MALicious INTent Dataset and Inoculating LLMs for Enhanced Disinformation Detection

Arkadiusz Modzelewski[1,2,3], **Witold Sosnowski**[2], **Eleni Papadopulos**[1,4], **Elisa Sartori**[1], **Tiziano Labruna**[1], **Giovanni Da San Martino**[1], **Adam Wierzbicki**[2]

[1]University of Padua, Italy
[2]Polish-Japanese Academy of Information Technology, Poland
[3]NASK National Research Institute, Poland
[4]Politecnico di Torino, Italy

Correspondence: contact@amodzelewski.com

Abstract

The intentional creation and spread of disinformation poses a significant threat to public discourse. However, existing English datasets and research rarely address the intentionality behind the disinformation. This work presents MALINT, the first human-annotated English corpus developed in collaboration with expert fact-checkers to capture disinformation and its malicious intent. We utilize our novel corpus to benchmark 12 language models, including small language models (SLMs) such as BERT and large language models (LLMs) like Llama 3.3, on binary and multilabel intent classification tasks. Moreover, inspired by inoculation theory from psychology and communication studies, we investigate whether incorporating knowledge of malicious intent can improve disinformation detection. To this end, we propose intent-based inoculation, an intent-augmented reasoning for LLMs that integrates intent analysis to mitigate the persuasive impact of disinformation. Analysis on six disinformation datasets, five LLMs, and seven languages shows that intent-augmented reasoning improves zero-shot disinformation detection. To support research in intent-aware disinformation detection, we release the MALINT dataset with annotations from each annotation step.

1 Introduction

The creation, dissemination, and consumption of online disinformation are increasing concerns, driven by easy access to false content and limited public awareness of its misleading nature (Shu et al., 2020a). The High Level Expert Group, established by the European Commission, defines disinformation as "*false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit*" (de Cock Buning, 2018). Researchers in communication theory further stress the importance of intentionality in defining disinformation (Hameleers, 2023). Uncovering intentions can help future research detect more effectively goal-driven attempts to influence public beliefs (Hameleers, 2023). Although English resources support disinformation research (Wang, 2017; Shu et al., 2020b; Ahmed et al., 2018), none address the varying types of intent behind malicious agents. To fill this gap, we introduce **MALINT**, the first English corpus that annotates disinformation and the most common **MAL**icious **INT**ention types of disinformation agents. The MALINT dataset is a high-quality resource developed in collaboration with fact-checking experts from organizations accredited by the International Fact-Checking Network (IFCN)[1]. We use MALINT to pursue two core objectives.

The first one is **Intent Classification**. We present the first investigation into how well different language models (LMs) can detect malicious intent in English texts. We evaluate small and large language models on binary and multilabel classification tasks. The second objective is **Intent-Augmented Disinformation Detection**. Inoculation theory in psychology suggests that exposing individuals to weakened forms of disinformation can build resistance to deception (Traberg et al., 2022; Roozenbeek et al., 2020). Building on this idea, we explore whether weakening disinformation via integration of malicious intent knowledge can enhance the LLMs' zero-shot disinformation detection. To evaluate it, we propose intent-based inoculation (IBI) and conduct experiments on five established disinformation datasets that include only disinformation labels, as well as on MALINT. In analysis, we use three data splits: (a) a genre-based (articles vs. posts), (b) a temporal split separating texts published before and after the LLMs' knowledge cutoff

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

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3125–3148
March 24-29, 2026 ©2026 Association for Computational Linguistics

dates, and (c) a language split employed to assess the usefulness of intent-based reasoning in a multilingual context, covering even low-resource languages such as Estonian and Polish. We demonstrate that IBI outperforms competitive methods by an average of 9% in English and achieves even larger gains in other languages. In summary, our main contributions are:

- Our MALINT, is the first English corpus annotated for *malicious intent and disinformation* comprising comprehensive stepwise annotations.
- We evaluate malicious intent classification capabilities of 12 different language models.
- By leveraging IBI, we show that intent reasoning improves LLM disinformation detection across diverse datasets and languages.

We release our dataset, prompts, and codebase[2].

2 The MALINT Dataset

MALINT is a novel corpus of online news articles designed to advance research on disinformation and the malicious intents behind it. During annotation, annotators first assess each article's credibility. Articles deemed disinformative are then annotated for the underlying malicious intent of the disinformation agents. This approach is grounded in the recognition that disinformation is deliberately crafted to serve specific malicious objectives (Hameleers, 2023). Credible content, by its nature, is free of such intent. This perspective draws on the growing consensus in disinformation research that malicious intent is a defining feature of disinformative content (Zhou et al., 2022; Appelman et al., 2022; Hameleers, 2023; Wang et al., 2024a).

2.1 Data Sources and Collection

To build a representative dataset, we collected articles from about 50 online sources spanning mainstream media, outlets promoting alternative or incidental narratives. Sources were reviewed by fact-checking experts and classified by consensus into one of three categories: *Reliable*, *Unreliable*, or *Mixed/Biased*. Classification was based on systematic content review and cross-checking with fact-checking tools such as Media Bias/Fact Check[3]. Articles were collected from all sources and subsequently provided for disinformation and malicious intent annotation. To prevent annotation bias, source categories were hidden from annotators. We release the full list of sources.

2.2 Annotation Methodology and Guidelines

A rigorous, multi-stage annotation approach was used to ensure high-quality and consistent annotations of the dataset. The following outlines the creation of guidelines, annotator training, and the step-by-step workflow adopted for article labeling.

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

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

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

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

Credibility Annotation. Each article was reviewed using a methodology that incorporated the debunking technique, complemented by fact-checking principles, as outlined by the NATO Strategic Communications Centre of Excellence (Pamment and Kimber, 2021). Anno-

[2] Repository with data, prompts and codebase: https://github.com/ArkadiusDS/MALINT

[3] An initiative where domain experts perform a careful manual analysis based on clear guidelines (Nakov et al., 2024) Link: https://mediabiasfactcheck.com/

tators assigned one of three labels. Two main annotations are: *Credible Information* and *Disinformation*. We used the definition of disinformation proposed by the European Commission's High-Level Expert Group (de Cock Buning, 2018), widely applied in recent research (Hameleers, 2023; Modzelewski et al., 2024; Sosnowski et al., 2024). Moreover, we introduced an additional annotation label: *Hard-to-say*, for cases where annotators could not agree on veracity. Articles falling into the latter class were excluded from our experiments.

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

Our categories and intent definition are based on the study proposed by Modzelewski et al. (2024) and refined in collaboration with fact-checking experts to better reflect the current disinformation landscape. Figure 1 shows malicious intent definition and detailed descriptions for each category.

2.3 Annotation and Data Quality Control

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

In the first stage, two annotators achieved an agreement of approximately 85.31% on the credibility task. They reached 65.19% agreement on the more complex multilabel intent task. These figures reflect pre-consensus agreement and indicate how challenging it was to prepare the final consensus annotation rather than measuring the final label quality (Piskorski et al., 2023).

In the second stage, the supervisor performed a third annotation. Disagreements were resolved through consensus, and expert input was utilized when necessary. At this stage, the annotation process concluded with agreement exceeding 95% between supervisors for both tasks. This improved the reliability and quality of the dataset. If consensus could not be reached for credibility analysis, the article was assigned a *Hard-to-say* label and excluded from further steps.

MALICIOUS INTENT

Malicious intent is a generalization of a disinformation narrative that we can define as a repeating pattern found in several disinformative articles. Malicious intent encapsulates the broader goal of the author, which guides the specific narratives used to achieve that goal (Modzelewski et al., 2024). In our study, we categorize the following types of malicious intent:

Undermining the Credibility of Public Institutions [UCPI]

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

Changing Political Views [CPV]

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

Undermining International Organizations and Alliances [UIOA]

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

Promoting Social Stereotypes/Antagonisms [PSSA]

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

Promoting Anti-scientific Views [PASV]

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

Figure 1: Definition and categories of malicious intent.

Note: We publish annotations from each stage.

2.4 Dataset Statistics

Table 1 presents the key statistics of the **MALINT** dataset, which consists of 1,600 news articles, providing a substantial corpus for analysis.

Credibility and Malicious Intent Labels. The dataset includes two primary credibility labels: *Credible*, comprising 63.5% instances, and *Disinformation*, which accounts for the remaining 584 articles (36.5%). Table 2 details the distribution of the five malicious intent categories in dataset.

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

Table 1: Overview of the **MALINT** corpus.

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

Table 2: Malicious intent types distribution in MALINT.

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

3 Intent Classification

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

- **Binary Detection Per Class** - Models are evaluated on their ability to detect the presence of a specific malicious intent. Each intent is treated as an independent binary classification problem. This setup allows us to analyze how well models can identify individual intent categories.
- **Multilabel Detection** - Evaluating a model's ability to simultaneously identify multiple intent types that may co-occur in a given input. The task is formulated as a multilabel classification problem, where models must assign all relevant intent labels to each input.

As shown in Table 2, our tasks are challenging due to class imbalance across intent categories.

3.1 Experimental Setup

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

Setup for SLMs. We fine-tuned a range of pre-trained SLMs, selected to represent different architectures and computational requirements. We have chosen *BERT* (Devlin et al., 2018), *RoBERTa* (Liu et al., 2019), *DeBERTa V3* (He et al., 2021b,a) and *DistilBERT* (Sanh et al., 2019). DistilBERT was included to assess the model suitable for environments with limited computational resources. Each model was fine-tuned for the two tasks across 42 model-task combinations, testing multiple hyperparameter settings, totaling around 2,000 experiments. All experiments were run on an NVIDIA L40 GPU. Details and the optimal hyperparameters for each experiment are provided in Appendix B.

Setup for LLMs. We evaluated five cutting-edge LLMs via different APIs: *GPT 4o Mini*, *GPT 4.1 Mini*, *Gemini 2.0 Flash*, *Gemma 3 27b it*, *Llama 3.3 70B*. To ensure as deterministic results as possible, we prompted all models with the temperature parameter set to zero. All evaluations were conducted in a zero-shot setting, as many documents were too long for few-shot prompting within the LLM context limits. The full set of experiments involved approximately 15,000 API calls. Prompts for all tasks are provided in Appendix D.

Further details regarding APIs used and each model, including their knowledge cut-off dates and the rationale for their selection, are available in Appendix C.

Baselines. We implemented two baselines for all tasks: a random classifier and logistic regression. For binary classification, the logistic regression model was trained on bag-of-words features represented as sparse token count vectors, with English stop words removed. For the multilabel task, logistic regression was applied using a one-vs-rest strategy (Murphy, 2018).

3.2 Evaluation Results

Binary Detection Per Each Class. As shown in Table 3, DeBERTa V3 Large and RoBERTa models consistently achieved the highest F_1 scores across most intent categories among SLMs. Among LLMs, GPT 4.1 Mini performed best for *UCPI* and *PSSA*, while Llama 3.3 70B for *PASV* and *CPV* categories. LLMs achieved superior results compared to fine-tuned SLMs across three intent categories. The logistic regression baseline with bag-of-words outperformed random predictions but remained below most LMs, serving as a simple baseline.

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

Table 3: LMs' F_1 scores on binary intent classification across five categories, compared to random and BoW logistic regression baselines.

Multilabel Detection. We show results for this task in Table 4. DeBERTa V3 and RoBERTa continued to perform best among the SLMs, achieving the highest weighted F_1 scores. The best-performing LLM, LlaMA 3.3 70B, lagged noticeably behind the top SLMs. Surprisingly, the logistic regression baseline (using a one-vs-rest strategy) outperformed most LLMs. However, fine-tuned SLMs demonstrated superior ability to capture the complexity of co-occurring intent labels, underscoring the effectiveness of supervision.

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

Table 4: Performance of LMs on multilabel intent classification compared to baselines: a random classifier and a one-vs-rest logistic regression using BoW approach.

4 Intent-Augmented Disinformation Detection

Inoculation theory, introduced by McGuire (1964), uses a biological metaphor. It suggests that, just as people can be protected against viruses through vaccines, they can also be "vaccinated" to resist persuasive messages (McGuire, 1964). An inoculation message has two parts: a *threat* and *refutational preemption*. The threat alerts individuals that a persuasive attack is coming (Lewandowsky et al., 2017). Refutational preemption (or prebunking) involves providing people with arguments or tools to resist persuasive attacks, helping them better recognize and respond to such attempts (Pfau et al., 2005). Building on this theory and its applicability to improve disinformation detection (Traberg et al., 2022), we pose the following research question: *Does inoculating LLMs against malicious intent improve their disinformation detection performance in a zero-shot setting?*

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

4.1 Datasets

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

- **ISOT FakeNews**: Thousands of real/fake news articles from reputable sources and sites flagged by PolitiFact[4] (Ahmed et al., 2018, 2017).
- **CoAID**: COVID-19 misinformation dataset with news and social media posts (Cui and Lee, 2020).
- **EUDisinfo**: The latest English disinformation dataset collected from the EUvsDisinfo database[5] (Modzelewski et al., 2025).
- **ECTF**: COVID-19 fake post detection dataset from Platform X (Twitter) (Bansal et al., 2021).
- **EUvsDisinfo**: Multilingual EUvsDisinfo's texts with pro-Kremlin propaganda (Leite et al., 2024).

We used five datasets (including MALINT) to assess the usefulness of intent-based reasoning across

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

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

genres and temporal splits in English. To evaluate its cross-lingual generalizability, we also used the EUvsDisinfo texts, splitting it into six language-specific datasets: German, French, Polish, Estonian, Russian, and Spanish.

4.2 Intent-based Inoculation Design

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

$$X_T = (T, K_I, G_A) \tag{1}$$

The model M produces a structured output:

$$A_I(T) = \{i_j : (r_{i_j}, R_{i_j}) \mid i_j \in I\}, \tag{2}$$

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

Formally, we define:

$$A_I(T) \sim M(X_T) = M(T, K_I, G_A). \tag{3}$$

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

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

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

$$Z_T = (T, \theta, A_I(T), G_I). \tag{4}$$

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

$$\hat{y}_T \sim M(Z_T) = M(T, \theta, A_I(T), G_I). \tag{5}$$

This design of experiments allows us to answer our research question. By integrating the threat component and a LLM-generated refutational preemption into the prompt, the IBI leverages ideas built upon inoculation theory to prepare the model for detecting disinformative content.

4.3 Experimental Setup

We created five test sets by randomly sampling about 400-500 texts from each of the datasets. Moreover, we sampled approximately 3,000 texts from EUvsDisinfo, creating test sets of about 500 per language across six languages. Table 5 reports class proportions across all test sets.

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

Table 5: Class distribution across test datasets. The percentages represent the original distribution of disinformation and credible content within each dataset.

Our experiments were conducted on the same five LLMs that we used for experiments in section 3. In these experiments, we again set the temperature hyperparameter to zero for all models. We focused on zero-shot settings with LLMs to evaluate the effectiveness of the IBI when texts lack explicit information about malicious intent. Additionally, prior studies show that LLMs can outperform supervised models such as BERT in disinformation detection (Pelrine et al., 2023; Bang et al., 2023). Lucas et al. (2023) also found that fine-tuned BERT models perform worse on unseen data compared to zero-shot with LLMs, which was later confirmed by Modzelewski et al. (2025).

To thoroughly evaluate IBI, we used two data splits on English data: (a) a genre-based split, separating long-form news articles from social media posts (from Platform X, formerly Twitter), and (b) a temporal split, comparing texts published before and after the LLMs' knowledge cutoffs. The temporal split is possible because EUDisinfo and MALINT include post-cutoff articles not seen during model training. Following Lucas et al. (2023), this setup enables a rigorous evaluation of IBI across two genres and, crucially, on unseen data, providing a more realistic test of its generalization. IBI was further evaluated on six languages to highlight its cross-lingual generalization.

In our study, we compare IBI to three competitive methods that were best on human-annotated datasets from study by Lucas et al. (2023). Below short description of chosen methods:

- *VaN* – A minimal baseline prompt with direct instructions to the LLM (Lucas et al., 2023).

- *Z-CoT* – Extends *VaN* with zero-shot chain-of-thought reasoning (Kojima et al., 2022).
- *DeF-SpeC* – Emphasizes contextual, deductive, and abductive reasoning (Lucas et al., 2023), improving multi-step reasoning (Bang et al., 2023).

To learn more about these methods, prompts, and how we adapted them to our IBI, see Appendix E.

We use F_1 for the positive class as evaluation metric. To assess significance between IBI and baselines, we used McNemar's test (see section 4.4 and Appendix F for details), a standard method for comparing two models on the same binary task (Dietterich, 1998; Dror et al., 2018), widely used in NLP (Blitzer et al., 2006; Card et al., 2020).

4.4 Results and Discussion

MALINT: Dataset-Specific Results Table 7 compares baseline prompting strategies (Base) with their intent-based inoculation (IBI) counterparts on the MALINT dataset. Across all models, IBI consistently improves disinformation detection, with average gains ranging from around 2% for GPT-4o Mini up to 8% for Gemini 2.0 Flash.

We analyzed the effect of intent prediction correctness on disinformation detection for each intent type separately. For each model and intent type, we computed F_1 scores separately for instances where the intent was predicted correctly versus incorrectly. This allowed us to measure the potential impact of separate intent knowledge on the model's ability to detect disinformation with IBI. The results show that correct intent predictions generally improve F_1 scores, indicating that accurate intent understanding boosts disinformation detection with IBI. Some exceptions exist, such as UIOA for certain models, where higher F_1 scores with IBI occur even when intent is mispredicted, suggesting that other correctly predicted intents can partially compensate.

Appendix H shows the ablation on intent's influence on disinformation prediction, and remaining dataset-specific results are in Appendix G.

Genre and Temporal Split Results Table 8 compares the baseline prompting strategies (Base) with their IBI counterparts. Across 75 evaluation scenarios (5 models × 3 prompting strategies × 5 settings: overall comparison, articles, social media posts, prior-cutoff, and post-cutoff), IBI leads to improved performance in approximately 90% of cases. On average, the overall performance increases by 9%. McNemar's test indicates that, in nearly all scenarios, IBI significantly outperforms the baselines on the overall dataset at the 0.01 significance level. For Llama 3.3 70B, the difference is still significant, though at the 0.05 level.

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

Notably, IBI improves performance not only on data that may have been present during LLM pretraining, but also on unseen content published after the models' knowledge cutoffs. While performance gains are evident in both pre- and post-cutoff subsets, LLMs show greater difficulty with the latter.

Overall, these findings support our central hypothesis: incorporating intent-based inoculation enhances zero-shot disinformation detection. Improvements are consistent across data types, temporal settings, and prompting strategies.

Multilingual Evaluation Table 6 shows the averaged F_1 scores across all models and methods for each language on the EUvsDisinfo dataset. IBI consistently improves performance over the Base setup for all six languages, with the largest gains observed in Estonian. These results indicate that intent-augmented reasoning enhances model disinformation detection capabilities across languages. Overall, IBI demonstrates a clear advantage in cross-lingual disinformation detection, achieving on average a 20% improvement over baseline methods. More detailed results in Appendix I.

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

Table 6: Comparison of Base vs IBI performance across languages. Results present averaged F_1 scores over all models and methods. See Appendix I for the table with standard deviations.

5 Related Work

Intent. Intent (or *intention*) discovery is multifaceted problem addressed from different perspectives. It can be purely textual (Xu et al., 2023) or multimodal, where text is analyzed alongside images (Kruk et al., 2019) or videos (Maharana et al., 2022). Some studies focus on uncovering the relationship between intention and behavior (Conner and Norman, 2022) or how intentions guide people

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

Table 7: F_1 scores on MALINT for competitive prompting methods and their improvement with IBI.

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

Table 8: Results with F_1 scores for five LLMs. The *Base* columns shows the competitive method results, while the *IBI* columns presents results for prompts adapted to the Intent-based Inoculation.

to achieve a goal (von Suchodoletz and Achtziger, 2011), while others aim to identify and categorize intentions. Hameleers (2022) defines a conceptualization that connects actors, intentions, and techniques for creating and disseminating disinformation content. They identify four literature-based categories of intentions: *delegitimization*, *mobilization*, *ideological motivations*, and *financial gain*. They present a framework that can be used to conceptualize disinformation, but they do not annotate a dataset nor provide a baseline for intention detection and any experiments with LMs. Wang et al. (2024b) use public datasets of real/fake articles and annotate them using a T5-based model with agent intent classes: *Public*, *Emotion*, *Individual*, *Popularize*, *Clout*, *Conflict*, *Smear*, *Bias*, *Connect*. Moreover, they show that incorporating intent features improves misinformation detection performance in their T5-based framework.

Gupta et al. (2021) explore user intent behind a query and proposes a fact-checking chatbot to counter the spread of fake news, but does not investigate whether intent could be leveraged to improve disinformation detection. Instead, Zhou et al. (2022) focuses on the intent of spreading fake news. After labelling a dataset algorithmically, they evaluate the effectiveness of employing user propagation intent to detect fake news using Heterogeneous Graph Neural Networks, achieving a slight improvement concerning the state-of-the-art. They also found that most fake news spreaders don't do it intentionally, so focusing on disinformation agents' intent could be more meaningful.

Disinformation. Prior work on text-based disinformation detection has relied mainly on supervised classification of false or misleading content. Early datasets include LIAR (Wang, 2017), with short statements labeled for veracity, and FakeNewsNet (Shu et al., 2020b), which links news with social and temporal context. Most systems model disinformation as binary classification (real vs. fake), using fine-tuned models like BERT (Khan et al., 2021). Surveys highlight that fake news is intentionally misleading (Shu et al., 2017). As a result, RMDM (Nguyen et al., 2023), a Vietnamese dataset, adds four labels (real, misinfo, disinfo, malinfo) to separate unintentional errors from deliberate harm. The only dataset annotated with intent and disinformation was introduced by Modzelewski et al. (2024). However, the dataset is limited to Polish, and their study reports only simple baseline results for intent detection in that language. Moreover, recent work shows that pretraining on related tasks (e.g., fine-grained sentiment) can enhance disinformation detection (Pan et al., 2024).

To the best of our knowledge, MALINT is the first human-annotated English dataset to label disinformation and malicious intents of disinformation agents (see Appendix J for full comparison to other datasets). We are the first to evaluate intent classification with SLMs and LLMs, establishing baselines for intent detection in English. Our study also demonstrates, for the first time, the utility of intent-augmented reasoning with different LLMs for zero-shot disinformation detection.

6 Conclusions

In this study, we present **MALINT**, the first human-annotated English-language dataset that includes disinformation and malicious intent annotations. MALINT is a high-quality dataset created with accredited disinformation and fact-checking experts.

Our research provides the first systematic evaluation of malicious intent identification, comparing seven fine-tuned SLMs (e.g., BERT) and zero-shot with LLMs across binary and multilabel settings. In the multilabel setup, SLMs outperform LLMs, reaching a weighted F_1 of 82.1%. In contrast, binary classification reveals that LLMs outperform fine-tuned SLMs on three intent categories.

Inspired by inoculation theory from psychology and communication studies, we designed an experiment on intent-based inoculation with LLMs. Our results show that exposure to knowledge about malicious intent significantly enhances disinformation detection performance in a zero-shot setting. IBI improves performance by 9% on average across five English datasets and LLMs, and enhances disinformation detection in other six languages.

Limitations

Datasets and Annotation. The MALINT dataset features five malicious intent categories, in addition to a binary classification for disinformation. While this framework provides a comprehensive representation, we acknowledge that the dataset may not exhaustively cover all possible intents. While this taxonomy offers a broad and representative framework, we acknowledge that it may not capture the full spectrum of possible intent types. Our categorization draws on prior research (Modzelewski et al., 2024) as well as official reports from an independent non-profit organisations, like EU DisinfoLab (e.g. Sessa (2023)). Furthermore, the taxonomy was developed in close collaboration with experts in fact-checking and debunking, ensuring informed coverage of malicious intent types that are especially relevant to the current global information landscape.

Biases. Human annotation inherently involves some degree of subjective interpretation. In order to mitigate this issue, annotation was conducted under the supervision of professional fact-checkers. Annotators underwent extensive training, following the detailed annotation guidelines developed with experts. Each article was independently annotated by each annotator and subsequently reviewed by a supervisor. When clarification was required, the supervisor consulted the original annotator to ensure consistency and accuracy. For what concerns the other datasets employed in this study, they inevitably retain any biases present in their original annotation process.

Experiments. The intent-based reasoning experiments conducted in this study rely on the fixed taxonomy of malicious intents adopted in the proposed dataset. Although this specific framework may contain inherent bias, our categorization was rigorously validated by professional fact-checkers. It builds on established research (Modzelewski et al., 2024) and incorporates insights from independent non-profit organizations such as EU DisinfoLab (e.g. Sessa (2023)), ensuring both empirical grounding and practical relevance.

Although we worked with a limited set of models, we deem this selection adequately diverse. It includes both Small and Large Language Models, incorporating open-source and closed models and covering multiple providers such as OpenAI, Google and Meta.

Ethics

Dataset and Annotation. The MALINT dataset comprises data extracted from publicly accessible online sources and is free from copyright restrictions. It does not contain personally identifiable information and is intended solely for research applications. The dataset will be released under the CC BY 4.0 license.

No crowdsourcing was used at any stage of the annotation process. All annotators were hired by affiliated institutions and received appropriate compensation. The annotation process was designed to ensure fair and unbiased work practices, with expert oversight throughout and approval obtained from the relevant ethics board. We also made efforts to maintain gender balance among annotators to promote diverse perspectives.

Our research aims to support society by helping fact-checkers, researchers, and public-interest efforts. However, the dataset and insights we provide could also be misused. For example, malicious actors might study the MALINT dataset to understand detection patterns and create more convincing or harder-to-spot disinformation. We strongly encourage ethical use. Any system built using our data should include transparency, oversight, and safeguards to reduce the risk of abuse.

Computational resources. The deployment of large language models raises environmental concerns, due to significant computational requirements. Our methodology mitigates these issues by focusing on model inference rather than training from scratch, thereby substantially reducing computational demands. Our experiments primarily relied on third-party API services, with computational resources managed by external providers while fine-tuning was conducted only on small language models. All computing resources were provided by university and reserved exclusively for research purposes.

Acknowledgements

This research was supported by the projects: Infotester4Education (full title: Development and implementation of AI education methods and digital tools supporting tackling disinformation, number: 2023-2-PL01-KA220-HED-000180856) within the framework of Cooperation Partnership for Higher Education, ERASMUS+, and EUonAIR project (number 101177370, ERASMUS-EDU-2024-EUR-UNIV-1), within the framework of the EUonAIR Centre of Excellence in Responsible AI in Education, co-funded by the European Commission.

Giovanni Da San Martino would like to thank the Qatar National Research Fund, part of Qatar Research Development and Innovation Council (QRDI), for funding this work by grant NPRP14C0916-210015. He also would like to thank the European Union under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of March 15, 2022 of Italian Ministry of University and Research – NextGenerationEU; Code PE00000014, Concession Decree No. 1556 of October 11, 2022 CUP D43C22003050001, Progetto "SEcurity and RIghts in the CyberSpace (SERICS)" - Spoke 2 Misinformation and Fakes - DEcision supporT systEm foR cybeR intelligENCE (Deterrence) for also funding this work.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9.

Naomi Appelman, Stephan Dreyer, Pranav Manjesh Bidare, and Keno C Potthast. 2022. Truth, intention and harm: Conceptual challenges for disinformation-targeted governance. *Internet Policy Review*.

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. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718.

Rachit Bansal, William Scott Paka, Nidhi, Shubhashis Sengupta, and Tanmoy Chakraborty. 2021. Combining exogenous and endogenous signals with a semi-supervised co-attention network for early detection of covid-19 fake tweets. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 188–200. Springer.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9263–9274.

Mark Conner and Paul Norman. 2022. Understanding the intention-behavior gap: The role of intention strength. *Frontiers in Psychology*, 13.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1383–1392.

European Commission. 2022a. Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions on the european democracy action plan. *Publications Office of the European Union*. Available at: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0790`.

European Commission. 2022b. The strengthened code of practice on disinformation 2022. *Publications Office of the European Union*. Available at: https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

Ankur Gupta, Yash Varun, Prarthana Das, Nithya Muttineni, Parth Srivastava, Hamim Zafar, Tanmoy Chakraborty, and Swaprava Nath. 2021. Truthbot: An automated conversational tool for intent learning, curated information presenting, and fake news alerting. *Preprint*, arXiv:2102.00509.

Michael Hameleers. 2022. 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.

Michael Hameleers. 2023. 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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in Instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632, Hong Kong, China. Association for Computational Linguistics.

João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2024. Euvsdisinfo: a dataset for multilingual detection of pro-kremlin disinformation in news articles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5380–5384.

Stephan Lewandowsky, Ullrich KH Ecker, and John Cook. 2017. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition*, 6(4):353–369.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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

Adyasha Maharana, Quan Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, and Mohit Bansal. 2022. Multimodal intent discovery from livestream videos. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 476–489, Seattle, United States. Association for Computational Linguistics.

William J McGuire. 1964. Inducing resistance to persuasion. some contemporary approaches. *CC Haaland and WO Kaelber (Eds.), Self and Society. An Anthology of Readings, Lexington, Mass.(Ginn Custom Publishing) 1981, pp. 192-230.*

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

Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. 2025. Pcot: Persuasion-augmented chain of thought for detecting fake news and social media disinformation. *arXiv preprint arXiv:2506.06842*.

Kevin P Murphy. 2018. Machine learning: A probabilistic perspective (adaptive computation and machine learning series). *The MIT Press: London, UK*.

Preslav Nakov, Jisun An, Haewoon Kwak, Muhammad Arslan Manzoor, Zain Muhammad Mujahid, and Husrev T Sencar. 2024. A survey on predicting the factuality and the bias of news media. In *ACL (Findings)*.

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

James Pamment and Anneli Lindvall Kimber. 2021. *Fact-checking and debunking: a best practice guide to dealing with disinformation*. NATO Strategic Communication Centre of Excellence.

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

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

Michael Pfau, Bobi Ivanov, Brian Houston, Michel Haigh, Jeanetta Sims, Eileen Gilchrist, Jason Russell, Shelley Wigley, Jackie Eckstein, and Natalie Richert. 2005. Inoculation and mental processing: The instrumental role of associative networks in the process of resistance to counterattitudinal influence. *Communication Monographs*, 72(4):414–441.

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

Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. 2020. Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *The Harvard Kennedy School (HKS) Misinformation Review*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

MG Sessa. 2023. Connecting the disinformation dots: Insights, lessons, and guidance from 20 eu member states. *EU Disinfolab, Friedrich Naumann Foundation for Freedom*.

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

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

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

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

Cecilie S Traberg, Jon Roozenbeek, and Sander van der Linden. 2022. Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, 700(1):136–151.

Antje von Suchodoletz and Anja Achtziger. 2011. Intentions and their limits. *Social Psychology*, 42:85–92.

Bing Wang, Ximing Li, Changchun Li, Bo Fu, Songwen Pei, and Shengsheng Wang. 2024a. Why misinformation is created? detecting them by integrating intent features. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2304–2314.

Bing Wang, Ximing Li, Changchun Li, Bo Fu, Songwen Pei, and Shengsheng Wang. 2024b. Why misinformation is created? detecting them by integrating intent features. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2304–2314, New York, NY, USA. Association for Computing Machinery.

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

Hua Xu, Hanlei Zhang, and Ting-En Lin. 2023. *Intent Recognition*, pages 7–29. Springer Nature Singapore, Singapore.

Xinyi Zhou, Kai Shu, Vir V. Phoha, Huan Liu, and Reza Zafarani. 2022. "this is fake! shared it by mistake":assessing the intent of fake news spreaders. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3685–3694. ACM.

A Annotatin Guidelines

Our methodology and annotation guidelines were designed to standardize the assessment of articles for disinformation content, aiming to reduce subjectivity and enable comprehensive analysis. Utilizing these annotation guidelines, we analyzed numerous articles to identify disinformation. The methodology was developed in cooperation with analysts (fact-checking and debunking experts) employed in the project based on their experience as experts, scientific knowledge available on the subject, and the experience of other institutions and organizations involved in research and detection of disinformation. The methodology improved throughout the project and subsequent testing to best reflect the disinformation environment. All authors of this methodology have at least three years of experience working for fact-checking or debunking organizations accredited by the International Fact-Checking Network. Moreover, our methodology and annotation guidelines draw on similar work on the annotation of disinformation, such as the guidelines presented by Modzelewski et al. (2024).

Main Assumptions of the Methodology. Creating a uniform methodology and guidelines aims to guarantee the quality of the assessments made by annotators and minimize their subjectivity.

The analysis of articles is carried out mainly via the debunking technique, with the auxiliary use of the fact-checking technique. These terms for this methodology are defined in a manner analogous to the methodology developed for the NATO Strategic Communication Centre of Excellence (Pamment and Kimber, 2021). Fact-checking is the long-standing process of checking that all facts in a piece of writing, news article, or speech are correct. Debunking refers to exposing falseness or manipulating systematically and strategically (based on a chosen topic, classifications of selected techniques, narrative).

Preparation of Articles for Evaluation. The first step is to select web portals from which articles on particular topics will be taken. Among them are both mainstream media and those presenting the alternative current. This is to ensure access to enough reliable as well as unreliable content. Each portal will be assigned to one of three categories, determining its credibility. This will be done by a team of experts by consensus. Assessing the credibility of a website requires an in-depth analysis of the content posted on it regularly, as well as checking it in reliable sources, including via the Media Bias/Fact Check search engine. The source's rating will not be visible to annotators. The analysis consists in selecting the category that best suits a given domain:

- **Reliable** - sources that are reliable/publishing reliable content on a specific topic, in particular traditional news portals.
- **Unreliable** - sources publishing unreliable content, typically disinformation, e.g., all domains financed by the Kremlin, sites containing conspiracy theories, etc.
- **Mixed/Biased** - partially or potentially biased websites that may present false information on specific issues, e.g., typically political websites, and blog collections.

Thematic Category. Before the analysis begins, articles will be assigned to eight topics. This will be done manually with the help of keywords through searches on selected web portals. Thematic categories were pre-defined. The selection of topics was based on EU DisinfoLab's cross-cutting report on disinformation in Europe (Sessa, 2023). It is based on expert studies from 20 countries.

- Anti-Europeanism and anti-Atlanticism (anti-EU, anti-NATO)
- Anti-migration and xenophobia
- Climate change and the energy crisis
- Health (including COVID-19 and vaccines)
- Institutional and media distrust (public institutions)

- Gender-based disinformation
- Ukraine war and refugees
- Disinformation about LGBTQIA+

Content Analysis. The next step requires analyzing the entire article's content and recognizing whether the information is accurate or disinformative. If the article provides only factual information, it is marked as "credible information." Selecting this category ends the assessment of the article. When information in the article is unreliable and misleads the recipients, content is considered disinformative. The unintentional dissemination of false information is known as misinformation. However, even unintentional dissemination of false information without the goal of manipulating recipients can fuel disinformation. Disinformation is particularly difficult to detect as the author's intention is usually unspecified, and in most cases, it can only be presumed. Therefore, for this study, we assume that any form of false or manipulative information is considered disinformation.

For these guidelines, the definition of disinformation provided by the European Commission High-Level Group of Experts on False News and Disinformation on the Internet (HELG) will be used, as it covers all four aspects and does not exclude potentially harmful content presented in the form of political advertising or satire, as presented in the EU Code of Practice. The definition is as follows (de Cock Buning, 2018):

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

However, a necessary supplement to this definition is taking into account the European Union Code of Practice on Disinformation, according to which disinformation is defined as: "verifiable false or misleading information which, cumulatively, (a) is created, presented and disseminated for economic gain or to intentionally deceive the public; and (b) may cause public harm, intended as threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment, or security". (European Commission, 2022b). The detected information must be verifiable, which means that it can be proved untrue, and, therefore, it cannot be, for example, a yet unproven theory or opinion, as long as it is not intended to mislead the recipients. In summary, disinformation is intentionally misleading by providing misleading or false information (European Commission, 2022a). Unlike disinformation, misinformation is *misleading information shared by people who do not recognize it as such* (de Cock Buning, 2018). However, as noted earlier, misinformation and disinformation are treated as a single category under "disinformation."

When a given content is not verifiable (reliable/-disinformative), it is marked as the "Hard to say" category. Indicating this category ends the assessment. Below, we present the main categories:

- Credible information
- Disinformation
- Hard to say

Annotation of Malicious Intent. The study of the malicious intentions of the disinformation content creators is potentially the most subjective element of the analysis, and therefore it is particularly important to develop precise components of the assessment. This allows for maintaining uniformity of the analysis carried out by different annotators.

In this methodology, understanding the intention behind disinformation is crucial for effectively analyzing it. Disinformation, according to our definition, is always spread intentionally, emphasizing the significance of comprehending the motives driving its dissemination. It encapsulates the broader goal of the author, which guides the specific narratives they employ to achieve that goal. Authors of disinformation have some purpose in creating it. It is in this category that we try to answer the question: what is the purpose of spreading disinformation by a particular author? The task type is defined as an exhaustive list with multiple choice options (multilabel). Below are possible choices:

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

- **Changing political views** - Influencing voter preferences is a common procedure used by disinformation authors. Changing political beliefs is aimed at strengthening one side of a political dispute and arousing resentment against the others. It usually involves the simultaneous promotion of politicians from extremist movements, which are treated as an alternative to the major parties. It is often based on the portrayal of mainstream politicians as corrupt and evil to the bone (e.g., portraying them as traitors to the nation, dependent on the outside influence of global elites).
- **Undermining international organizations and alliances** - Undermining the credibility of international institutions is often part of disinformation activities carried out by external forces (e.g., Russia). These are aimed at breaking up alliances of democratic states to facilitate propaganda efforts by authoritarian states (e.g., portraying NATO as an aggressor that will drag peaceful states into war). Of course, numerous extreme political movements also have an interest in shattering trust in international institutions. This is part of a populist influence on society and a way to gain power. International institutions are then most often portrayed as entities that take away the sovereignty of member states (e.g., presenting the EU as an authoritarian organization that imposes its will on others).
- **Promoting social stereotypes/antagonisms** - Deepening social divisions is a frequent goal of disinformation efforts. A strongly divided society is less resistant to manipulation, and mutual distrust also promotes a collapse of confidence in the institution of the state and democracy. This causes internal problems to absorb most of the attention, giving room for external centers of influence to operate. This can take the form of reinforcing xenophobia (e.g., stirring up resentment against Ukrainian refugees and portraying them as dangerous). Aversion to specific social groups can also be exploited (e.g., portraying homosexuals as pedophiles).
- **Promoting anti-scientific views** - Science is a frequent enemy of disinformation authors. Science enhances critical thinking and is an important part of the strength of Western democracies. Presenting it as an enemy aids in undermining the system under which Western countries operate. Reinforcing anti-scientific attitudes also enables short-term financial gain (e.g., selling pseudo-medical remedies for various diseases). The fight against science can be based on a direct attack on scientists (e.g., the claim that vaccines are designed to depopulate humanity), but is also a significant element of conspiracy theories (e.g., medicine is not used to cure people, but to make money).

Double Evaluation and Consensus Establishment According to this methodology, all content must undergo a double evaluation. Articles are evaluated two times by two annotators, working independently of each other. The first is the student, and the second is the supervisor. The supervisor does not read the first performed assessment, but only evaluates the content according to the methodology, independently of the results of the first evaluation. The supervisor then compares the two performed assessments and makes the final decision on the choices made in the analysis process. Discrepancies spotted by the double-verification analyst are discussed by the team. Then, a common, consistent approach to content classification is established. When necessary, the lead annotator, an expert in fact-checking and debunking, can be consulted to discuss the evaluation. The final registered assessment is therefore a consensus based on the first and second assessment, and can include elements of both independent evaluations. The purpose of double verification is therefore not only to avoid the human errors but also to the standardization of the methodology's application.

B Small Languages Models and Hyperparameters for Malicious Intent Classification Tasks

This appendix provides details on the experimental setup and optimal hyperparameters used for multilabel classification of malicious intent categories. All models were trained and evaluated using the same dataset splits, with five intent categories: *Undermining the Credibility of Public Institutions* (UCPI), *Changing Political Views* (CPV), *Undermining International Organizations and Alliances* (UIOA), *Promoting Social Stereotypes/Antagonisms* (PSSA), and *Promoting Anti-scientific Views* (PASV).

Tokenization and Data Preprocessing For all models and tasks, we used the Hugging Face Transformers library to load both the tokenizer and the model. The input content was tokenized with the following settings:

- Truncation: True
- Padding: True
- Max Length: 256 tokens

Training and Hyperparameter Search The training procedure involved feeding data into a training loop using the `Trainer` API. Hyperparameter tuning was performed over the following grid:

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

B.1 Binary Detection Per Each Class.

All models were trained for 5 epochs with model checkpointing based on the F_1 score over the positive class. Other fixed hyperparameters included:

- Batch size (train/eval): 8
- FP16 training: Enabled
- Evaluation strategy: every 50 steps
- Save total limit: 2 checkpoints
- Load best model at end: True

Optimal Hyperparameters per Model The best-performing hyperparameter configuration for each model and each binary detection task is listed in Table 9. Identifiers of all models given in Table 9 as of 21.07.2025. All models were evaluated using F_1 over the positive class.

B.2 Multilabel Detection.

All models were trained for 5 epochs with model checkpointing based on the macro-weighted F_1 score. Other fixed hyperparameters included:

- Batch size (train/eval): 4
- FP16 training: Enabled
- Evaluation strategy: every 50 steps
- Save total limit: 2 checkpoints
- Load best model at end: True

Optimal Hyperparameters per Model The best-performing hyperparameter configuration for each model is listed in Table 10. Identifiers of all models given in Table 10 as of 21.07.2025. All models were evaluated using macro-weighted F_1 to account for class imbalance and the multilabel nature of the task.

C Overview of LLMs from Experiments

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

Table 11 lists the Large Language Models used in our experiments, detailing their knowledge cutoff dates, access methods, licenses, and sizes. To enable evaluation on both prior and post cutoff content, we used a knowledge cutoff date of September 2024 to split the EUDisinfo and MALINT datasets accordingly. All other datasets contain only texts published prior to this date.

D Prompts for Malicious Intent Classification Tasks

D.1 Binary Detection Per Each Class.

To perform binary detection of specific malicious intent categories, we designed a prompt that conditions the LLM on a single target intent and asks for a strict `Yes`/`No` decision. The model is instructed to be conservative and only respond `Yes` when confident (see Figure 2).

D.2 Multilabel Detection.

For the multilabel setting, we use a single prompt that asks the model to evaluate the presence of all five malicious intent categories simultaneously. The model provides a `Yes`/`No` decision for each category independently and is instructed to be conservative in its judgments (see Figure 3).

E Baseline Methods and Prompts used for Intent-Based Inoculation Experiments

E.1 Baseline Methods

For the disinformation detection stage, we selected three strong baseline methods identified by Lucas et al. (2023) as top performers, particularly on human-annotated datasets such as CoAID (Cui and Lee, 2020) and FakeNewsNet (Shu et al., 2020b). The selected methods are as follows:

- *VaN* – A basic prompt that provides minimal, direct instructions to the LLM, serving as a foundational baseline (Lucas et al., 2023).
- *Z-CoT* – Builds on *VaN* by encouraging step-by-step reasoning, following the zero-shot chain-of-thought prompting strategy introduced by Kojima et al. (2022).
- *DeF-SpeC* – A more advanced prompt designed to elicit contextual, deductive, and abductive reasoning (Lucas et al., 2023), addressing limita-

Binary Detection Per Each Class

System Prompt:
You are an assistant who detects malicious intention in text. Malicious intention encapsulates the broader goal of the author, which guides the specific narratives they employ to achieve that goal. Your expertise lies in detecting one malicious intention:
<*Here name of the malicious intent category [shortcut]*>
User Prompt:
Analyze the text and decide if the text contains any malicious intention: Undermining the credibility of public institutions [UCPI] You are very conservative in your final decisions and when you are not fully sure you answer No.
Give your answer in the form of dictionary:

```
{
    "<[shortcut]>": "Your answer if text include <Here name of the malicious intent category> intent. Use only Yes or No"
}
```

Text: <*Text to analyze*>

Figure 2: Prompt template used for binary classification of malicious intent categories with LLMs. In each instance, placeholders <*Here name of the malicious intent category*> and <*[shortcut]*> were replaced with one of the following categories and their respective abbreviations: Undermining the Credibility of Public Institutions [UCPI], Changing Political Views [CPV], Undermining International Organizations and Alliances [UIOA], Promoting Social Stereotypes/Antagonisms [PSSA], and Promoting Anti-scientific Views [PASV].

tions in LLMs' ability to perform multi-step or inductive inference (Bang et al., 2023).

We used these methods as baselines to evaluate the effectiveness of our intent-based inoculation (IBI) approach (Figure 4 illustrates the baseline prompt template). To adapt them to our setting, we modified the original prompts to incorporate malicious intent analysis, as introduced in the first stage of our pipeline (Section 4.2). This allowed us to examine whether IBI is robust across different prompting strategies or sensitive to prompt.

Our evaluation was conducted on five datasets spanning multiple domains and genres, including news and social media. This diversity ensures a broad test of IBI's generalizability and enables a meaningful comparison between standard baselines and our intent-augmented reasoning framework.

E.2 Prompts used for IBI Experiments

In this section, we outline the prompt design used in our study of intent-based reasoning for disinformation detection and present templates corresponding to each stage of the IBI experiment. Due to the number and length of the prompts, we do not reproduce them in full here. The complete set of prompts is available in our online repository.

Baselines Figure 4 presents the baseline prompt template used for zero-shot disinformation detection. We focus on three methods introduced by Lucas et al. (2023): *VaN*, *Z-CoT*, and *DeF-SpeC*, which were selected based on their strong performance on human-annotated data. While Lucas et al. (2023) conducted a comprehensive evaluation across both human-annotated and LLM-generated datasets, our study considers only human-annotated examples. Accordingly, we include the top-performing methods in this setting.

Intent Analysis Figure 5 shows the final prompt template used in the first stage of the IBI experiment, which focuses on identifying malicious intent. The prompt integrates the category names and definitions from our intent taxonomy to guide model reasoning. For transparency and reproducibility, we release the full set of final prompts in our public repository.

Disinformation Detection with IBI Figure 6 presents the final prompt template used in the second stage of the IBI experiment, which targets disinformation detection. This prompt builds on the malicious intent analysis produced in the first stage. For each test set, we evaluated three adapted prompt variants based on the *VaN*, *Z-CoT*, and *DeF-SpeC* methods introduced by Lucas et al. (2023). These adaptations align the original methods with our IBI framework. To assess their effectiveness, we compare the adapted methods against their original counterparts as baselines.

F McNemar's Test for Intent-Based Inoculation Experiments

To evaluate the statistical significance of intent-base inoculation, we conducted McNemar's test comparing each prompting method to its IBI-adjusted counterpart across various language models. The results, presented in Table 12, show that IBI improves performance primarily at the significance level of 0.01 across all models and methods

Multilabel Detection

System Prompt:
You are an assistant who detects malicious intention in text. Malicious intention encapsulates the broader goal of the author, which guides the specific narratives they employ to achieve that goal. Your expertise lies in detecting five different malicious intentions:

1. Undermining the credibility of public institutions [UCPI]
2. Changing political views [CPV]
3. Undermining international organizations and alliances [UIOA]
4. Promoting social stereotypes/antagonisms [PSSA]
5. Promoting anti-scientific views [PASV]

User Prompt:
Analyze the text and decide if the text contains any malicious intentions from the following:
Undermining the credibility of public institutions [UCPI],
Changing political views [CPV],
Undermining international organizations and alliances [UIOA],
Promoting social stereotypes/antagonisms [PSSA],
Promoting anti-scientific views [PASV].

You are very conservative in your final decisions and when you are not fully sure you answer No. No.

Give your answer in the form of dictionary:

```
{
   "UCPI": "Your answer if text include Undermining the credibility of public institutions intent. Use only Yes or No",
   "CPV": "Your answer if text include Changing political views intent. Use only Yes or No",
   "UIOA": "Your answer if text include Undermining international organizations and alliances intent. Use only Yes or No",
   "PSSA": "Your answer if text include Promoting social stereotypes/antagonisms intent. Use only Yes or No",
   "PASV": "Your answer if text include Promoting anti-scientific views intent. Use only Yes or No"
}
```

Text: <*Text to analyze*>

Figure 3: Prompt used for multilabel classification of malicious intent with LLMs. The system is instructed to detect five predefined categories of malicious intent within a given text. The model evaluates all categories simultaneously and returns a dictionary of binary Yes/No decisions for each. The prompt emphasizes a conservative decision-making policy: the model is instructed to respond Yes only when confident.

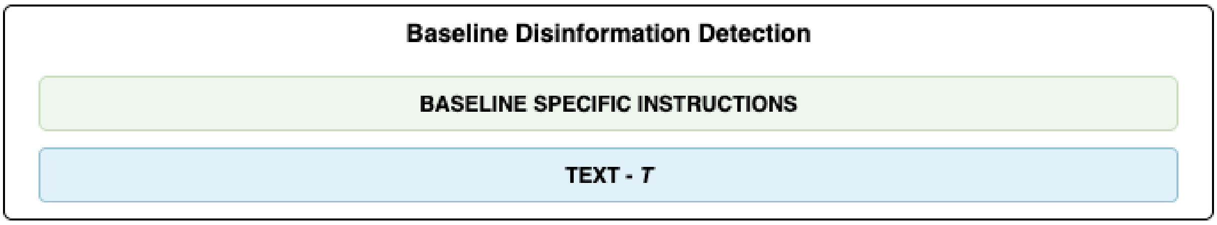
Baseline Disinformation Detection
BASELINE SPECIFIC INSTRUCTIONS
TEXT - T

Figure 4: The prompt template for each baseline method in disinformation detection, namely, *VaN*, *Z-CoT*, and *DeF-SpeC*. Each baseline method differs in the *Baseline Specific Instructions* block. Generally, it provides method-specific guidelines defining the task and requests for structured output. The text T represents the content passed for disinformation evaluation.

in the overall evaluation. However, certain cases, such as experiments on twitter posts exhibit significance for each adjusted prompting method only on GPT-4.1 Mini model. Overall, siginificance on 0.05 level or higher can be observed on about 79% different scenarios.

G Intent-Based Inoculation Experiments Results on Different Datasets

Table 13 reports F_1 scores across four disinformation datasets for various models and prompting strategies, further enhanced with intent-based inoculation (IBI). The table compares the Base setup with IBI, showing that IBI consistently improves performance across nearly all models and datasets. The largest gains are observed on datasets containing longer-form news articles, while improvements on ECTF are more modest, likely due to limited context in shorter social media posts. Overall, these results demonstrate that intent-augmented reasoning effectively enhances disinformation detection, particularly for longer-form texts, across diverse datasets and model architectures.

H Impact of Intent Prediction on Disinformation Detection with IBI

Table 14 shows the average F_1 scores for disinformation detection, split by whether the predicted intent matches the gold labels ("Correct") or not ("Incorrect"). Overall, accurate intent predictions generally improve disinformation detection performance, as evidenced by higher F_1 scores in the

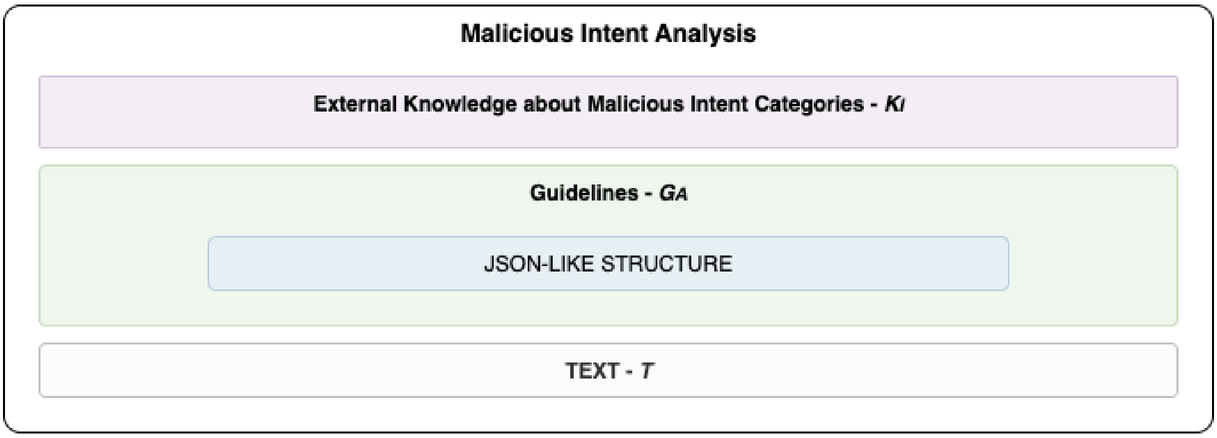
Malicious Intent Analysis
External Knowledge about Malicious Intent Categories - K_I
Guidelines - G_A
JSON-LIKE STRUCTURE
TEXT - T

Figure 5: The prompt template for first stage of IBI experiment, namely for intent analysis. The component K_I encapsulates knowledge about a predefined set of malicious intent categories. Guidelines G_A determine the task and specify the structure of the expected response. Finally, the text T represents the content passed for intent analysis.

"Correct" subsets for most intents and models. Notable exceptions, such as UIOA for some models, indicate that even mispredicted intents can still provide useful signals for detection. While the analysis evaluates each intent independently, the underlying multi-label nature of intent prediction means correct predictions in other intents could offset errors, making the actual impact of intent on F_1 more nuanced. These results highlight that intent-augmented reasoning can enhance detection, but the relationship between intent accuracy and disinformation detection is complex and context-dependent.

Table 14 presents average results across all three competitive baseline methods improved to intent-based resoning. In contrast, Tables 15, 16, and 17 report results for each specific baseline method further enhanced with intent-based reasoning.

I Cross-lingual results with IBI

Table 19 presents F_1 scores computed on approximately 500 test texts per language across six languages in the EUvsDisinfo dataset. Overall, IBI consistently improves performance across all models and languages, demonstrating the effectiveness of intent-augmented reasoning in disinformation detection. These results confirm that incorporating intent information provides robust performance gains even for low-resource languages like Estonian.

Table 18 shows that the IBI approach consistently outperforms the Base model across all languages, achieving higher averaged F_1 scores with generally lower variability. The reported values are averaged over all models and methods and are presented together with their standard deviations, providing a clear indication of performance stability.

J MALINT Comparison with Existing Datasets

Table 20 provides a comparison of MALINT with existing datasets and papers that contain intent analysis for disinformation. The comparison highlights several key aspects:

- **Dataset origin:** While some datasets are enhanced versions of existing collections Zhou et al. (2022), MALINT and Modzelewski et al. (2024) are fully original datasets. Moreover, one study experiments with intent behind misinformation, but does not present any novel dataset Wang et al. (2024a).
- **Intent categorization:** MALINT introduces 6 categories focused on current global disinformation, whereas other datasets either provide binary labels or local/national intent categories.
- **Annotation quality:** Unlike some datasets relying partially on algorithmic or weak annotation, MALINT is fully human-annotated and additionally documents each step of the annotation process, ensuring transparency and reproducibility.
- **Language coverage:** MALINT is in English and covers global disinformation, expanding beyond datasets that are either language-specific or region-specific.

Overall, MALINT stands out by combining original data, comprehensive human annotation, multi-step transparency, and a focus on globally relevant disinformation intents.

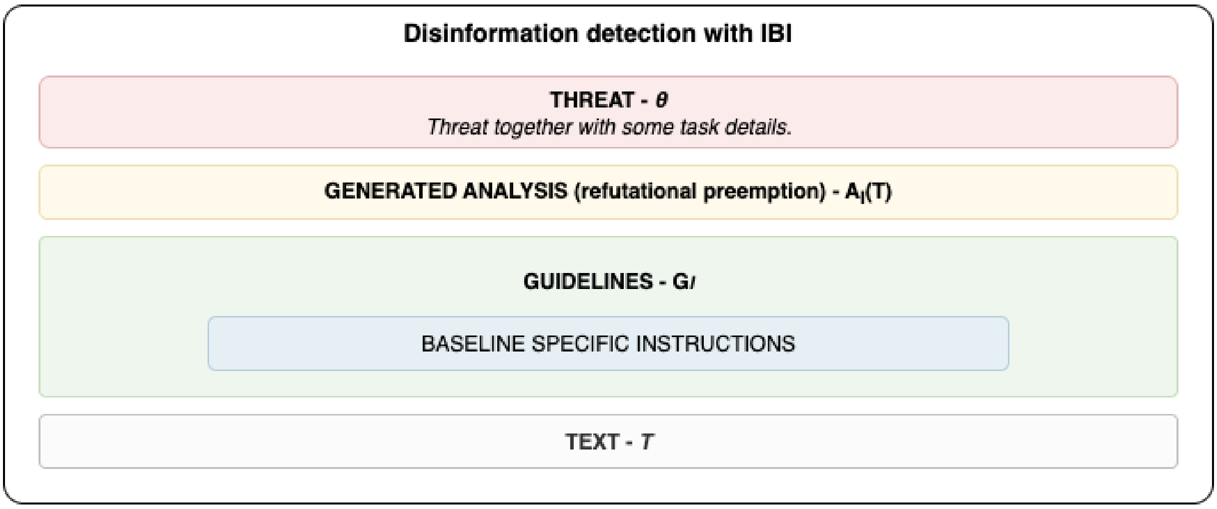
Disinformation detection with IBI
THREAT - θ
Threat together with some task details.
GENERATED ANALYSIS (refutational preemption) - A_I(T)
GUIDELINES - G_I
BASELINE SPECIFIC INSTRUCTIONS
TEXT - T

Figure 6: The prompt template for second final stage of IBI experiments, namely for disinformation detection step. The component θ provides threat against malicious intents and gives some task details. Next component is the generated analysis $A_I(T)$ from the output of first step of IBI experiment and finally, the text T represents the content passed for disinformation evaluation. G_I fully determine the task and specify the structure of the expected response. The *Baseline Specific Instructions* block is a part of guidelines and includes different instructions depending on which baseline method was adapted to IBI experiment, namely it can be instruction from *VaN*, *Z-CoT*, or *DeF-SpeC*

K Generalizability of Intent Categories and IBI Experiments

The generalizability of the intent categories used in this work was empirically validated. This validation demonstrates that the categories are applicable across multiple disinformation datasets.

Experiments with the Intent-Based Inoculation (IBI) approach were conducted on five disinformation datasets, including CoAID, ISOT Fake News, and ECTF. To demonstrate that these datasets encompass global information and context beyond Europe, topic modeling was performed using BERTopic with GPT-4o-mini.

Example topics for three of the datasets are provided below:

ECTF - Global Topics: COVID-19 updates Canada; Sat Bhakti and Health; Israel coronavirus vaccine; Global COVID-19 Solidarity; Bioweapons in Wuhan; Wuhan virus theories; Vaccine funding initiatives; Trump and COVID-19

ISOT Fake News - Global Topics: Rohingya refugee crisis; Venezuela political crisis; Mugabe's Political Crisis; Brazilian Corruption Scandal; North Korea Sanctions; Duterte's War on Drugs; Saudi anti-corruption purge

CoAID - Global Topics: Global COVID-19 Response; Coronaviruses and Animals; Racial health disparities; Economic impact of COVID-19; COVID-19 Vaccine Distribution; Tuberculosis Awareness and Response

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

Table 9: Optimal hyperparameters for binary classification of all malicious intent categories.

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

Table 10: Optimal hyperparameters for each model in multilabel malicious intent classification.

API Model Name	**Knowledge Cutoff Date**	**Access Details**	**License**	**Model Size**
`gpt-4o-mini`	October 2023	OpenAI API 07.2025	Commercial	Not Disclosed
`gemini-2.0-flash`	June 2024	Google API 07.2025	Commercial	Not Disclosed
`gpt-4.1-mini-2025-04-14`	June 2024	OpenAI API 07.2025	Commercial	Not Disclosed
`meta-llama/Llama-3.3-70B-Instruct`	December 2023	DeepInfra API 07.2025	Meta Llama 3 Community	70B
`google/gemma-3-27b-it`	August 2024	DeepInfra API 07.2025	Gemma Terms of Use	27B

Table 11: Large Language Models used in our experiments.

Method	Split	Gemini 2.0 Flash	GPT 4.1 Mini	GPT 4o Mini	Llama3.3 70B	Gemma 3 27B
VaN	Overall	0.010	0.010	0.010	0.050	0.010
VaN	News Article	0.010	0.010	0.010	0.010	0.010
VaN	Twitter Post	N.S.	0.010	0.010	N.S.	N.S.
VaN	Prior Cutoff	0.010	0.010	0.010	N.S.	0.010
VaN	Post Cutoff	0.010	0.010	0.010	0.050	N.S.
Z-CoT	Overall	0.010	0.010	0.010	0.010	0.010
Z-CoT	News Article	0.010	0.010	0.010	0.010	0.010
Z-CoT	Twitter Post	N.S.	0.010	0.010	N.S.	0.010
Z-CoT	Prior Cutoff	0.010	0.010	0.010	0.010	0.010
Z-CoT	Post Cutoff	0.010	0.010	N.S.	0.010	N.S.
DeF_Spec	Overall	0.010	0.010	0.010	0.010	0.010
DeF_Spec	News Article	0.010	0.010	0.050	0.010	0.010
DeF_Spec	Twitter Post	N.S.	0.010	0.010	N.S.	N.S.
DeF_Spec	Prior Cutoff	0.010	0.010	0.010	0.010	0.010
DeF_Spec	Post Cutoff	N.S.	N.S.	N.S.	0.010	N.S.

Table 12: Results of McNemar's test, comparing each prompting method (*VaN*, *Z-CoT*, and *DeF-Spec*) against its IBI-adjusted counterpart across various language models. The values represent significance levels for different evaluation metrics, with *N.S* as *Non-Significant* indicating no statistically significant difference at the 0.05 threshold.

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

Table 13: F_1 scores on four disinformation datasets for competitive prompting methods and their enhancement with Intent-Based Inoculation.

Model	CPV		PSSA		UIOA		PASV		UCPI	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
gpt-4o-mini	0.914	0.787	0.882	0.809	0.860	0.864	0.866	0.850	0.905	0.796
Llama-3.3-70B-Instruct	0.889	0.825	0.876	0.839	0.856	0.899	0.862	0.888	0.836	0.941
gpt-4.1-mini	0.872	0.869	0.886	0.831	0.862	0.908	0.862	0.905	0.876	0.863
gemini-2.0-flash	0.878	0.821	0.855	0.866	0.845	0.902	0.854	0.871	0.865	0.842
google/gemma-3-27b-it	0.853	0.806	0.834	0.799	0.824	0.824	0.815	0.862	0.876	0.722

Table 14: Average F1 scores for disinformation detection across three methods (VaN, Z-CoT, DeF-Spec), split by correct and incorrect intent prediction for each intent.

	CPV		PSSA		UIOA		PASV		UCPI	
Model	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
gpt-4o-mini	0.910	0.783	0.883	0.792	0.852	0.867	0.860	0.848	0.903	0.788
Llama-3.3-70B	0.881	0.829	0.870	0.843	0.855	0.889	0.857	0.889	0.834	0.933
gpt-4.1-mini	0.870	0.879	0.882	0.850	0.864	0.912	0.861	0.921	0.871	0.877
gemini-2.0-flash	0.869	0.830	0.849	0.871	0.843	0.896	0.856	0.853	0.871	0.821
gemma-3-27b-it	0.853	0.800	0.831	0.792	0.828	0.800	0.812	0.857	0.879	0.707

Table 15: F_1 scores with VaN + IBI, split by intent prediction correctness.

	CPV		PSSA		UIOA		PASV		UCPI	
Model	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
gpt-4o-mini	0.911	0.766	0.872	0.796	0.850	0.847	0.857	0.832	0.913	0.759
Llama-3.3-70B	0.891	0.818	0.873	0.843	0.854	0.904	0.860	0.889	0.833	0.945
gpt-4.1-mini	0.862	0.860	0.874	0.825	0.844	0.931	0.854	0.889	0.865	0.855
gemini-2.0-flash	0.847	0.819	0.833	0.847	0.825	0.879	0.828	0.866	0.840	0.830
gemma-3-27b-it	0.837	0.786	0.818	0.774	0.805	0.809	0.794	0.857	0.858	0.707

Table 16: F_1 scores with Z-CoT + IBI, split by intent prediction correctness.

	CPV		PSSA		UIOA		PASV		UCPI	
Model	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
gpt-4o-mini	0.921	0.812	0.892	0.839	0.876	0.878	0.880	0.869	0.899	0.841
Llama-3.3-70B	0.894	0.829	0.886	0.829	0.861	0.904	0.868	0.885	0.840	0.945
gpt-4.1-mini	0.885	0.867	0.901	0.819	0.878	0.881	0.871	0.906	0.892	0.857
gemini-2.0-flash	0.918	0.814	0.882	0.879	0.866	0.932	0.877	0.896	0.885	0.874
gemma-3-27b-it	0.868	0.832	0.852	0.830	0.840	0.864	0.840	0.871	0.892	0.752

Table 17: F_1 scores with DeF-Spec + IBI, split by intent prediction correctness.

Language	**Base**	**IBI**
German	0.794 ± 0.06	0.911 ± 0.02
Spanish	0.683 ± 0.08	0.828 ± 0.04
Estonian	0.716 ± 0.11	0.892 ± 0.04
French	0.611 ± 0.06	0.749 ± 0.05
Polish	0.709 ± 0.1	0.846 ± 0.05
Russian	0.619 ± 0.07	0.735 ± 0.02

Table 18: Comparison of Base vs IBI performance across languages. Results present averaged F_1 scores together with standard deviations over all models and methods.

	Spanish		German		Polish	
	Base	IBI	Base	IBI	Base	IBI
GPT 4o Mini						
VaN	0.715	0.868 ↑21%	0.806	0.925 ↑15%	0.734	0.842 ↑15%
Z-CoT	0.710	0.877 ↑24%	0.813	0.932 ↑15%	0.727	0.840 ↑16%
DeF-Spec	0.742	0.825 ↑11%	0.850	0.913 ↑7%	0.696	0.775 ↑11%
Llama 3.3 70B						
VaN	0.719	0.817 ↑14%	0.813	0.892 ↑10%	0.809	0.847 ↑5%
Z-CoT	0.710	0.820 ↑15%	0.814	0.897 ↑10%	0.733	0.844 ↑15%
DeF-Spec	0.738	0.805 ↑9%	0.833	0.909 ↑9%	0.860	0.855 ↓1%
GPT 4.1 Mini						
VaN	0.561	0.775 ↑38%	0.693	0.887 ↑28%	0.579	0.775 ↑34%
Z-CoT	0.549	0.775 ↑41%	0.671	0.882 ↑31%	0.546	0.775 ↑42%
DeF-Spec	0.717	0.796 ↑11%	0.824	0.895 ↑9%	0.690	0.796 ↑15%
Gemini 2.0 Flash						
VaN	0.560	0.797 ↑42%	0.737	0.885 ↑20%	0.556	0.873 ↑57%
Z-CoT	0.568	0.791 ↑39%	0.756	0.886 ↑17%	0.611	0.851 ↑39%
DeF-Spec	0.744	0.801 ↑8%	0.845	0.926 ↑10%	0.802	0.904 ↑13%
Google Gemma 3-27B						
VaN	0.711	0.871 ↑23%	0.792	0.941 ↑19%	0.809	0.903 ↑12%
Z-CoT	0.722	0.869 ↑20%	0.814	0.942 ↑16%	0.813	0.901 ↑11%
DeF-Spec	0.782	0.880 ↑13%	0.846	0.953 ↑13%	0.860	0.912 ↑6%
	French		**Russian**		**Estonian**	
	Base	IBI	Base	IBI	Base	IBI
GPT 4o Mini						
VaN	0.618	0.757 ↑22%	0.644	0.749 ↑16%	0.732	0.945 ↑29%
Z-CoT	0.632	0.759 ↑20%	0.657	0.750 ↑14%	0.727	0.933 ↑28%
DeF-Spec	0.684	0.739 ↑8%	0.711	0.738 ↑4%	0.705	0.859 ↑22%
Llama 3.3 70B						
VaN	0.629	0.773 ↑23%	0.626	0.745 ↑19%	0.756	0.902 ↑19%
Z-CoT	0.628	0.755 ↑20%	0.643	0.752 ↑17%	0.751	0.907 ↑21%
DeF-Spec	0.661	0.763 ↑15%	0.664	0.764 ↑15%	0.805	0.908 ↑13%
GPT 4.1 Mini						
VaN	0.545	0.676 ↑24%	0.509	0.718 ↑41%	0.546	0.835 ↑53%
Z-CoT	0.533	0.676 ↑27%	0.505	0.730 ↑45%	0.551	0.833 ↑51%
DeF-Spec	0.648	0.715 ↑10%	0.667	0.722 ↑8%	0.637	0.847 ↑33%
Gemini 2.0 Flash						
VaN	0.502	0.729 ↑45%	0.489	0.705 ↑44%	0.579	0.871 ↑50%
Z-CoT	0.514	0.711 ↑38%	0.523	0.696 ↑33%	0.601	0.857 ↑43%
DeF-Spec	0.641	0.736 ↑15%	0.655	0.733 ↑12%	0.767	0.852 ↑11%
Google Gemma 3-27B						
VaN	0.621	0.800 ↑29%	0.644	0.736 ↑14%	0.849	0.939 ↑11%
Z-CoT	0.626	0.807 ↑29%	0.655	0.738 ↑13%	0.857	0.943 ↑10%
DeF-Spec	0.688	0.839 ↑22%	0.699	0.744 ↑6%	0.874	0.952 ↑9%

Table 19: F_1 scores on the EUvsDisinfo dataset separated by six languages computed on about 500 test texts per language. Results report F_1 for competitive prompting methods as the baseline (*Base*) and their enhancement with intent-based reasoning (*IBI*).

Paper	Dataset	Intent Categories	Fully Annotated	Language	Algorithmic Annotation	Annotation from Each Step
Zhou et al. (2022)	Enhanced from Existing	Intentional / Unintentional	No	English	Yes	No
Modzelewski et al. (2024)	Original	9 categories focused on Polish disinformation	Yes	Polish	Fully annotated by humans	No
Wang et al. (2024a)	No dataset presented	Hierarchy of expressed intents of articles (lowest-level labels: Popularize, Clout, Smear, Conflict, Connect)	No	English	Yes	No
Our MALINT	Original	6 categories focused on current global disinformation	Yes	English	Fully annotated by humans	Yes

Table 20: Comparison of datasets related to MALINT, highlighting the novelty of the MALINT dataset