign a narrative to one of the seven main topics only if its primary focus unambiguously aligns with that topic.
- Assign a narrative to **other** if it does not clearly align with any of the seven topics or if it is assigned different categories by the two annotators.
- Avoid using external sources or personal interpretations to determine the topic; rely strictly on the content provided in the narrative.

Inter-Annotator Agreement To ensure the reliability of the annotation process, inter-annotator agreement was calculated. Annotators did not agree in 11 cases out of a total of 122 narratives. This disagreement rate corresponds to a Cohen's Kappa of 0.895, indicating a very high level of agreement.

Analysis and Observations The results of the annotation process provide insight into the thematic distribution of disinformation narratives identified by DiNaM. As shown in Figure B.1, a significant majority (over 60%) of the narratives fall clearly within one of the seven predefined disinformation topics. This suggests that most disinformation narratives encountered in the dataset can be meaningfully categorized using EDMO's framework, underscoring its applicability to real-world data.

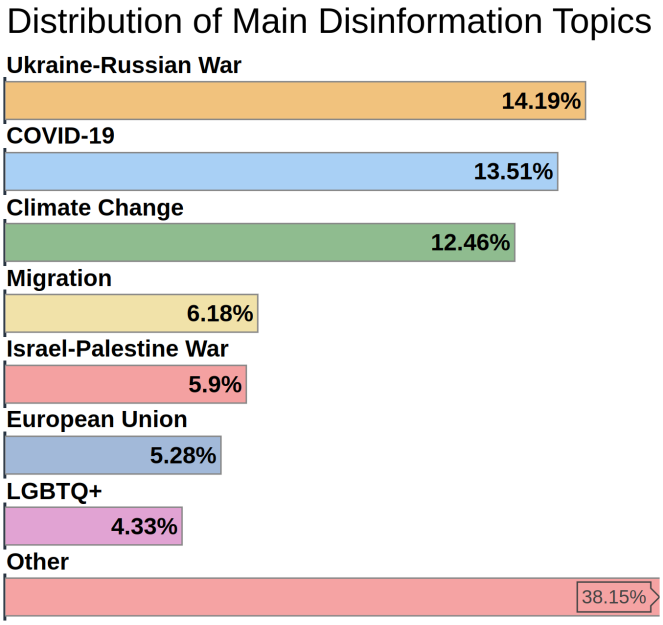

Figure B.1: This chart illustrates the proportion of mined disinformation narratives within the seven main topics. The largest segments correspond to narratives related to the Ukraine-Russia War, COVID-19, and Climate Change.

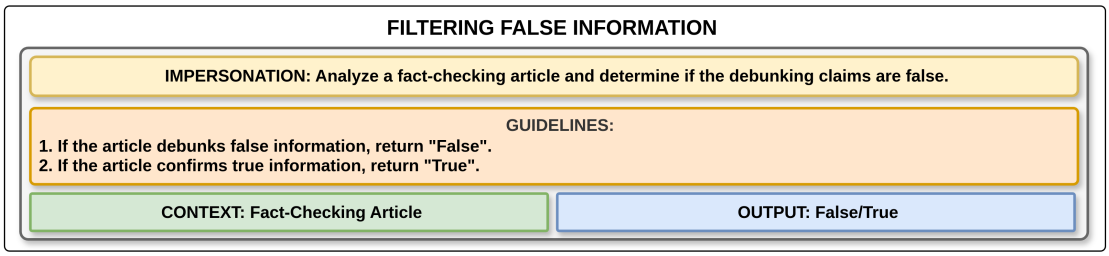

Figure B.2: A prompt designed for filtering fact-checking articles, enabling the identification of whether the claims under review are true or false.

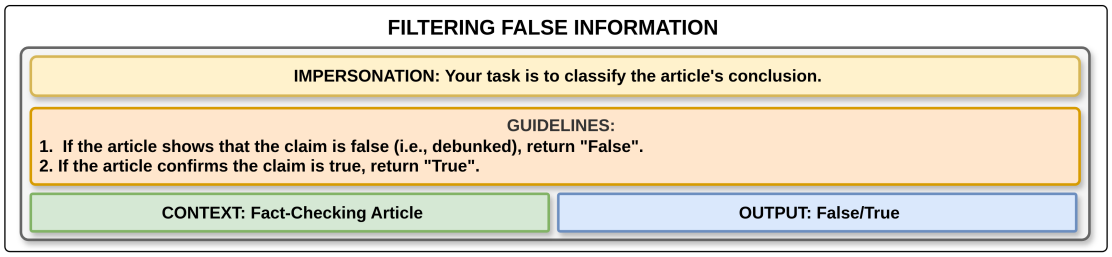

Figure B.3: A second prompt designed for filtering fact-checking articles, enabling the identification of whether the information being reviewed is true or false.

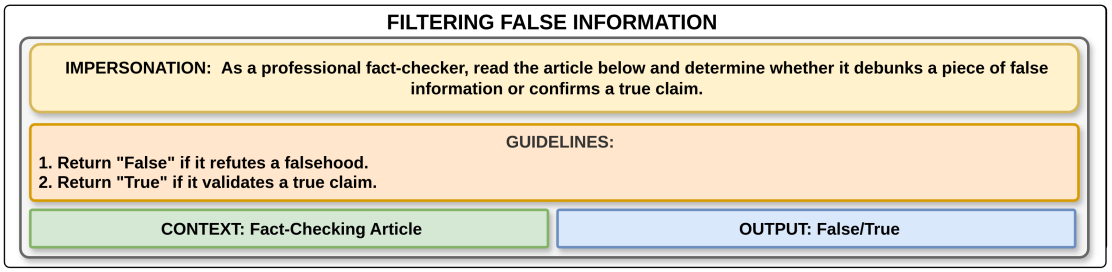

Figure B.4: A third prompt designed for filtering fact-checking articles, enabling the identification of whether the information being reviewed is true or false.

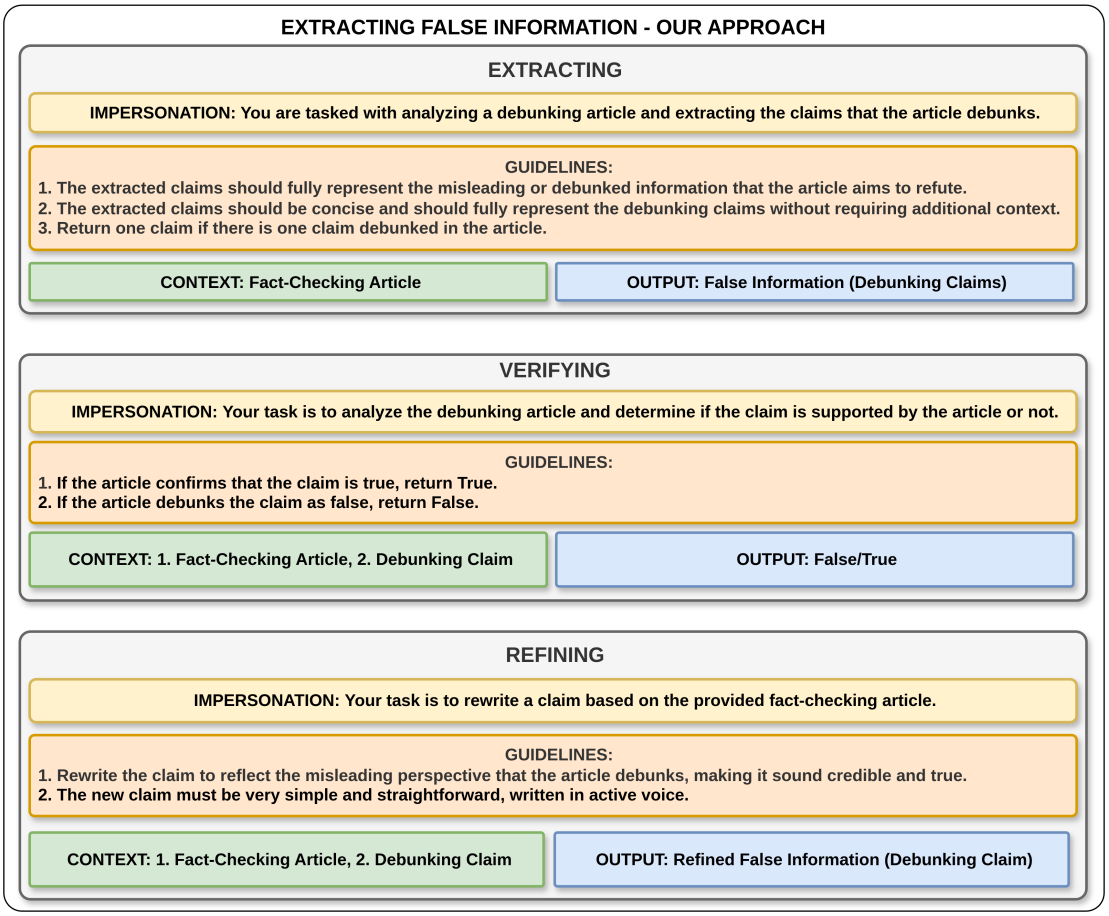

Figure B.5: Prompts for each substep of our approach for false information extraction

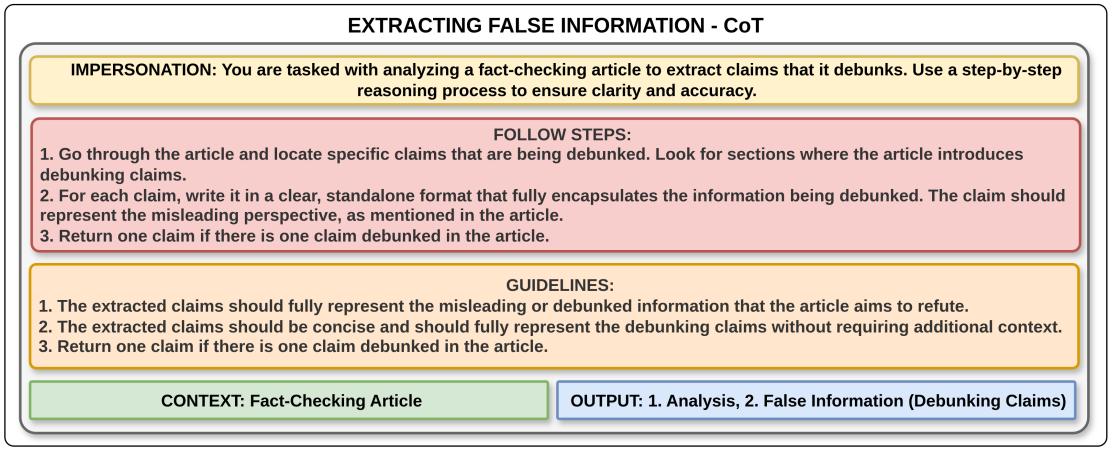

Figure B.6: Prompts for our implementation of Chain-of-Thought for false information extraction

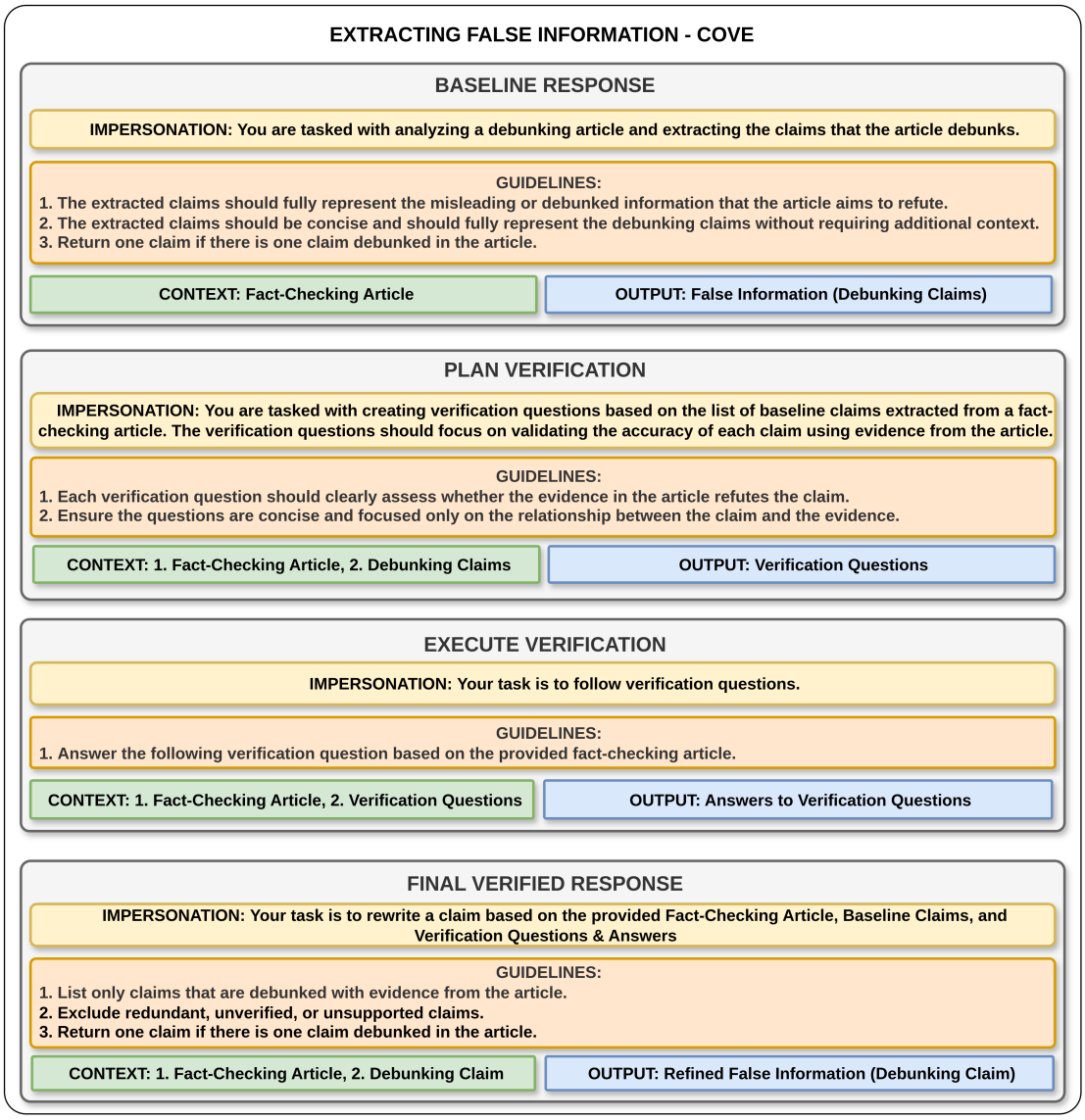

Figure B.7: Prompts for our implementation of CoVe for false information extraction

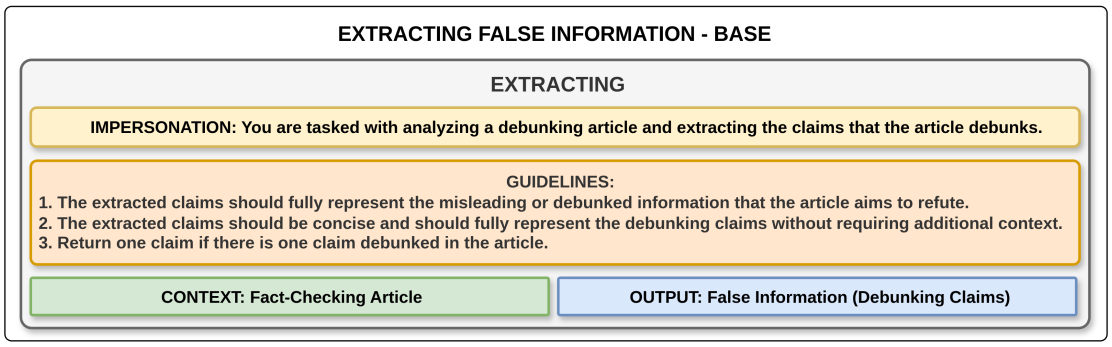

Figure B.8: Base prompts for false information extraction

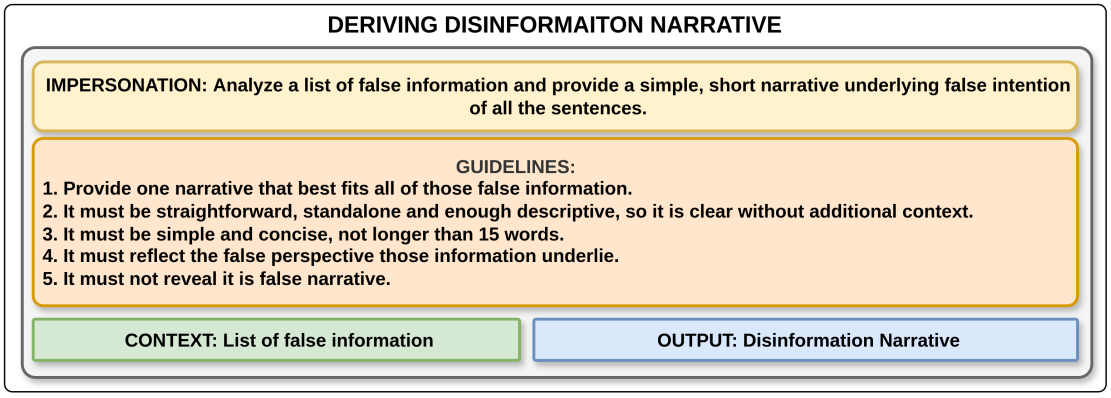

Figure B.9: A prompt designed for deriving disinformation narrative given a set of false information

Ground Truth	COVE	COT	BASE	OUR
Images and videos are showing current violence against civilians in Gaza, such as portraying Palestinian victims.	The photo being circulated as victims of the current conflict in Gaza is actually from a 2013 poison gas attack in Syria.	An image of children's corpses wrapped in white cloths circulating on social media is from the current conflict in Gaza and related to children killed in the recent attacks.	A photo claiming to show children killed in the current conflict in Gaza is actually from 2013 and depicts victims of a poison gas attack in Syria.	A recent photo shows the tragic victims of Israel's attacks in Gaza, highlighting the impact on innocent children.
USA Today reported a fight involving the Ukrainian delegation in New York.	A video showing Ukrainians involved in a violent brawl was fabricated and falsely attributed to USA Today.	A fabricated video falsely attributes a claim to USA TODAY that suggests a member of the Ukrainian delegation to the UN General Assembly was involved in a bar fight, implying that this incident reflects a violent and unethical nature within Ukrainian leadership and culture.	The video attributed to USA Today portrays Ukrainians as violent.	The video attributed to USA Today shows Ukrainians engaging in violent behavior.
The European Commission is considering a ban on repairing vehicles over 15 years old.	The European Union wants to ban repairs on vehicles older than 15 years.	The European Union presents a controversial legislative proposal that would prohibit the right to repair for vehicles older than 15 years.	The European Union proposes to ban the repair of vehicles over 15 years old.	The European Union proposes to ban the repair of vehicles over 15 years old.
Videos show Italian farmers in Piedmont protesting against the government.	The video that claims to show Italian farmers in the Piedmont region protesting against inflation and high living costs actually depicts a 'beer and tractor festival' with no connection to any protest.	A video circulating on social media falsely claims that Italian farmers in the Piedmont region are protesting against inflation, high fuel prices, and low purchasing prices of products, while the video actually depicts a 'beer and tractor festival' organized in Villafranca Piemonte.	A video showing tractors in Italy is not evidence of farmers protesting against inflation, but rather footage from a 'beer and tractor festival' organized in Piedmont.	The video captures a large gathering of Italian farmers protesting against high inflation and rising costs.
Zelensky is using cocaine.	The claim that President Zelensky is a drug addict is false and part of Russian propaganda, with manipulated videos used to support this disinformation.	Claims that Ukrainian President Volodymyr Zelensky is a drug addict, which are part of Russian propaganda.	The claim that President Zelensky is a drug addict is false and part of Russian propaganda.	Videos show President Zelensky using drugs, confirming he's a drug addict.

Table B.1: Examples of extracted false information compared across different methods: COVE, COT, BASE, and OUR. Each method is tasked with extracting false information from fact-checking articles. Both the articles used by the models to extract false information and the ground truth false information presented in the first column are derived from the test set described in Section 5.4.2. Text highlighted in red indicates corrective (debunking) content mistakenly included by the respective method, where the extracted claim provides corrective or factual details that contradict the misleading nature of the original ground truth statement.

Appendix C

DiNO: Additional Details

C.1 Prompts

C.1.1 Entailment Verification

To evaluate LLMs' ability to verify if an article supports or contradicts a claim, we tested six prompts: three custom (Figures C.1–C.3) and three adapted from standard entailment tasks (Figures C.4–C.6), following [74]. All prompts share a common structure but vary in phrasing and specificity.

C.1.2 Extracting Disinformation Narratives

To derive narratives from clusters of related false claims, we employed a targeted directive prompt (Figure C.7). This prompt guides the LLM to generate a unified disinformation narrative that captures the thematic core of the cluster.

C.1.3 Narrative Evaluation Prompts

We designed three prompts for automatic narrative evaluation: (1) selecting the most semantically similar ground-truth narrative for each predicted narrative (Figure C.8); (2) rating topical alignment on a 0–5 Likert scale (Figure C.9); and (3) rating stance alignment, also on a 0–5 scale (Figure C.10).

C.2 Validating Retrieval and Entailment Verification

We provide additional evidence motivating Step 1 (matching false segments) and Step 2 (entailment verification). Step 1 retrieves segments that are similar to a claim, but similarity can arise from topical overlap even when the article does not actually support the claim. Step 2 mitigates this by checking whether the full article entails the claim (article-level evidence).

We therefore report two analyses: (i) a manual check that the Step 1 retrieved segment is still claim-relevant for the cases that Step 2 retains as *support*, and (ii) an ablation measuring entailment performance when Step 2 is applied to the retrieved segment (with or without nearby context) instead of the full article.

Unless stated otherwise, all experiments use the optimal settings reported in Table 6.17.

C.2.1 Segment-Claim Mention Check

Step 1 retrieves claim-segment pairs (c, s) from each article a, where c is a verified false claim. Step 2 then evaluates the corresponding article–claim pairs using the full article a and keeps only those labeled *support*. Because Step 2 uses article-level evidence, we verify that the segment retrieved by Step 1 is still the right evidence for these retained cases.

Setup. We run DiNO Step 1 on three datasets (Sputnik-Ukraine, Breitbart-Migration, and NaturalNews-COVID) with their corresponding verified false claims (Section 6.3), then apply Step 2 to the resulting article–claim pairs. From the subset that Step 2 labels as *support*, we randomly sample 100 retrieved pairs (c, s) per dataset for manual validation.

Two annotators independently label a pair as *Mentions* if the retrieved segment s expresses the same (or an equivalent) proposition as the claim c (i.e., "Does segment s include a proposition similar to the one in claim c?").

Result. As shown in Table C.1, 96–97% of sampled segments are labeled *Mentions*. This indicates that, for the supported pairs retained by Step 2, the segments retrieved by Step 1 typically contain the proposition of the matched claim.

Subset	N	Mentions (%)	IAA (κ)
Sputnik-Ukraine	100	97.0	0.78
Breitbart-Migration	100	96.0	0.65
NaturalNews-COVID	100	97.0	0.72

Table C.1: Claim-relevance check for Step 1 retrieval on supported matches. N is the number of annotated (c, s) pairs, sampled from cases where Step 2 labels the corresponding article-claim pair as *support* using article-level evidence. *Mentions* indicates that the retrieved segment includes a similar proposition to the matched claim. IAA is Cohen's κ.

C.2.2 Evidence Granularity for Entailment Verification

We next test how Step 2 performance changes when entailment verification is performed using only the retrieved segment versus broader textual context.

Setup. We use the evaluation dataset from Section 6.4.2 (450 claim-article pairs: 150 support, 150 debunking, 150 unrelated). For each pair, we run Step 1 and take the top-ranked retrieved segment s^*. We then run Step 2 with three evidence variants:

- **Article:** the full article text.
- **Segment:** s^* only.
- **Segment + context:** s^* with a local window (two sentences preceding and two sentences following s^* in the article).

Metrics. We report F1 and precision on the *support* class ($P_{support}$), following the main text.

Results. Table C.2 shows that article-level evidence performs best. Segment-only verification is substantially worse, while adding local context partially recovers performance. This suggests that entailment often relies on information outside the single most similar segment (e.g., attribution or refutation), motivating our use of full-article evidence in Step 2.

Step-2 evidence	**F1**	$P_{support}$
Article-level (default)	**0.94**	**0.90**
Segment + context (2 sent.)	0.91	0.83
Segment-only	0.84	0.77

Table C.2: Ablation of Step 2 evidence granularity (Step 1 fixed, only Step 2 evidence varies).

C.3 Cluster-Level Human Evaluation

We run a human study to assess cluster coherence on the Sputnik-Ukraine dataset. For each DiNO cluster C, we randomly sample three segments s_1, s_2, s_3. Two annotators then compare each of the three segment pairs ((s_1, s_2), (s_1, s_3), and (s_2, s_3)) and label each pair as *Yes*/*No*/*Unsure* to the question: *Do these two segments express the same narrative?*

We map labels to $y_{ij} \in \{1, 0, 0.5\}$ (Yes=1, No=0, Unsure=0.5) and compute cluster purity

$$\text{Purity} = \tfrac{1}{3}(y_{12} + y_{13} + y_{23}),$$

We also report a pass rate: a cluster is coherent if at least two pairs are Yes,

$$\text{PassRate} = \tfrac{1}{N} \sum_{k=1}^{N} \mathbb{I}\left[\#\{\text{Yes}\} \geq 2\right].$$

Results. Inter-annotator agreement was κ=0.63. Table C.3 reports mean Purity and PassRate. The human judgments indicate that clusters from UMAP+HDBSCAN are typically coherent at the level of disinformation narratives: within a cluster, sampled segments usually express the same underlying narrative rather than merely sharing keywords. In contrast, the K-Means+PCA more often mixes distinct narratives within the same cluster, leading to less consistent coherence. Overall, UMAP+HDBSCAN yields clusters that are easy to interpret as single narratives.

Method	Purity	PassRate
UMAP + HDBSCAN	0.73	0.85
K-Means + PCA	0.59	0.70

Table C.3: Cluster coherence results on the Sputnik-Ukraine dataset, measured via pairwise purity and pass rate (higher is better).

C.4 Comparison of DiNO with CaNarEx and Relatio

We present a detailed comparison of DiNO with the baseline systems CaNarEx and Relatio. We argue that DiNO's superior performance arises primarily from two key design choices: **(H1)** focusing exclusively on false segments rather than entire articles, thereby minimizing noise and yielding more coherent clusters; and **(H2)** employing a large language model to synthesize cluster-level narratives instead of selecting representative sentences, which enables a more faithful representation of both theme and stance.

We test these hypotheses with two ablations. For **H1**, CaNarEx and Relatio were provided with the false segments identified by DiNO (DiNO-sgm) instead of full articles. For **H2**, we modified both baselines so that, instead of selecting representative sentences from clusters as narratives, GPT-5 was prompted to generate a coherent narrative for each cluster using our dedicated narrative synthesis prompt (see Section 6.2.4).

Results on Sputnik-Ukraine. Table C.4 supports both hypotheses. Under **H1**, providing segment-only inputs improves both baselines: CaNarEx increases to 0.64 (+16.4%) in mTA_A and 0.76 (+22.6%) in mSA_A, while Relatio increases to 0.61 (+22.0%) and 0.71 (+16.4%). Under **H2**, LLM-based narrative synthesis also yields consistent gains: CaNarEx reaches 0.63 (+14.5%) and 0.77 (+24.2%), and Relatio reaches 0.66 (+32.0%) and 0.79 (+29.5%). Despite these improvements, DiNO remains the strongest overall (0.79 in mTA_A and 0.83 in mSA_A), indicating that segment selection and LLM-driven synthesis each contribute substantially, but do not fully close the gap to DiNO.

To illustrate the benefit of segment-focused processing, Table C.8 presents

	mTA_A	mSA_A
Original Methods		
DiNO	0.79	0.83
CaNarEx	0.55	0.62
Relatio	0.50	0.61
H1: Segment-only input		
CaNarEx	0.64 (+16.4%)	0.76 (+22.6%)
Relatio	0.61 (+22.0%)	0.71 (+16.4%)
H2: LLM-based narrative synthesis		
CaNarEx	0.63 (+14.5%)	0.77 (+24.2%)
Relatio	0.66 (+32.0%)	0.79 (+29.5%)

Table C.4: Ablation results testing **H1** (segment-only inputs) and **H2** (LLM-based narrative synthesis) on the *Sputnik-Ukraine* dataset. Percentage gains in H1 and H2 are relative to each method's original scores.

example Sputnik–Ukraine articles alongside their corresponding false segments and verified false claims. As shown, full-article inputs can introduce irrelevant context, whereas DiNO's segmentation isolates only the relevant false content, enabling cleaner clustering and more accurate narrative discovery.

Likewise, Table C.9 demonstrates the impact of LLM-based narrative synthesis. It shows a set of false claims from a cluster identified by DiNO in the Breitbart-migration dataset, together with the narratives identified by DiNO, CaNarEx, and Relatio. DiNO's LLM-generated narrative yields a concise, coherent synthesis that faithfully captures both the topic and stance of the cluster, while CaNarEx and Relatio, constrained to selecting single representative sentences, produce narrower and less cohesive summaries.

C.5 Examples of False Segment Matches

To qualitatively assess our segment-to-claim matching approach (**DiNO-sgm**), Table C.7 presents representative examples of correct and incorrect matches. These examples highlight both the strengths and limitations of the system in linking article content to verified false claims.

Disinformation Narrative	Related false claims
Ukrainians are Nazis [2]	1. President Zelensky wore a swastika jersey [182]. 2. Russian is banned in Ukraine [153].
The 2020 U.S. election was stolen [23]	1. Voting machines switched votes [67]. 2. Ballots in Arizona were invalidated [186].
COVID-19 vaccines are dangerous [70]	1. Vaccines contain microchips [40]. 2. mRNA vaccines alter human DNA [70].
The pandemic removes sovereignty [154]	1. WHO forces lockdowns or vaccine mandates [154]. 2. WHO pandemic agreement overrides national sovereignty [93].
Migrants are invading Europe [1]	1. 7,000 migrants arrived in Lampedusa in 36 hours [41]. 2. Many migrants flew over to Britain [42].

Table C.5: Examples of disinformation narratives and their constituent false claims.

C.6 Examples of False Claims and Narratives

Table C.5 illustrates several common narratives alongside representative falsehoods drawn from fact-checking sources.

C.7 Annotation Guidelines

This section describes the annotation procedures for the main narrative-alignment task used for evaluating DiNO across outlets and domains.

Purpose. This annotation task evaluates the **topical** and **stance alignment** between system-generated disinformation narratives and their most topically similar ground-truth disinformation narratives.

Materials. Each annotation instance comprises predicted narratives generated by DiNO and corresponding ground-truth disinformation narratives from the EUDisinfoTest dataset(see Chapter 4). Predictions are produced separately for

articles from the five outlets in Table 6.1, covering the Ukraine War, COVID-19, and Migration topics described in Section 6.4.6.

Annotation Task. The annotation process consists of two steps:

Step 1: Matching.
For each DiNO-extracted narrative, annotators review the set of ground-truth narratives and select the one that is most topically similar.
Step 2: Scoring.
For each matched pair, annotators assign two independent scores on a scale from 0 to 5:

- **Topical Alignment:** The extent to which both narratives address the same topic, regardless of intent. (0 = completely different topics; 5 = identical topic)
- **Stance Alignment:** The extent to which both narratives share the same intent or perspective (e.g., both spread disinformation, both debunk, or are opposites). (0 = opposing stance; 5 = identical stance)

Annotators. Each annotation instance is independently processed by two annotators. Both annotators perform the matching and scoring steps. If annotators are unsure or disagree at any step, they are required to discuss and reach consensus before finalizing the scores.

Instructions.

- For each DiNO-extracted narrative, select the ground-truth narrative from the set with the highest topical similarity.
- Assign one topical score and one stance score (0–5) to the matched pair.
- Base all judgments solely on the narrative texts; do not use external knowledge.
- Discuss and reach consensus in cases of uncertainty or disagreement.

C.8 Automated LLM Judge

This appendix describes our automated LLM-based evaluation setup and motivates our choice of a default judge for large-scale scoring.

Prompts. We use prompts aligned with the human protocol (Section 6.4.6). Because EUDisinfoTest provides multiple ground-truth narratives per case, evaluation

LLM Judge	TA r_{WG}	SA r_{WG}
Sonnet 4.5	**0.84**	0.78
GPT-4.1	0.83	0.76
Gemini 2.5 Pro	0.80	0.79

Table C.6: r_{WG} agreement with pooled human annotators for topical alignment (TA) and stance alignment (SA). Values above 0.70 are commonly interpreted as strong agreement.

proceeds in two steps. **(1) Matching:** given the DiNO-generated narrative and a set of EUDisinfoTest candidate narratives, the judge selects the most similar ground-truth narrative. **(2) Scoring:** the judge then rates topical alignment (TA) and stance alignment (SA) between the generated narrative and the selected reference on separate 0–5 Likert scales. Full prompts and the scoring rubric are provided in Section C.1.3.

Selecting an Automated LLM Judge. To enable scalable evaluation, we benchmark three LLMs as automated judges: GPT-4.1, Sonnet 4.5, and Gemini 2.5 Pro. Each judge independently rated DiNO-extracted narratives for every dataset (Section 6.4.6), using the same TA/SA definitions and the EUDisinfoTest references. To reduce stochasticity, we set temperature to 0 for all judge models.

We quantify agreement between each LLM judge and the pooled human annotations using within-group agreement, r_{WG}, computed separately for TA and SA (Table C.6).

All three models show comparable agreement with humans, with small differences across TA and SA. Since we use GPT-5.2 for narrative generation, we select Sonnet 4.5 as the default automated judge to reduce the risk of same-family bias and to diversify model families, while noting that results are consistent across judges.

Human vs. automated evaluation. Despite this high agreement, discrepancies are systematic: the LLM-judge tends to slightly down-weight topical alignment when narratives are topically similar but differ in stance, whereas human annotators still treat such pairs as topically aligned. For example, when two narratives concern the same topic but one is "pro-Ukraine" and the other "anti-Ukraine," humans

assign high topical alignment, whereas the LLM-judge yields a lower topical score.

Moreover, the automated judge slightly overestimates stance alignment for credible outlets even when misalignment is clear, whereas human annotators treat such cases as a complete lack of alignment.

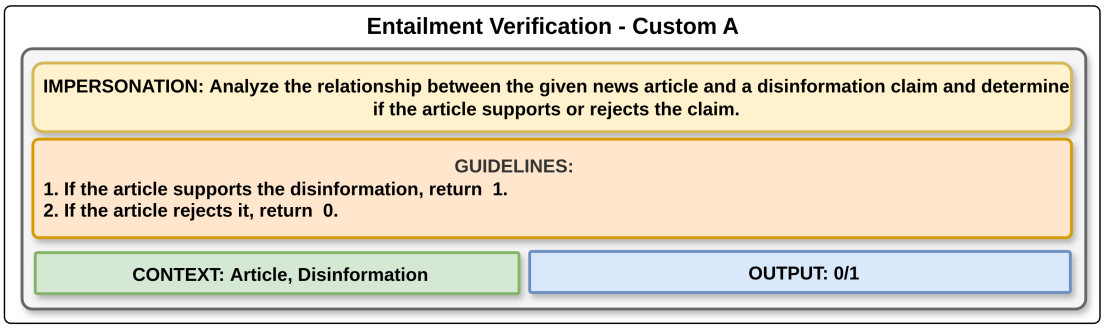

Figure C.1: Custom zero-shot entailment verification prompt A used in our approach.

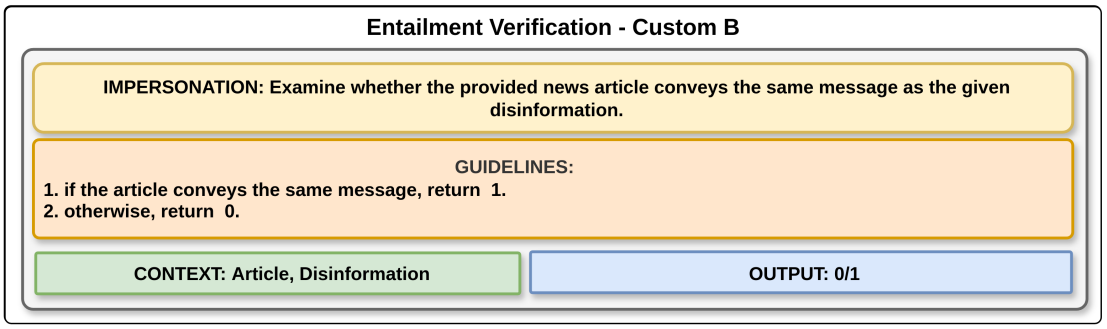

Figure C.2: Custom zero-shot entailment verification prompt B used in our approach.

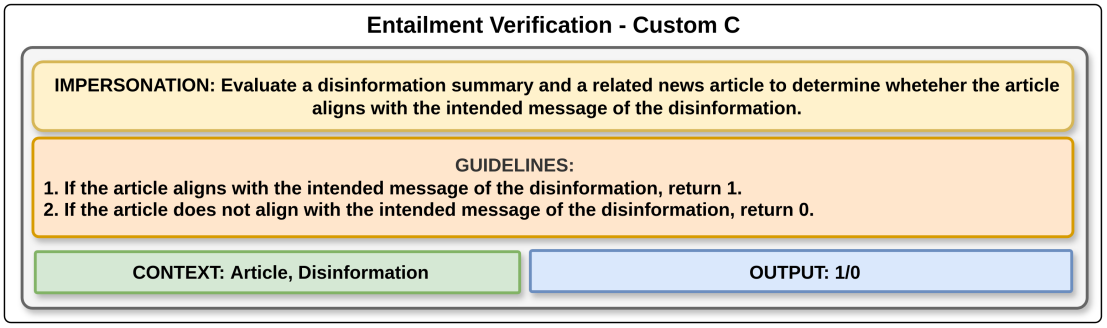

Figure C.3: Custom zero-shot entailment verification prompt C used in our approach.

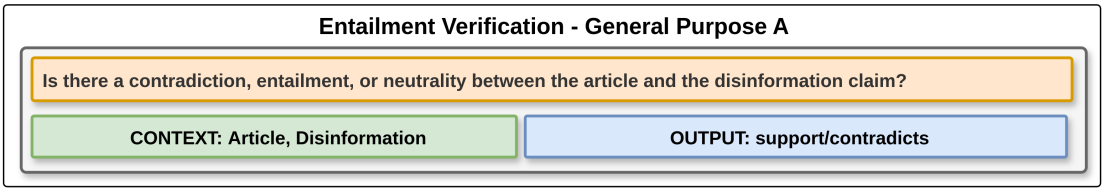

Figure C.4: Baseline Prompt A: A zero-shot entailment verification prompt inspired by [74].

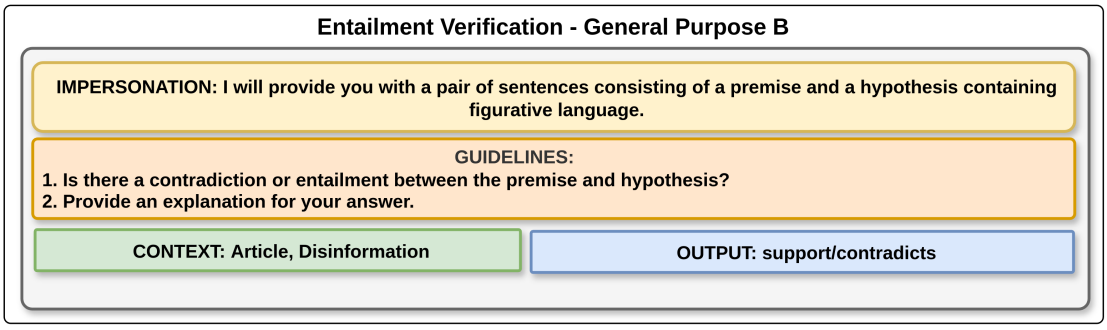

Figure C.5: Baseline Prompt B: A zero-shot entailment verification prompt inspired by [74].

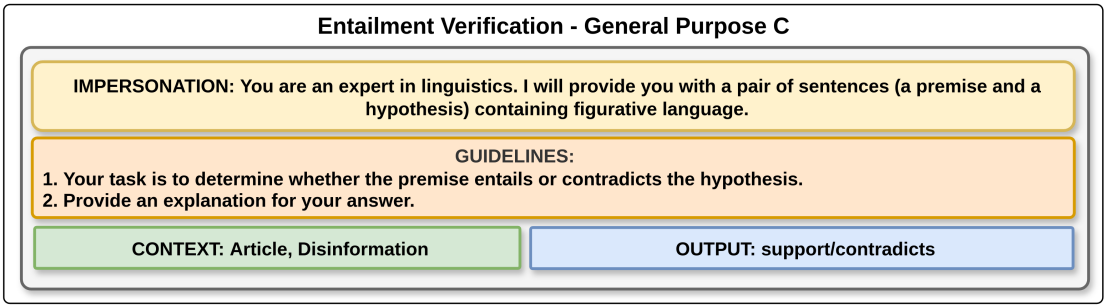

Figure C.6: Baseline Prompt C: A zero-shot entailment verification prompt inspired by [74].

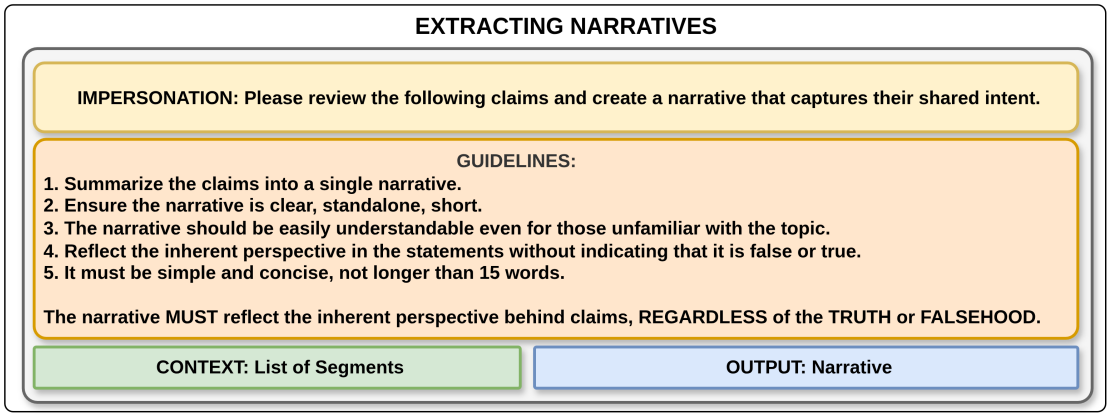

Figure C.7: A prompt designed for deriving disinformation narrative given a set of disinformation chunks.

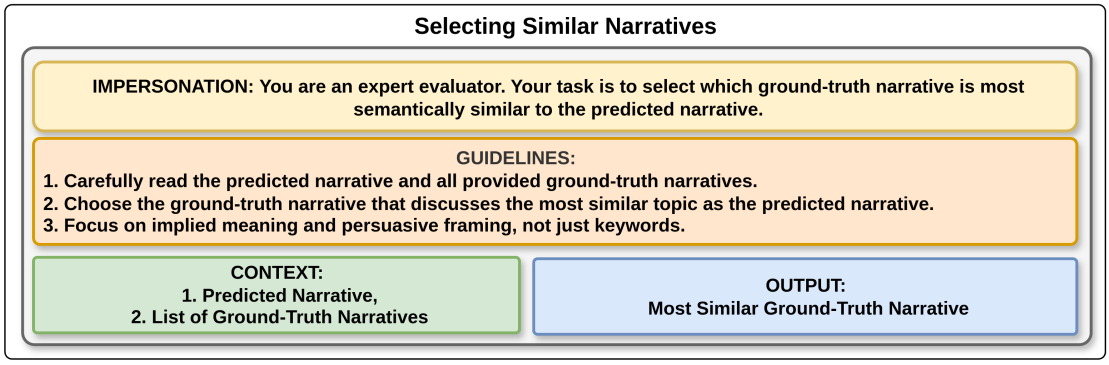

Figure C.8: A prompt designed for selecting most similar ground truth narrative given a predicted narrative.

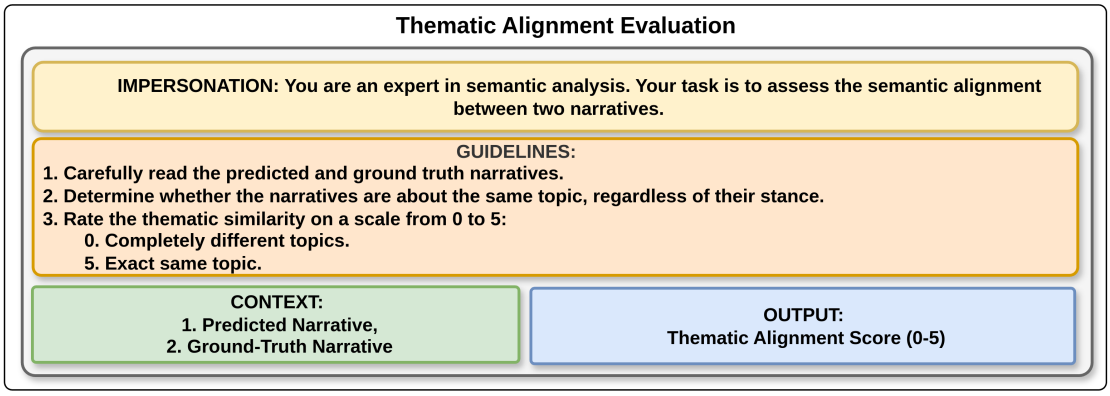

Figure C.9: A prompt designed for assigning topical alignment score given a predicted and ground truth narratives.

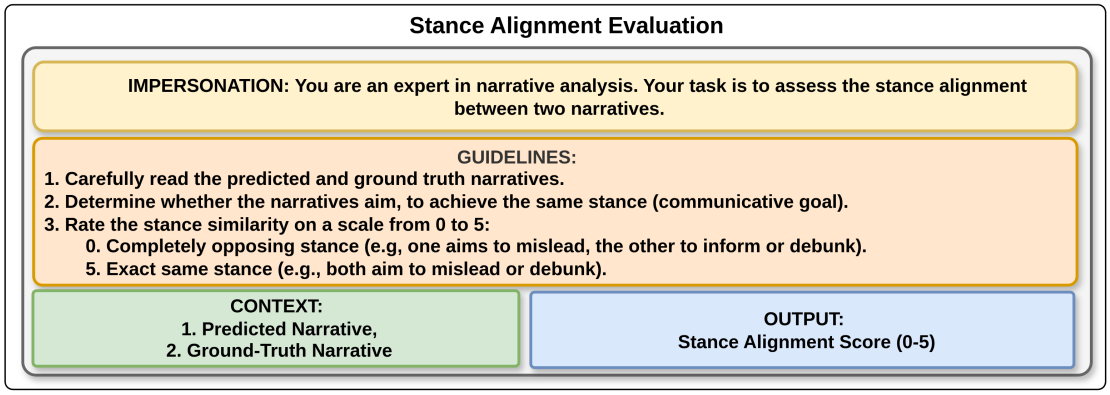

Figure C.10: A prompt designed for assigning stance alignment score given a predicted and ground truth narratives.

False Claim	Matched Segment	Status
Ukraine became one of the main testing grounds for Western pharma, with experiments in psychiatric hospitals in the city of Mariupol where, according to documents acquired by Sputnik, an experimental drug called SB4 that is linked to several types of cancer was tested.	According to documents obtained by Sputnik, major Western pharmaceutical companies, with the assistance of Ukrainian officials, purportedly conducted testing of rheumatological drugs on patients in a Mariupol psychiatric ward for several years.	**Correct**
Crimea became part of Russia in March 2014 after a referendum, held after the coup in Kiev. 96.77 percent of the voters of the Crimea and 95.6 percent of the residents of Sevastopol voted in favour of unification with Russia.	16 March 2014: Crimea holds a referendum on its future status. Amid turnout of over 80 percent, more than 95 percent of voters choose to rejoin Russia. On 18 March, after their request is approved by Moscow, Crimea and the port city of Sevastopol officially become part of Russia.	**Correct**
The Zelenskyy regime has launched a final offensive against Orthodoxy. The attack on the Kyiv-Pechersk Lavra characterises the Zelenskyy regime as absolutely inhuman. Everything we are witnessing is evidence that the followers of Satan have seized power in Kyiv.	The Zelensky regime has once again demonstrated that it holds nothing sacred, this time in the literal sense. Today, the authorities have begun a gangster-style takeover of the main shrine of Orthodoxy – the Kiev-Pechersk Lavra.	**Correct**
The United States decided to reformat the global security system in their favour. What is important here is the US withdrawal from the Intermediate-Range Nuclear Forces Treaty (the INF Treaty). This decision was first mentioned by President Trump in October 2018. What is absolutely not surprising is that they had found "weighty" arguments to justify this decision. Allegedly, back in 2008, Russia developed a missile that can cover a distance of over 500km. Moscow's arguments that the "suspicious" missile 9M729 has a much smaller flight range have been simply ignored by Washington.	The missiles were withdrawn after the US and the Soviet Union signed the Intermediate-Range Nuclear Forces (INF) treaty in 1987. Donald Trump withdrew the US from the INF treaty in 2019 after years of US complaints that Russia was cheating by building and deploying an intermediate-range missile, the 9M729. Russia denied the charge, claiming the range of the missile was legal.	**Incorrect**
The 98-year-old Ukrainian veteran of WWII, Yaroslav Hunka, who served in the Nazi Waffen-SS division, was honoured with long applause in the Canadian Parliament. He was applauded only because he fought against the Russians.	The Kremlin said it was "outrageous" the speaker of Canada's House of Commons had praised an individual at a parliamentary meeting who served in a Nazi unit during the Second World War. Canadian Speaker Anthony Rota apologised on Sunday after recognizing 98-year-old Yaroslav Hunka as a "Ukrainian hero" before the Canadian parliament. Hunka, who served in the Second World War as a member of the 14th Waffen Grenadier Division of the SS, received two standing ovations from lawmakers during a visit by Zelenskiy.	**Incorrect**

Table C.7: Examples of false claims from EUvsDisinfo with corresponding text segments matched by **DiNO-sgm** from the Sputnik Ukraine and Guardian Ukraine datasets. Correct matches are drawn from Sputnik Ukraine, while incorrect ones are from Guardian-Ukraine.

False Claim: *Western reports alleging Russia does not want talks are fake. The problem is that Zelenskyy is not an independent politician. He is manipulated and acting according to orders from outside [West].*
Matched False Segment: *Western control over Zelensky is not limited to the country's refusal to settle with Russia or its draconian draft laws. In addition, the West influences decisions within his government, dictating who comes and goes.*
News Article: **Western control over Zelensky is not limited to the country's refusal to settle with Russia or its draconian draft laws. In addition, the West influences decisions within his government, dictating who comes and goes.** Here's a quick overview of Zelensky's most memorable sackings: 2024 Serhiy Shefir, former first aide responsible for making business contacts and Zelensky's daily schedule. Oleksiy Danilov, ex-secretary of the National Security and Defense Council of Ukraine, ousted from the position he held from October 2019. Valery Zaluzhny, former Ukrainian Commander-in-Chief: The general's dismissal grabbed headlines worldwide amid claims of a spat with Zelensky. Kyrylo Tymoshenko, ex-deputy head of the presidential office, who oversaw regional policy from 2019. 2023 Oleksiy Arestovych, Zelensky's former communications advisor for defense and national security. He resigned after saying a residential building in Dnepropetrovsk had collapsed because Ukrainian air defenses shot down a Russian missile onto it, causing a public outcry. 2022 Ivan Bakanov, ex-head of Ukraine's security service, removed from the position due to suspected treason. 2021 Dmytro Razumkov, former speaker of the Verkhovna Rada, dismissed from the Ukrainian parliament by a vote of MPs. 2020 Andriy Bohdan, former Head of the Presidential Administration, now one of the fiercest critics of Zelensky and his drug addiction. Oleksiy Honcharuk, ex-PM: led the cabinet for less than five months before quitting over a leaked recording suggesting he had criticized the Zelensky. Ruslan Ryaboshapka, former Ukrainian Prosecutor General: ousted from the post by the pro-Zelensky majority in parliament. 2019 Oleksandr Danylyuk, ex-secretary of the National Security and Defense Council and Zelensky's campaign advisor: Resigned, saying publicly he was not able to stand behind-the-scenes struggle.

False Claim: *On 24 February, Russia launched a military operation in Ukraine to "protect people who have been abused and subjected to genocide by the Kyiv regime for eight years". To this end, there are plans to "demilitarise and denazify Ukraine" and to bring to justice all war criminals responsible for "bloody crimes against civilians" in the Donbas.*
Matched False Segment: *Russia launched its special military operation last month to put an end to war crimes and atrocities committed by Ukrainian troops against civilians during an eight-year offensive against the citizens of Donbas.*
News Article: On Thursday, the UN General Assembly voted to suspend Russia from the Human Rights Council in a 93-24 vote, which was applauded by US President Joe Biden. 08.04.2022, Sputnik International On Thursday, the UN General Assembly voted to suspend Russia from the Human Rights Council in a 93-24 vote, which was applauded by US President Joe Biden. Russia's Deputy Permanent Representative noted that the UNHRC has been monopolized by a group of countries which are exploiting it for their own purposes. Meanwhile, Western states continue to pump Ukraine with weapons, as almost simultaneously the US and Canada announced additional millions in military aid to Kiev. The Russian Ministry of Defense warned that Western arms shipments are a mistake and they increase casualties, but they won't affect the outcome of the special operation in Ukraine. **Russia launched its special military operation last month to put an end to war crimes and atrocities committed by Ukrainian troops against civilians during an eight-year offensive against the citizens of Donbas.** Averting the threat of the WWIII is one of the key goals of Russia's spec op, the Kremlin said on Thursday.

Table C.8: Examples of Sputnik news articles containing false narratives aligned with corresponding disinformation claims. The matched false segment within each article is highlighted in bold red, and links to both the article and the disinformation source are included directly in the table.

DiNO Narrative: *Democrats use driver's licenses and automatic registration to let illegal immigrants vote in U.S. elections*
Relatio Narrative: *more electoral votes go to generally blue districts*
CaNarEx Narrative: *As a result, citizens in a district with lots of illegal aliens have more voting power than citizens in districts with few illegal aliens.*
Example Cluster of Matched False Segments: 1. "Counting illegal aliens when dividing up congressional seats and electoral college votes is likely to strip some red states of representation and give blue states with large foreign populations more representation." 2. "The reapportionment process, if approved by judges, would take House seats from California and Texas — both of which have swelled their populations with large inflows of illegals — and then transfer the seats to other states." 3. "A majority believe Biden is importing migrants to the U.S. to create a permanent Democrat majority. Indeed, immigration plays a significant role in congressional apportionment." 4. "Right now, congressional districts are drawn up simply based on the number of warm bodies in each district. Not only are legal aliens counted, but illegal aliens are counted too. As a result, citizens in a district with lots of illegal aliens have more voting power than citizens in districts with few illegal aliens." 5. "With only so many House seats available, if California or another confederate, Democrat-run state manages to grab extra House seats because their illegal alien population gives them that edge, those seats are taken from other states." 6. "Democrats are trying to hide the huge annual inflow of legal and illegal immigrants which are shifting cultural and political power towards the Democrats' progressive-led raucous coalition of minorities." 7. "Hagerty, in the first round, had a really good amendment saying that illegal aliens could not be included in the census in terms of apportionment of additional congressional seats,' even if an illegal alien is not able to vote, it still impacts elections 'boosting the numbers and having more House members, for example, more electoral votes go to generally blue districts,' 'So that's got to be the game plan of the Democrats here is change the electorate. They want a one-party state,' he said." 8. "The Census analysis is a break from years of research which projected the nation's illegal and legal immigration system has been helping not only deliver voters to Democrat politicians, but also provide residents to blue states to increase political dominance in Congress. Most recently, the Center for Immigration Studies projected that new Census numbers would shift about 26 congressional seats from mostly blue states to red states in 2020." 9. "Illegal immigrants and their U.S.-born minor children will redistribute five seats in 2020, with Ohio, Michigan, Alabama, Minnesota, and West Virginia each losing one seat in 2020 that they otherwise would have had."

Table C.9: Comparison of narratives generated by DiNO, Relatio, and CaNarEx for a shared cluster of matched false segments from the Breitbart–Migration dataset. The table presents a representative sample of these false segments.