EU DisinfoTest: a Benchmark for Evaluating Language Models' Ability to Detect Disinformation Narratives

Witold Sosnowski[1], Arkadiusz Modzelewski[1], Kinga Skorupska[1], Jahna Otterbacher[2], Adam Wierzbicki[1]

[1]Polish-Japanese Academy of Information Technology
[2]Open University of Cyprus

Correspondence: witold.sosnowski.pw@gmail.com

Abstract

As narratives shape public opinion and influence societal actions, distinguishing between truthful and misleading narratives has become a significant challenge. To address this, we introduce the EU DisinfoTest, a novel benchmark designed to evaluate the efficacy of Language Models in identifying disinformation narratives. Developed through a Human-in-the-Loop methodology and grounded in research from EU DisinfoLab, the EU DisinfoTest comprises more than 1,300 narratives. Our benchmark includes persuasive elements under Logos, Pathos, and Ethos rhetorical dimensions. We assessed state-of-the-art LLMs, including the newly released GPT-4o, on their capability to perform zero-shot classification of disinformation narratives versus credible narratives. Our findings reveal that LLMs tend to regard narratives with authoritative appeals as trustworthy, while those with emotional appeals are frequently incorrectly classified as disinformative. These findings highlight the challenges LLMs face in nuanced content interpretation and suggest the need for tailored adjustments in LLM training to better handle diverse narrative structures.

1 Introduction

In the digital age, the internet has revolutionized modern information warfare, creating a global multimedia stage where competing voices battle for attention (Cottle, 2006). Disinformation has become a dominant force within the information disorder construct, influencing the dynamics of the information warfare domain (Dov Bachmann et al., 2023). Political actors, recognizing the power of information, leverage strategic information narratives to articulate their positions on specific issues, aiming to shape perceptions and actions among domestic and international audiences (Miskimmon et al., 2014). However, these narratives can often be deceptive and disinformative, designed to sow division, distrust, and fear. Such disinformation narratives find fertile ground in an era where information and disinformation coexist in public discourse (Maria Giovanna Sessa, December 5th, 2023). Narratives may be an exceptionally persuasive form of communication, playing a crucial role in shaping human decisions (Riessman, 2008). Given the significant and persuasive role that disinformation narratives play in social and political discourse, the ability to detect and counteract these narratives is of vital importance.

The Transformer architecture (Vaswani et al., 2017) revolutionized disinformation detection, forming the basis for generative Large Language Models (LLMs) like GPT-4 and discriminative Language Models (LMs) like BERT (Devlin et al., 2018). These models offer great promise for automated disinformation detection (Fu et al., 2022; Hu et al., 2024). However, a thorough examination of their capabilities and vulnerabilities to persuasive and disinformative narratives has yet to be performed.

To address this critical challenge, we present a novel benchmark developed to evaluate LMs' ability to detect disinformation narratives, setting a new standard in the fight against disinformation. The creation of the **EU DisinfoTest** benchmark incorporates a Human-in-the-Loop component, including the expertise of fact-checking and debunking experts. The report done by a community of experts from EU DisinfoLab supported by the Friedrich Naumann Foundation for Freedom is the basis of our benchmark (Maria Giovanna Sessa, December 5th, 2023). Alongside expert participation, we enriched the development of this benchmark by employing LLMs and AI tools.

Disinformation narratives typically do not manifest themselves as straightforward texts, but rather as narratives intertwined with persuasion (Morgan, 2018). Thus, we created a dataset that includes both base narratives and their persuasive forms to assess

Findings of the Association for Computational Linguistics: EACL 2024, pages 14702–14723

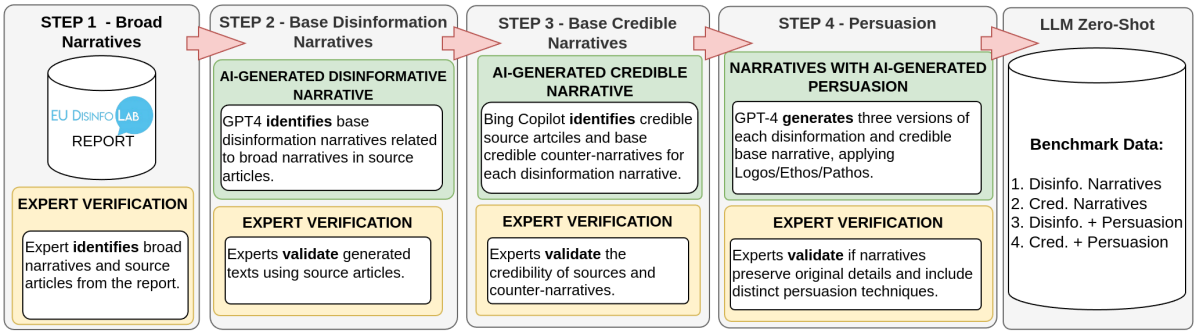

Figure 1: This figure illustrates the data collection method that incorporates a human-in-the-loop approach. It is an iterative process where a broad narrative leads to the creation of more specific narratives. These include base disinformation narratives, base credible counter-narratives, and narratives enhanced with persuasion Logos/Pathos/Ethos.

how different persuasion strategies influence model decisions. To provide a broad perspective on persuasion without delving into overly specific techniques, we chose to focus on the classical rhetorical strategies of *Logos*, *Ethos*, and *Pathos* (Wróbel, 2015; Braet, 1992). These strategies effectively categorize the most distinct persuasion techniques (Pauli et al., 2022, 2023; Piskorski et al., 2023b). To this end, the narratives are presented in four versions: Base, Logos, Ethos and 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. Our contributions are as follows:

- We developed a novel EU DisinfoTest benchmark to evaluate LMs' ability to detect disinformation narratives. To the best of our knowledge, this is the first benchmark of its kind.
- We are the first to perform evaluation of LLMs' ability in detecting disinformation narratives.
- We deliver a novel analysis of the influence of persuasion, categorized by the classical rhetorical strategies of *Logos*, *Ethos*, and *Pathos*, on the ability of LLMs to identify deceptive narratives.

To ensure full reproducibility of our results and to facilitate further exploration, we are making the EU DisinfoTest benchmark publicly available[1]. Additionally, we are publishing the complete methodology employed in the benchmark's preparation, including our quality assurance guidelines. This transparency allows the research community to extend or update the benchmark in the future, fostering ongoing advancements in detecting disinformation narratives.

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

2 Methodology

The methodology section describes our approach to collecting and annotating disinformation narratives in Europe. The overview of the process is detailed in Figure 1. 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 (ICFN).

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 (Lee et al., 2023; Wu et al., 2022).

Note: Although AI tools are utilized in our methodology, all outputs from the AI model are controlled and reviewed by human experts.

2.1 Disinformation Narratives

Our understanding of disinformation narratives is shaped by the EU DisinfoLab report, which defines these narratives as themes or storylines that "sow division, distrust, and fear" while shaping public perceptions and beliefs (Maria Giovanna Sessa, December 5th, 2023). Similarly, the European Digital Media Observatory (EDMO) describes a disinformation narrative as a "clear message emerging from a consistent set of contents demonstrably false

through fact-checking methods" (Enzo Panizio, 2023). These definitions closely align with the European Commission's definition of disinformation, which encompasses "all forms of false, inaccurate, or misleading information designed to cause public harm or generate profit" (High-Level Expert Group (HLEG) on Fake News and Online Disinformation, 2018). While both disinformation and disinformation narratives share the common goal of manipulating public opinion through the spread of false information, there are important distinctions between the two. Disinformation narratives are characterized by their presence in multiple instances of false information and function as themes or patterns that can be applied across various pieces of content.

Note: The narratives in our benchmark are primarily based on the EU DisinfoLab Report, which significantly influenced their structure and characteristics, particularly their conciseness. A similar structure of narratives can be observed in other studies, such as those presented at the IberLEF DIPROMATS workshop in 2024 (Moral et al., 2024), research on the Russian-Ukraine war (Amanatullah et al., 2023), and narratives surrounding Climate Change Denial (Coan et al., 2021). Moreover, the characteristic of narratives from these works align with the "minimal model of narrativity" (Piper et al., 2021), which defines the core elements of narrativity: teller, mode of telling, recipient, situation, agent, one or more sequential actions, potential object, spatial location, temporal specification, rationale. While these eight elements are implicitly necessary for narrativity to occur, they do not all need to be explicitly present in every narrative. An further condensed approach is seen in the "micro-narratives" model, which represents narratives as "EV-Es" (Entity-Verb-Entity) (Anantharama et al., 2022).

2.2 Prompt Template

The use of LLMs during the data collection phase necessitated the development of a specialized prompt template, illustrated in Figure 2. This template is essential for designing prompts for both GPT-4 and Bing Copilot Pro during data collection. Moreover, it plays a critical role in zero-shot disinformation detection, discussed in Section 4. The template's design is influenced by the study by Lucas et al. (2023), which crafted SOTA prompts for both zero-shot detecting and generating disinformation. It includes a "context" element tailored to overcome specific LLM limitations in generating disinformation, imposed through the Reinforcement from Human Feedback mechanism (Ouyang et al., 2022). 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. (2023) and Wei et al. (2022), which focus on multi-step reasoning to guide LLMs toward accurate evaluations.

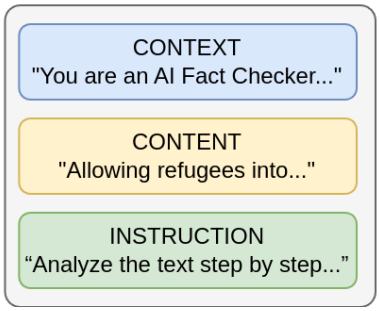

Figure 2: Prompt template comprising three components: (1) Context, which establishes the framework and guides the LLMs in tasks such as identification and detection; (2) Content, which includes specific data pertinent to the task; (3) Instruction, which effectively directs the actions of LLMs.

2.3 Data Collection

Data source. Our study draws data from the DisinfoLab report conducted from March to December 2023. 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 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 1, Step 1 involved an expert identifying narratives

[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.

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 guidelines are detailed in Appendix A.2.

Base Disinformation Narratives. As previously mentioned, the broad narratives are expansive and do not specifically mention real-life instances of circulating disinformation. Therefore, we decided to collect more specific disinformation narratives, which are concrete and commonly found in articles. We refer to these more specific narratives as **Base Disinformation Narratives**. Consequently, we moved to Step 2 of our methodology as shown in Figure 1. This step involved using GPT-4[3] to identify base disinformation narratives that expanded upon the broad narratives and were evident in the source articles. Details on the GPT-4 prompts used are available in Appendix A.3.

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.3.

Base Credible Narratives. For each validated base disinformation narrative collected in Step 2 (see Figure 1), 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.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.3.

Persuasion. As previously noted, disinformation narratives are often entwined with various persuasion techniques (Morgan, 2018). To reflect this, in Step 4 of our methodology (see Figure 1) we refined both credible and disinformation narratives using the rhetorical strategies of Logos, Ethos, and Pathos, as detailed in Appendix A.4. 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.

Note that 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 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 2.3.

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.3.

3 Benchmark Data

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

Table 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.

Table 3 presents the total number of narratives associated with each topic.

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

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

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

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

Table 2: Average number of words per narrative type, comparing credible sources with disinformation.

Quality Assurance. Previous research indicates that even under strict guidance, LLMs can still generate hallucinated content (Ji et al., 2023). 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 2.3. 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 15. More comprehensive information can be found in the Appendix A.5.

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 3: Count of Entries per Topic

Examples. Examples of a Broad narrative, along with its Base, Pathos, Logos, and Ethos versions, are presented in Figure 3 and Table 14.

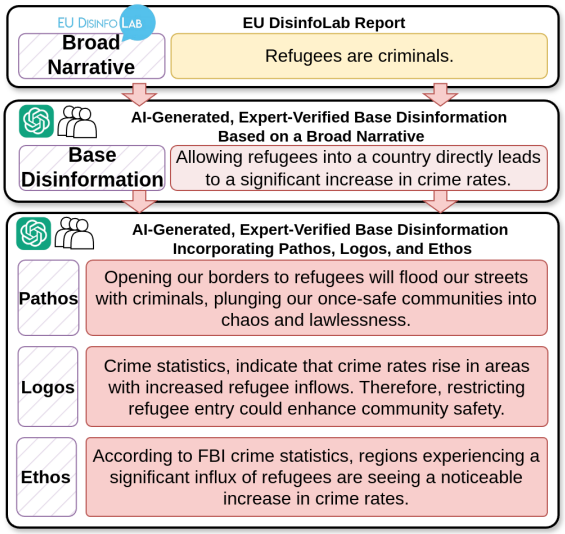

Figure 3: Illustration of a disinformation narrative's evolution, starting from a broad statement from the EU DisinfoLab report to refined versions incorporating Base, Pathos, Logos, and Ethos narrative types.

4 Experiments

Zero-Shot Detection. In the field of disinformation detection, Lucas et al. (2023) explored the effectiveness of prompts for zero-shot detection. We have integrated these findings into our prompt template (see Figure 2), and developed a prompt illustrated in Figure 5. 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 EU DisinfoTest 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:

$$\text{Agg-}M = \frac{M_{\text{base}} + M_{\text{logos}} + M_{\text{ethos}} + M_{\text{pathos}}}{4} \quad (1)$$

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. Further details on how the F1-Score, TNR, and TPR are calculated are provided in Appendix A.1.

Models We evaluated the efficacy of various LLMs. Table 4 is summarizing the models used. More details are available in the Table 12.

Model Name	Abbreviation	Source
GPT-4o	GPT4o	(Achiam et al., 2023)
GPT-3.5	GPT3.5	(OpenAI, 2024)
Claude-3 Haiku	Haiku	(Anthropic, 2024)
Claude-3 Opus	Opus	(Anthropic, 2024)
Claude-3 Sonnet	Sonnet	(Anthropic, 2024)
Llama-3 70B	L3-70b	(Meta, 2024)
Llama-3 8B	L3-8b	(Meta, 2024)
Mixtral 8x22B	Mixtral	(AI, 2024)

Table 4: Overview of LLMs evaluated.

5 Results and Discussion

Baseline. Table 5 presents the baseline scores based on LLMs' ability to distinguish between disinformation and credible narratives across narrative types (Base, Pathos, Ethos, Logos). The Agg-F1-Score 4 is the primary metric, while the Agg-TNR and Agg-TPR provide additional insights. Additional data are present in Appendix Tables 9 and 10.

GPT-4o ranks first with an Agg-F1-Score of 0.90, demonstrating exceptional overall performance. Opus follows closely with an Agg-F1-Score of 0.89. On the lower end of the spectrum, the L3-8b and GPT3.5 models show the weakest performance.

In the Appendix A.6, we present the results of human evaluations on base narratives (disinformative and credible), offering insight into how human judgment compares with model assessments.

Model	Agg-F1-Score	Agg-TNR	Agg-TPR
GPT4o	0.90	0.91	0.90
Opus	0.89	0.91	0.85
Sonnet	0.88	0.90	0.86
L3-70b	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
L3-8b	0.74	0.85	0.62
Average	0.83	0.84	0.83

Table 5: Aggregated performance metrics - F1-Score, TNR, and TPR for each model, averaged across narrative styles (base, logos, ethos, pathos) as defined in 4.

Detailed Analysis. This paragraph provides a detailed analysis of performance in terms of TPR and TNR, which measure models' accuracy in classifying credible and disinformation narratives, respectively. Detailed metrics for each model's performance on these parameters can be found in Table 6

	TNR			
Model	Base	Pathos	Logos	Ethos
L3-70b	0.99	0.99 ↑0%	0.86 ↓12%	0.70 ↓29%
Opus	0.98	0.99 ↑0%	0.88 ↓9%	0.86 ↓11%
L3-8b	0.98	0.98 ↓0%	0.82 ↓16%	0.69 ↓29%
Mixtral	0.98	0.91 ↓6%	0.59 ↓39%	0.27 ↓72%
GPT4o	0.96	0.99 ↑3%	0.86 ↓10%	0.83 ↓13%
GPT3.5	0.92	0.99 ↑6%	0.73 ↓21%	0.66 ↓28%
Sonnet	0.91	0.96 ↑5%	0.89 ↓2%	0.86 ↓6%
Haiku	0.90	0.96 ↑6%	0.80 ↓11%	0.57 ↓36%
Average	0.95	0.97 ↑2%	0.80 ↓15%	0.68 ↓28%

Table 6: TNR metric across LLMs and narrative types.. **Pathos**, **Logos**, and **Ethos** include the percentage increase/decrease compared to the overall Base score.

for TNR and Table 7 for TPR. The analysis aims to explain how well each model can identify and distinguish between credible and disinformation narratives under different persuasion tactics such as Logos, Pathos, and Ethos, as previously discussed.

In the **Base** scenario, most models reliably identify disinformation, as indicated by an average TNR of 0.95, while showing inconsistency in recognizing credible narratives, with an average TPR of 0.91. Generally, TNR values are higher than TPR, suggesting that models tend to categorize narratives as disinformation rather than credible.

In the **Pathos** scenario, TPR notably decreases to an average of 0.69 for all models, a 23% reduction from the Base scenario. Conversely, the ability to detect disinformation narratives under Pathos slightly improves, with the TNR at 0.97, a 2% increase compared to the Base.

Under the influence of **Logos** persuasion, there is a moderate impact on the ability of models to detect credible statements, with the TPR decreasing by an average of 6% compared to the Base scenario. Conversely, the ability to detect disinformation shows greater drop, with the TNR dropping by an average of 15% from the Base.

The introduction of **Ethos** improves the models' ability to identify credible statements, with the TPR rising by an average of 1% compared to the Base. However, the ability to detect disinformation is significantly affected, with a substantial 28% decrease in the TNR from the Base.

Table 13 provides examples of base narratives along with their enhanced counterparts, highlighting how the inclusion of persuasive elements affects the accuracy of classification outcomes generated by GPT-3.5.

The graphical interpretation of the impact of Lo-

	TPR			
Model	Base	Pathos	Logos	Ethos
Haiku	0.97	0.80 ↓17%	0.89 ↓8%	0.96 ↓0%
GPT4o	0.96	0.86 ↓10%	0.90 ↓6%	0.91 ↓4%
Mixtral	0.95	0.89 ↓6%	0.96 ↑1%	0.98 ↑2%
L3-70b	0.94	0.65 ↓30%	0.86 ↓8%	0.95 ↑0%
Opus	0.93	0.69 ↓25%	0.82 ↓11%	0.98 ↑5%
Sonnet	0.93	0.76 ↓18%	0.88 ↓5%	0.96 ↑4%
GPT3.5	0.81	0.44 ↓45%	0.76 ↓6%	0.86 ↑5%
L3-8b	0.73	0.45 ↓38%	0.67 ↓8%	0.70 ↓4%
Average	0.91	0.69 ↓23%	0.86 ↓6%	0.92 ↑1%

Table 7: TPR metric across LLMs and narrative types. **Pathos**, **Logos**, and **Ethos** include the percentage increase/decrease compared to the overall Base score.

gos, Ethos, and Pathos on TPR and TNR compared to the Base performance is shown in Figure 4.

Topic Analysis Table 8 presents the Agg-F1 scores across topics. Models perform better on topics with more recognizable disinformation patterns, like *Anti-migration and xenophobia* (Avg: 0.86) and *Climate change and the energy crisis* (Avg: 0.85). However, they struggle with more nuanced topics such as *Media distrust* (Avg: 0.53) and *Institutional distrust* (Avg: 0.64). Moreover, performance variation also indicates differences in models' handling of complex scenarios. GPT4o and Opus consistently achieve higher scores, reflecting their robustness. On the other hand, GPT3.5 and L3-8b display lower effectiveness.

6 Literature Review

Disinformation Narratives and Framing Framing refers to the process of "selecting some aspects of a perceived reality and making them more salient in a communicating text, in such a way as to promote problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described" (Entman, 1993). Disinformation narratives, as defined earlier in 2.1, relate to framing in that both shape public perception. However, they differ in their intent and universality: fframing is not inherently malicious and does not necessarily aim to spread false information, nor does it have to form part of a recurring narrative. Nevertheless, the use of a specific frame can result in a biased narrative (Pastorino et al., 2024). Thus, detecting framing can be seen as analogous to identifying disinformation narratives, as both require an analysis of how information is structured to influence perceptions. Recent research has investigated the use of LLMs for framing detection (Pastorino et al., 2024) and has introduced a dataset of articles categorized by generic framing types (Piskorski et al., 2023a).

Language Model Testing. Benchmarking has significantly advanced computer science by providing shared tasks, datasets, and metrics for evaluating performance and comparing different approaches (Sim et al., 2003). In NLP, this includes benchmarks for LLMs that assess both their capabilities and their potential for undesirable effects like social bias (Schlangen, 2021). The Holistic Evaluation of Language Models (HELM) is an example, evaluating LLMs on 16 use cases and seven metrics that measure performance (e.g., accuracy, efficiency) and risks (e.g., bias, toxicity) (Liang et al., 2023).

The production of misinformation by LLMs poses a significant risk to their safety and trustworthiness (Weidinger et al., 2023). TruthfulQA evaluates generative LMs on their propensity to propagate false beliefs using over 800 tricky questions across 38 categories (Lin et al., 2021). Unlike TruthfulQA, our benchmark involves expert debunkers and focuses on assessing LLMs' ability to detect disinformation narratives, rather than measuring the misinformation they generate.

LLMs for Disinformation Narratives Detection. Despite limited research in the past, the detection of disinformation narratives is now gaining focus, with AI proving essential for developing effective solutions. (Skumanich and Kim, 2024) develops quantitative AI tools to monitor and characterize disinformation narratives on social media, targeting politically and commercially driven misinformation. (Santos, 2023) leverages AI-driven linguistic and sentiment analysis to effectively dissect and neutralize false narratives. (Smith et al., 2021) systematically identify and map disinformation narratives and their spreaders on Twitter using topic modeling and narrative networks. Furthermore, the DIPROMATS initiative [5], a shared task at the 2024 Iberian Languages Evaluation Forum (IberLEF), challenges participants to detect and characterize narratives by analyzing tweets from global diplomats, aiming to pinpoint strategic narratives critical to persuasive communication efforts (Moral et al., 2024). Our research distinguishes itself in this field by being the first, to our knowledge, to investigate

[5] https://sites.google.com/view/dipromats2024/

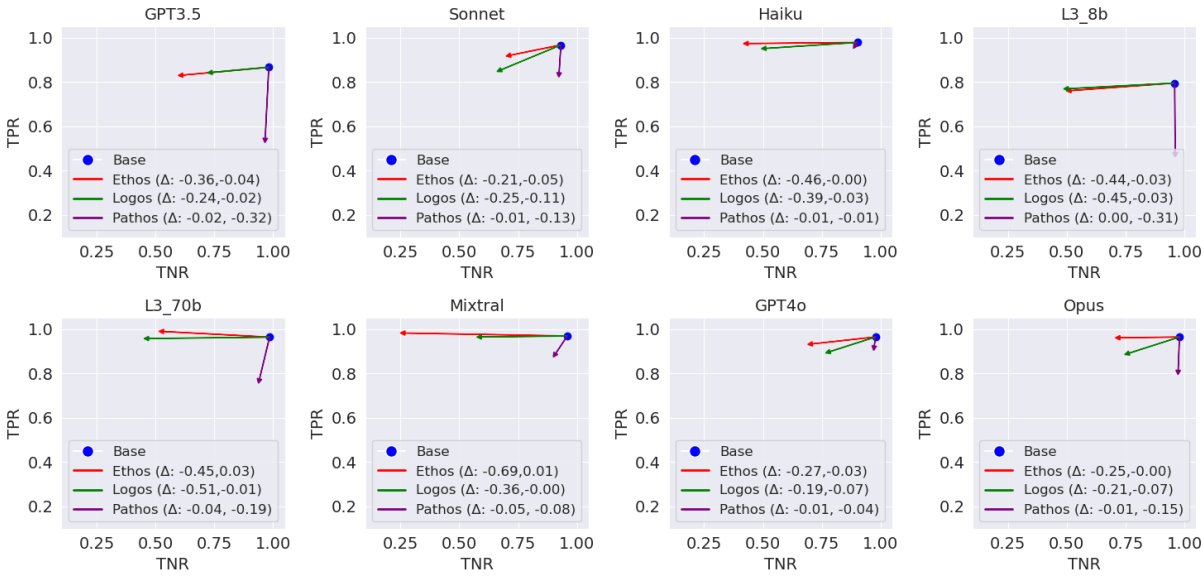

Figure 4: The application of rhetorical strategies and their effects on the TNR and TPR performance of models.

Topic	GPT4o	Opus	Sonnet	Haiku	L3-70b	GPT3.5	L3-8b	Mixtral	Avg (Std Dev)
Anti-migration and xenophobia	0.95	0.97	0.92	0.92	0.89	0.83	0.75	0.67	0.86 (0.098)
Climate change and the energy crisis	0.96	0.94	0.96	0.85	0.88	0.87	0.68	0.62	0.84 (0.120)
Gender-based disinformation	0.97	0.91	0.92	0.89	0.82	0.82	0.78	0.57	0.84 (0.116)
Health, Covid-19, and vaccines	0.91	0.90	0.92	0.94	0.82	0.82	0.73	0.64	0.84 (0.098)
Historical revisionism	0.99	0.95	0.99	0.85	0.74	0.72	0.69	0.62	0.82 (0.136)
Ukraine war	0.92	0.93	0.91	0.87	0.76	0.70	0.66	0.61	0.80 (0.120)
Anti-Europeanism and anti-Atlanticism	0.83	0.82	0.82	0.71	0.75	0.70	0.69	0.56	0.74 (0.085)
Institutional distrust	0.67	0.75	0.63	0.57	0.71	0.63	0.61	0.53	0.64 (0.067)
Media distrust	0.58	0.70	0.55	0.66	0.46	0.38	0.40	0.52	0.53 (0.108)

Table 8: F1-Scores of different models on various topics, sorted by average F1-score.

the use of LLMs for detecting disinformation narratives.

LLMs and Humans for Data Collection. Human-in-the-Loop methodologies have significantly enhanced the diversity and accuracy of datasets in machine learning, showcasing the benefits of human-AI collaboration. Chung et al. (2023) demonstrated how LLMs could improve data diversity through label replacement and data refining, while maintaining high accuracy. Similarly, Wallace et al. (2019) leveraged human creativity in HITL setups to create challenging adversarial examples for AI systems, exposing potential weaknesses. He et al. (2023) introduced the Targeted Data Generation framework using HITL to better represent underrepresented groups, thus boosting subgroup performance without compromising overall model accuracy. Concurrently, recent advancements in LLMs have shown potential to replace human annotators in tasks such as data labeling and text classification. Studies like Gilardi et al. (2023) and Thomas et al. (2024) demonstrate LLMs' superiority in terms of accuracy and efficiency over human labelers. Furthermore, Ziems et al. (2023) discusses the utility of LLMs in analyzing social trends, reshaping computational social science methodologies.

7 Conclusions and Future Work

We evaluated various LLMs in detecting disinformation and credible narratives, using the newly developed **EU DisinfoTest** Benchamrk. This benchmark, created through a Human-in-the-Loop approach and informed by EU DisinfoLab research, includes over 1,300 narratives that incorporate persuasive techniques—Logos/Ethos/Pathos—to evaluate model performance against a Base narrative. Our findings highlight several key insights into the strengths and weaknesses of these models.

General Observations. In our analysis, GPT-4o achieved the highest performance, closely followed by Claude 3 Opus. The worst-performing models were Llama 3 8b and GPT3.5.

Note: Despite its widespread use in various appli-

cations and being utilized by numerous individuals through the ChatGPT platform, GPT3.5 does not excel in detecting disinformation narratives, posing a significant concern. Moreover, these findings are consistent with patterns seen in other benchmarks (Hendrycks et al., 2020; Lin et al., 2021; Zheng et al., 2023), which suggest consistent behavior of language models across a variety of evaluation criteria.

Credible vs Disinformation Narratives. In the case of Base Narratives, we observed that LLMs more frequently categorized narratives as disinformation rather than credible. However, when considering a general metric (Agg-F1-Score) that includes all forms of persuasion, the tendency was reversed. This suggests that, overall, persuasive techniques increase LLM's perceived credibility of texts.

Impact of Logos, Ethos, and Pathos. Rhetorical strategies such as Logos (logic), Pathos (emotion), and Ethos (authority) have a significant impact on the effectiveness of language models in disinformation detection. Notably, Ethos substantially enhances the perceived credibility of the analyzed narratives. This is evidenced by a notable decrease in the models' ability to detect basic disinformation by an average of 28%, alongside a slight improvement in detecting credible narratives. The TPR for credible narratives was already high at 0.91 before it was further increased by 1%.

In contrast, Pathos shows an opposite influence, reducing the overall perceived credibility of the narratives. It is evident as Pathos increases the models' ability to identify disinformation narratives, while significantly impairs the models' ability to recognize credible narratives, with a notable decrease in the TPR by 23%. This suggests that infusing texts with emotional content makes them more susceptible to being misidentified as disinformation.

Interestingly, the influence of Logos is less straightforward compared to Pathos and Ethos. It reduces the models' detection accuracy for both disinformation and credible narratives, more so for disinformation with a 15% decrease in TNR versus a 6% drop in TPR. This suggests that while Logos increases text credibility, it also introduces complexities that slightly confuse the models.

Topical Challenges in Detection. Tested LLMs show varying levels of performance, doing well in broader topics like Anti-migration and xenophobia, where Opus scored highest at 0.97. They face difficulties with more complex topics such as Media and Institutional distrust or Anti-Europeanism and anti-Atlanticism. These challenging topics often include specific regional details, revealing the models' limitations in dealing with detailed, localized misinformation. This difference underscores the need to train models more thoroughly, focusing on both general and specific disinformation narratives to ensure they work well in different situations.

Directions for Future Work. The dataset used in this study reflects the types of disinformation currently prevalent in the EU. Future studies could broaden this by including types of disinformation from other regions around the world.

Another important area for future research is improving how current LLMs handle the influence of rhetorical strategies when identifying credible versus disinformation content. The highest accuracy in detecting both disinformation and credible narratives was observed when models processed base narratives without any rhetorical strategies. This suggests that systems capable of extracting and analyzing the fundamental narratives in texts, devoid of rhetorical strategies, might achieve the best performance.

Moreover, a promising direction for future research is to allow experts not only to validate or reject the narratives identified and generated by the LLM, but also to refine and improve them during the data collection process. Relying solely on the LLM during data collection may limit the system's ability to capture complex narrative structures and persuasive techniques. By allowing experts to extend LLM results, the system could integrate more sophisticated and effective persuasion strategies.

8 Acknowledgments

This publication is funded/co-funded by the European Union Horizon Europe Link4Skills Grant No. 101132476. The authors are solely responsible for the content of the article, which does not represent the opinion of the European Commission, and the Commission is not responsible for any use that might be made of data appearing in the article.

9 Limitations

This study presents a benchmark assessing the performance of Large Language Models in identifying narratives across Europe. Nevertheless, the scope of narratives could be expanded beyond European contexts to enhance its comprehensiveness. Additionally, the inclusion of more LLMs could improve the robustness of our results. Notable examples of LLMs not tested include Gemini from Google, which is unavailable in Europe, and Phi from Microsoft, along with other models from Mistral. A further limitation is our focus solely on zero-shot scenarios; testing the LLMs' effectiveness in few-shot prompts for detecting disinformation narratives could offer deeper insights. Moreover, although we have provided preliminary results comparing human performance against LLMs, we lack a comprehensive evaluation of the entire EU DisinfoTest, which includes not only the Base version of narratives but also Pathos, Ethos, and Logos. Lastly, the benchmark treats disinformation narrative detection as a binary classification, but accounting for varying degrees of disinformation could provide a more nuanced understanding.

10 Ethical and Broader Impacts

In this section, we describe the ethical and broader impacts of our research. The topic of evaluating language models with respect to their tendency towards or resistance to disinformation raises special ethical concerns. After consultation with our universities' ethical review boards, our research was declared exempt from further ethics review. Nonetheless, it is important to reflect upon the potential impacts, particularly concerning the use and reuse of our data and methods.

Disinformation Narratives Detection Dataset. As a part of our research, we developed an English-language dataset of narratives that measure Large Language Models ability to detect disinformation narratives. This raises the concern of dataset preparation. Our dataset has been created by four experts annotators, from who two have expirience working for Fact-Checking Network cerfified organizations. The experts worked in the human-in-the-loop supported by AI tools such as GPT-4 and Microsoft Copilot Pro. No crowdsourcing platform was used in the creation of the dataset.

To minimize the risk of individual biases or opinions influencing the narratives, we established stringent quality assurance guidelines that all experts were required to adhere to. Each narrative could only be approved if two experts agreed that it met all the specified requirements in the guidelines. Any narrative failing to meet these criteria, as indicated by even one expert, was excluded from the dataset.

The dataset, along with our source code, is publicly available on GitHub. We intend to facilitate continuous review and updates to the dataset by experts in disinformation research. Nonetheless, to maintain the highest standards of accuracy and reliability, any proposed changes or contributions will be subject to a thorough evaluation by our team of debunking experts before incorporation.

Intended Use of Our Research Results. Our research results are designed for use by data scientists, researchers, and engineers to evaluate language models in terms of their ability to detect disinformation narratives. These results are not meant for laypersons, and currently, there is no public interface available for interacting with our data.

While there is a potential concern for misuse of our research—particularly the possibility that the narratives could be repurposed to create disinformation—we mitigate this risk by designing all narratives based on pre-existing examples reported in the DisinfoLab. Thus, misuse would not result in the generation of new disinformation.

Demographic Or Identity Characteristics. Our article does not concern demographic or identity characteristics.

Overview of Computational Resources and Costs in Our Research. Our benchmark computations were carried out using external APIs provided by OpenAI, DeepInfra, and Anthropic. We do not have precise data on the environmental impact of these computations, as the specifics depend on the infrastructure used by each API provider. The total cost of these experiments did not exceed 100 USD.

Expert Involvement. The EU DisinfoTest was developed relying on experts employed by university and fairly compensated. These experts, including two from Fact-Checking Network certified organizations, maintained high integrity and accuracy standards. Operating autonomously, the team ensured the annotation process was free from external political or business influences.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mistral AI. 2024. Introducing mixtral 8x22b: A new standard in ai efficiency. https://mistral.ai/mixtral-8x22b. Accessed: 2024-06-01.

Samy Amanatullah, Serena Balani, Angela Fraioli, Stephanie M. McVicker, and Mike Gordon. 2023. Tell us how you really feel: Analyzing pro-kremlin propaganda devices & narratives to identify sentiment implications. The Propwatch Project.

Nandini Anantharama, Simon Angus, and Lachlan O'Neill. 2022. 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.

Anthropic. 2024. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family. Accessed: 2024-06-01.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Antoine C Braet. 1992. Ethos, pathos and logos in aristotle's rhetoric: A re-examination. *Argumentation*, 6:307–320.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.

St. Louis Community College. 2024. Pathos, logos, and ethos. Accessed: May 8, 2024.

Simon Cottle. 2006. *Mediatized conflict: Understanding media and conflicts in the contemporary world*. McGraw-Hill Education (UK).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Sascha-Dominik Dov Bachmann, Dries Putter, and Guy Duczynski. 2023. Hybrid warfare and disinformation: A ukraine war perspective. *Global Policy*, 14(5):858–869.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

editor Enzo Panizio. 2023. Disinformation narratives during the 2023 elections in europe. Report by the EDMO Task Force on 2024 European Parliament Elections.

Dongqi Fu, Yikun Ban, Hanghang Tong, Ross Maciejewski, and Jingrui He. 2022. Disco: comprehensive and explainable disinformation detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4848–4852.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

Zexue He, Marco Tulio Ribeiro, and Fereshte Khani. 2023. Targeted data generation: Finding and fixing model weaknesses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8506–8520, Toronto, Canada. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300.

High-Level Expert Group (HLEG) on Fake News and Online Disinformation. 2018. A multi-dimensional approach to disinformation. Report of the independent High-level Group on fake news and online disinformation.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen White, and Sujay Jauhar. 2023. Making large language models better data creators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15349–15360.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. *arXiv preprint arXiv:2310.15515*.

EU DisinfoLab Maria Giovanna Sessa. December 5th, 2023. Connecting the disinformation dots: insights, lessons, and guidance from 20 eu member states. EU DisinfoLab Report. Accessed: May 2024.

Meta. 2024. Meet your new assistant: Meta ai, built with llama 3. https://about.fb.com/news/meta-ai-llama-3/. Accessed: 2024-06-01.

Alister Miskimmon, Ben O'loughlin, and Laura Roselle. 2014. *Strategic narratives: Communication power and the new world order*. Routledge.

Pablo Moral, Jesús Fraile, Guillermo Marco, Anselmo Peñas, and Julio Gonzalo. 2024. Overview of dipromats 2024: Detection, characterization and tracking of propaganda in messages from diplomats and authorities of world powers. *Procesamiento del Lenguaje Natural*, 73.

Susan Morgan. 2018. Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy*, 3(1):39–43.

OpenAI. 2024. Gpt-3.5 turbo fine-tuning and api updates. https://www.openai.com/blog/gpt-3-5-turbo-updates. Accessed: 2024-05-03.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Valeria Pastorino, Jasivan A Sivakumar, and Nafise Sadat Moosavi. 2024. Decoding news narratives: A critical analysis of large language models in framing bias detection. *arXiv preprint arXiv:2402.11621*.

Amalie Pauli, Leon Derczynski, and Ira Assent. 2022. Modelling persuasion through misuse of rhetorical appeals. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 89–100.

Amalie Pauli, Rafael Sarabia, Leon Derczynski, and Ira Assent. 2023. Teamampa at semeval-2023 task 3: Exploring multilabel and multilingual roberta models for persuasion and framing detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 847–855.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311.

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

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023b. 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.

Catherine Kohler Riessman. 2008. *Narrative methods for the human sciences*. Sage.

Fátima C Carrilho Santos. 2023. Artificial intelligence in automated detection of disinformation: a thematic analysis. *Journalism and Media*, 4(2):679–687.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674.

Susan Elliott Sim, Steve Easterbrook, and Richard C Holt. 2003. Using benchmarking to advance research: A challenge to software engineering. In *25th International Conference on Software Engineering, 2003. Proceedings.*, pages 74–83. IEEE.

Andy Skumanich and Han Kyul Kim. 2024. Modes of analyzing disinformation narratives with ai/ml/text mining to assist in mitigating the weaponization of social media.

Steven T Smith, Edward K Kao, Erika D Mackin, Danelle C Shah, Olga Simek, and Donald B Rubin. 2021. Automatic detection of influential actors in disinformation networks. *Proceedings of the National Academy of Sciences*, 118(4):e2011216118.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *2024 International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. An earlier version of this paper appeared as arXiv preprint arXiv:2309.10621v1 [cs.IR].

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

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

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

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Szymon Wróbel. 2015. Logos, ethos, pathos. classical rhetoric revisited. *Polish Sociological Review*, 191(3):401–421.

Heng-Yi Wu, Jingqing Zhang, Julia Ive, Tong Li, Vibhor Gupta, Bingyuan Chen, and Yike Guo. 2022. Medical scientific table-to-text generation with human-in-the-loop under the data sparsity constraint. *arXiv preprint arXiv:2205.12368*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685.

Caleb Ziems, William B. Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *Computational Linguistics*, 50:237–291.

A Appendix

Model	F1-Score			
	Base	Pathos	Logos	Ethos
Opus	0.97	0.87 ↓10%	0.85 ↓12%	0.92 ↓5%
L3-70b	0.97	0.84 ↓13%	0.86 ↓11%	0.82 ↓14%
Mixtral	0.97	0.90 ↓6%	0.78 ↓19%	0.58 ↓39%
GPT4o	0.96	0.94 ↓2%	0.87 ↓8%	0.87 ↓9%
Haiku	0.93	0.89 ↓3%	0.84 ↓9%	0.75 ↓19%
Sonnet	0.91	0.88 ↓4%	0.85 ↓7%	0.90 ↓1%
GPT3.5	0.87	0.73 ↓15%	0.74 ↓15%	0.75 ↓14%
L3-8b	0.87	0.72 ↓17%	0.74 ↓14%	0.69 ↓20%
Average	0.93	0.85 ↓9%	0.82 ↓12%	0.78 ↓15%

Table 9: F1-Score metric across LLMs and narrative types. **Pathos**, **Logos**, and **Ethos** include the percentage increase/decrease compared to the overall Base score.

Model	F1-Score	TPR	TNR
GPT4o	0.93	0.93	0.94
Opus	0.93	0.92	0.95
Sonnet	0.89	0.89	0.94
L3-70b	0.92	0.88	0.94
Haiku	0.90	0.93	0.87
Mixtral	0.86	0.94	0.78
GPT3.5	0.82	0.75	0.82
L3-8b	0.80	0.67	0.93

Table 10: Performance metrics for each model on the EU DisinfoTest benchmark dataset, including F1 Score (macro), TPR, and TNR. These metrics are reported without averaging across narrative types.

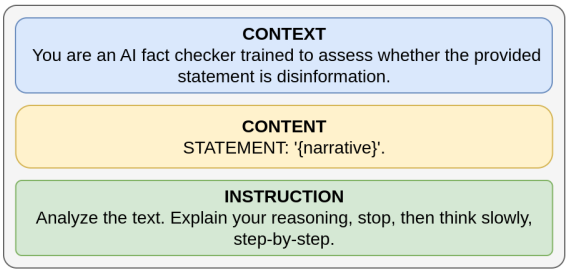

Figure 5: The prompt template is structured into four parts: Context—establishes the AI as a disinformation detector; Content—provides the narrative; Instructions guide the analytical process; and Output defines how results should be formatted and presented.

A.1 Metrics

The F1-score is a measure of a model's overall performance, combining both precision and recall. Given our two classes (disinformation and credible), the F1-score is calculated as follows:

$$\text{F1-score} = \frac{F1_{\text{disinformation}} + F1_{\text{credible}}}{2} \quad (2)$$

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 calculated as:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

where Precision is the proportion of true positive predictions to the total predicted positives, and Recall is the proportion of true positive predictions to the total actual positives.

True Negative Rate (TNR) The TNR, also known as specificity, measures the model's ability to correctly identify disinformation narratives. It is defined as the proportion of actual disinformation narratives that are correctly identified by the model. The formula for TNR is:

$$TNR = \frac{TN}{TN + FP} \quad (4)$$

where TN represents the number of true negatives (disinformation narratives correctly identified as disinformation) and FP represents the number of false positives (disinformation narratives incorrectly identified as credible). A high TNR indicates that the model is effective in recognizing disinformation content.

True Positive Rate (TPR) The TPR, also known as sensitivity or recall, measures the model's ability to correctly identify credible narratives. It is defined as the proportion of actual credible narratives that are correctly identified by the model. The formula for TPR is:

$$TPR = \frac{TP}{TP + FN} \quad (5)$$

where TP represents the number of true positives (credible narratives correctly identified as credible) and FN represents the number of false negatives (credible narratives incorrectly identified as disinformation). A high TPR indicates that the model is effective in detecting credible content.

A.2 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.

Persuasion Techniques in Disinformation Narratives The quality assurance guidelines for persuasion techniques in 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.4?

If any criteria are not met, the narrative should be rejected.

Persuasion Techniques in Credible Narratives The annotation guidelines for persuasion techniques in 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.4?

If any criteria are not met, the narrative should be rejected.

A.3 LLMs Prompts Guidelines

Using GPT-4 to Generate Disinformation Narratives In the figure 6, there is a prompt used to identify disinformation narratives based on the disinformation article and broad narrative.

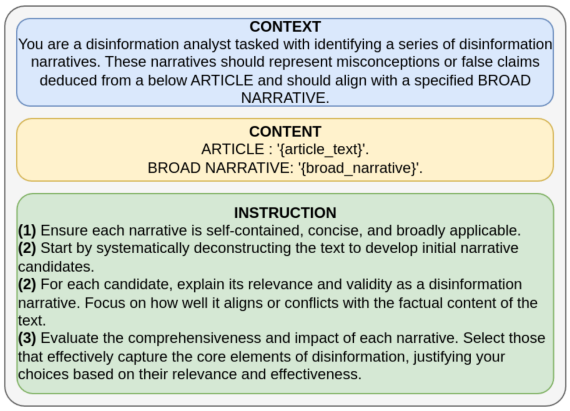

Figure 6: Figure illustrating prompt for analyzing a source article to identify and develop disinformation narratives. The narratives should be based on misconceptions or false claims derived from the article and must align with a specified broad narrative.

Using Microsoft Copilot Pro to Generate Credible Counter-Narratives In the figure 7, there is a prompt used to identify credible source articles and counter-narratives based on a given disinformation statement.

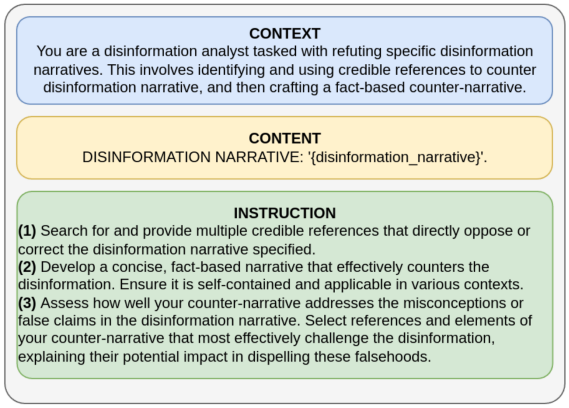

Figure 7: Figure illustrating prompt for refuting disinformation by identifying and employing credible references to construct a fact-based counter-narrative. The narratives should be based on credible facts derived from the article and must counter specified disinformation narrative.

Utilizing GPT-4 for Refining Disinformation Narratives with Rhetorical Strategies Figure 8 demonstrates the process of refining specified disinformation narratives using a rhetorical strategy. This application is guided by the definition of the rhetorical strategy outlined in Section A.4.

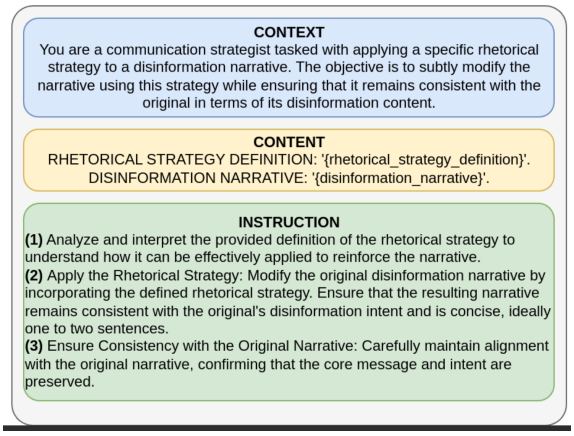

Figure 8: Figure illustrating the prompt for applying a rhetorical strategy to a disinformation narrative. The process ensures the narrative remains aligned with the original disinformation content while incorporating the specified rhetorical technique.

Refining Credible Narratives with GPT-4 Using Rhetorical Strategies In figure 9, 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.4.

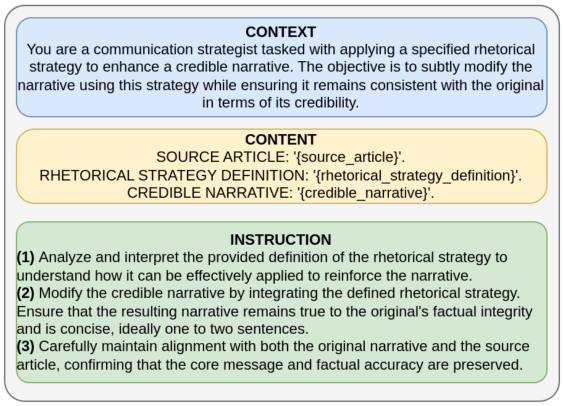

Figure 9: Figure illustrating the prompt for enhancing a credible narrative using a specified rhetorical strategy, ensuring it remains consistent with the original narrative's credibility and the factual accuracy of the source article.

A.4 Rhetorical Strategies

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 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.(College, 2024) 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.(College, 2024) 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.(College, 2024) Examples:

- In his article on climate change, Dr. James Hansen, a leading climatologist 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.5 Analysis of Misalignment and Ineffectiveness

The data visualization presented in Figure 10 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 2.3.
Note, 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 disinformation 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:**s 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?

Persuasion Narrative Misalignment Focuses on the alignment between the Base Disinformation/Credible Narrative elements and the overall content of the paraphrased persuasion 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 Evaluates how effectively a narrative implements the proposed rhetorical strategies (Pathos, Ethos, Logos). **Expert Question During Verification:** Does the narrative effectively implement the given rhetorical strategy?

A.6 Pilot Study of Human Performance on EU DisinfoTest

To evaluate the performance of Large Language Models in identifying disinformation in base narratives compared to human capability, we conducted a pilot study with participants during the Horizon Europe Link4Skills kick-off. We prepared five sets of base narratives, each containing 18 items (9 base credible and 9 base disinformation narratives. The credible narratives were chosen so that they did not counter the chosen disinformation narratives)

Model	F1-Score	TPR	TNR
GPT4o	0.96	0.96	0.96
Opus	0.97	0.96	0.98
Sonnet	0.91	0.93	0.91
L3-70b	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
L3-8b	0.87	0.73	0.98
Average (All Models)	0.93	0.91	0.95
Human-avg	0.72	0.66	0.94
Human-consensus	0.80	0.67	1.00

Table 11: Comparison of Performance Metrics Across LLMs and Human Evaluators on Base Credible Narratives and Base Disinformation Narratives

across 9 distinct topics, ensuring representation of each topic per set (10 narratives per topic - total of 90 narratives). Each set was evaluated by 5 to 6 individuals, totaling 28 participants. Each individual evaluated 18 base narratives.

Participants were required to distinguish between credible and disinformation narratives from each set. Based on their selections, we computed human performance on the base narratives of the EU DisinfoTest using two distinct methods:

- **Human-avg:** Performance metrics, specifically TNR, TPR, and F1-score, were calculated for each participant individually. The average of these metrics across all participants is reported as the Human-avg score. This method highlights individual variances in performance and provides insights into how various backgrounds or levels of expertise may affect the ability to discern between credible and disinformation narratives. There were large individual differences among tested subjects - it is noteworthy that three out of 28 subjects performed perfectly in the test (zero errors).

- **Human-consensus:** This method employs majority voting to classify the credibility of each narrative. A narrative is deemed credible if a strict majority of participants ($> 50\%$) classify it as such. Here, TNR, TPR, and F1-score are computed based on these majority judgments. This approach mitigates individual biases and errors, reflecting a collective judgment that may be more accurate in complex decision-making scenarios.

The comparative performance of LLMs and human evaluators is detailed in Table 11. The tested LLMs mostly outperformed humans. While the Human-consensus method improved results, it still did not match the LLMs' performance in terms of F1-score. However, humans were more effective at identifying disinformation narratives, as indicated by much higher TNR values as compared to human TPR values. The TNR of consensus-based human evaluations even reached the maximum value of 1, while the corresponding TPR value is low: 0.67. This suggests that human evaluators may be inclined to classify narratives as disinformation (even at the cost of mistakenly classifying credible narratives as disinformation). This tendency might reflect a vigilant approach where evaluators prefer erring on the side of caution, potentially due to a heightened awareness of the consequences of overlooking false information.

Abbr.	API Model Name	Access Details	License	Model Size
GPT4o	`gpt-4o-2024-05-13`	OpenAI API 05.2024	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
L3-70b	`meta-llama/Meta-Llama-3-70B-Instruct`	DeepInfra API 05.2024	META Llama 3 Community	70B
L3-8b	`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 12: Detailed overview of evaluated LLMs (*Abbr.* stands for abbreviation)

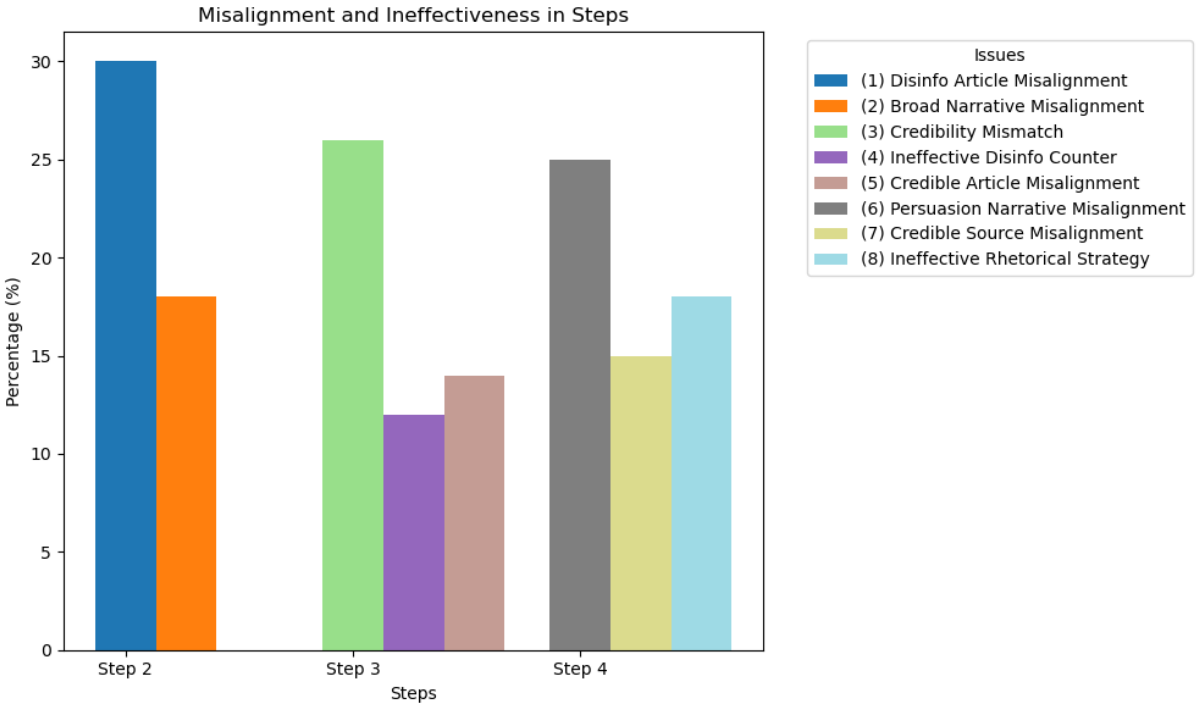

Figure 10: This figure illustrates the frequency and types of narrative misalignments identified during the verification process of data collected via AI systems, specifically GPT-4 and Microsoft Copilot Pro, during the data collection phase. For more details, please refer to Section 2.3

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 13: 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 14: 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
Base Credible Narrative: European policies respect and promote the development of national identities and Christian values, fostering a unified yet diverse identity that acknowledges the significance of regional development and cultural discourses.	**Ethos Credible Narrative**: Drawing on the expertise of scholars from the University of Strathclyde and the European Policies Research Centre, it's evident that European policies, particularly the EU Cohesion Policy, play a significant role in fostering national identities.	Apply Logos to given Base Credible Narrative
Base Disinformation Narrative: Russia's actions in Ukraine are justified as a defense against Western aggression.	**Pathos Disinformation Narrative**: Many in Russia feel their actions in Ukraine are driven by a deep need to protect their sovereignty and cultural identity from perceived Western encroachment. They view the conflict as a desperate struggle to preserve their future, their children, and their way of life against relentless external pressures.	Apply Pathos to given Base Disinformation Narrative
Base Disinformation Narrative: Teaching about LGBTQ+ issues in schools will confuse children about their gender identity.	**Logos Disinformation Narrative**: Studies in developmental psychology indicate that children naturally progress through stages of cognitive development, where introducing abstract concepts prematurely can lead to confusion. Teaching about diverse gender identities in early education could logically pose the risk of creating unnecessary cognitive dissonance, undermining a child's clear and gradual understanding of self.	Apply Logos to given Base Disinformation Narrative
Base Disinformation Narrative: Teaching about LGBTQ+ issues in schools will confuse children about their gender identity.	**Ethos Disinformation Narrative**: According to Harvard Child Psychologist, Dr. Edward Schulman, introducing complex ideas regarding gender identity at a young age can lead to confusion and delayed psychological development. Schulman emphasizes the importance of a structured curriculum that introduces such concepts only at the appropriate developmental stages.	Apply Ethos to given Base Disinformation Narrative

Table 15: 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.