Examiner's Independent Report
Candidate: Arkadiusz Modzelewski
Doctoral Dissertation Title: Leveraging Persuasion and Intent for Analysis and Reasoning-based Detection of Disinformation with Large Language Models
Examiner: Prof. Andreas Vlachos
Date of Report: 05/06/2026

Report

This thesis presents substantial progress in the field of analysing the potential of large language models in dealing with disinformation. The research objectives stated in section 1.2 focusing on creating novel resources and developing novel methods are worthwhile.

Chapter 1 presents the overall context and goals of the thesis.

Chapter 2 gives the background and definitions for concepts that the thesis builds upon. While it gives definitions to the terminology used in later chapters, I believe that a fundamental concept missing here is trust. A definition of trust that would be worth including here is due to Mayer et al. (1995), who decomposed it into three components: ability (whether an agent is able to do something, e.g. give us correct information), benevolence (whether an agent has our best interests at heart, corresponding to intent in the thesis) and integrity (whether an agent follows their principles). Furthermore, how can a piece of text earn credibility? The source of the text can earn credibility over a period of time, but the credibility of a text itself is fixed. Furthermore, in the context of persuasion, how is an overly simple explanation distinguished from an explanation that is as simple as it needs to be? Finally, if intent to generate profit is sufficient to characterise misinformation as disinformation, would this consider all advertisements as such?

Chapter 3 focuses on the background on NLP applied to disinformation and persuasion. However, in section 3.1, the datasets mentioned are about misinformation more broadly, not disinformation. E.g. Liar is derived from politifact which fact-checks even claims from the Onion, which a satirical website. It would be better to focus on disinformation datasets, or if such datasets do not exist, make clear what how the ones discussed relate to disinformation explicitly. Similarly, in section 3.2, how are the argumentation datasets being discussed related to persuasion? Some further work worth citing on persuasion by LLMs includes the one by Hackenburg et al. (2025). Section 3.3 repeats from non-NLP content from chapter 2 (second paragraph), and Section 3.4 refers to a second cluster of research gaps without having mentioned the first one.

Chapter 4 focuses on Polish disinformation detection. The dataset is interesting. I was wondering, how did you quantify bias? This would help appreciate your effort to minimise it. And you need to state the inter-annotator agreement between the independent annotations. If humans don't agree with it, how can we be sure we have a well-established task. It would be interesting to see how manipulation/intent detection transfer across topics, e.g. train on data from some topics and test on others. And it would have been interesting to see if in-context learning examples would help the decoder models. Also, why not have results for the LLMs in sections 4.4.2 and 4.4.3? Some minor points: Polish is mentioned to be the largest of the V4 countries, but in which sense? Number of speakers? And are the hard-to-say articles publicly available? The small language models would better be characterised as encoder models, as size characterisations change over time.

Chapter 5 focuses on persuasion-augmented reasoning or disinformation detection. Again the inter-annotator agreement needs to be reported among independent annotators, before discussion. How were the extra 400 articles post-2014 mentioned in section 5.1.3 labeled? The results are positive, but I was wondering if they are consistent across LLMs, e.g. those in Table 5.9. And the results of table 5.24 should be contextualised by stating which other table they can be compared against. Some minor points. Avoid repeating the EC-HLEG definitions in every chapter. Politifact checks claims, not whole articles; how is the ISOT fake news dataset used?

Chapter 6 focuses on intent, with the main contribution being the intent-based inoculation approach. I was wondering, how is this different from a combination of in-context learning with prompt engineering? Inoculation suggests that something changes in the model/human, but this doesn't seem to be the case here. And credible, correct, information can be given with malicious intent, this is the most common characterisation of malinformation. And the following are mentioned as malicious intents: Undermining the Credibility of Public Institutions, Changing Political Views, Undermining International Organizations and Alliances. However not sure that they are by default malicious. Especially changing political views can be benevolent.

Chapter 7 focuses on detecting persuasion by LLMs. I was wondering, why do we expect it to be different from human? Does the author believe that we are persuaded by LLMs different than we would be persuaded by humans? What is the evidence? As far as I am aware, studies such as the one by Costello et al. (2025) find that what matters are the facts and the arguments rather the persuasion techniques. And do we know how persuasive are the texts generated? Just using some technique, doesn't entail that a human would be persuaded. It could be using the technique but in a poor way. And I was confused by the labels used to generate the text; are they relative or absolute? The human texts seem to be absolute (persuasive or not), but the AI ones are relative (increasing or decreasing persuasion). It could be that decreasing persuasion nullifies it entirely.

Chapter 8 summarises the thesis.

Overall, the thesis is on the whole well-written, with a good flow of language. The work is substantial in both quantity and quality, published in good venues, and worthy of a PhD thesis, and thus my conclusion is positive.

References:

- Mayer et al. (1995): https://journals.aom.org/doi/abs/10.5465/AMR.1995.9508080335
- Hackenburg et al. (2025): https://www.science.org/doi/10.1126/science.aea3884
- Costello et al. (2025): https://osf.io/preprints/psyarxiv/h7n8u_v2