DiNaM: Disinformation Narrative Mining with Large Language Models

Witold Sosnowski[1], Arkadiusz Modzelewski[1,2,3], Kinga Skorupska[1], Adam Wierzbicki[1]

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

Correspondence: witold.sosnowski.pw@gmail.com

Abstract

Disinformation poses a significant threat to democratic societies, public health, and national security. To address this challenge, fact-checking experts analyze and track disinformation narratives. However, the process of manually identifying these narratives is highly time-consuming and resource-intensive. In this article, we introduce DiNaM, the first algorithm and structured framework specifically designed for mining disinformation narratives. DiNaM uses a multi-step approach to uncover disinformation narratives. It first leverages Large Language Models (LLMs) to detect false information, then applies clustering techniques to identify underlying disinformation narratives. We evaluated DiNaM's performance using ground-truth disinformation narratives from the EUDisinfoTest dataset. The evaluation employed the Weighted Chamfer Distance (WCD), which measures the similarity between two sets of embeddings: the ground truth and the predicted disinformation narratives. DiNaM achieved a state-of-the-art WCD score of 0.73, outperforming general-purpose narrative mining methods by a notable margin of 16.4–24.7%. We are releasing DiNaM's codebase and the dataset to the public.

1 Introduction

Disinformation has emerged as a powerful force in digital media, posing serious threats such as physical harm and the erosion of democracy (Dowse and Bachmann, 2022). Malicious actors have leveraged disinformation in campaigns that heavily influenced events such as the COVID-19 pandemic (Agley and Xiao, 2021) and the Russo-Ukrainian war (OECD, 2022). These campaigns use strategic, organized tactics to spread misleading content over time, shaping societal beliefs and attitudes (Suau and Puertas-Graell, 2023). A pointed example of this is the disinformation campaign and subsequent annulment of the first round of 2024 Romanian presidential elections (Erizanu). As Starbird et al. (2019) emphasizes, disinformation goes beyond isolated falsehoods, weaving interconnected pieces of information to serve broader agendas. Disinformation narratives, which unify multiple falsehoods into cohesive and memorable patterns, are a cornerstone of these campaigns (Suau and Puertas-Graell, 2023). Therefore, developing an automated method to uncover disinformation narratives is essential to understand, monitor and counteract their influence.

Recent NLP research has explored narrative classification (Moral et al., 2024) and broader narrative mining (Ash et al., 2021; Anantharama et al., 2022). However, disinformation narrative mining, which could offer a holistic understanding of interconnected falsehoods in coordinated campaigns, has not been explored. To address this gap, we introduce the **DiNaM**: **Di**sinformation **Na**rrative **M**ining algorithm, a novel approach designed to systematically uncover disinformation narratives, advancing the field beyond current methodologies. DiNaM leverages LLMs to identify false information from fact-checking articles and applies embedding-based clustering to group semantically similar information, ultimately deriving disinformation narratives. We rigorously evaluate DiNaM using ground truth disinformation narratives from the EUDisinfoTest dataset (Sosnowski et al., 2024), which contains the most prominent disinformation themes circulating within the EU. The ground-truth narratives were manually created by experts from numerous fact-checking articles, a highly time-consuming endeavor. We aim to expedite this process without a drop in quality. Furthermore, we introduce one of the first NLP applications of Weighted Chamfer Distance (WCD) (Barrow et al., 1977) to measure the similarity between sets of texts, specifically to assess disinformation narratives alignment with ground truth. Moreover, we use WCD to compare DiNaM with general narrative mining methods such as CaNarEx (Anantharama et al., 2022) and Relatio (Ash et al., 2021).

Our main contributions are as follows:

- We introduce the timely problem of disinformation narrative mining and propose a methodology to address it.
- We present DiNaM, a novel algorithm for mining disinformation narratives from a corpus of fact-checking articles.
- We propose a multi-step methodology that leverages LLMs to extract, verify, and refine instances of false information, combined with clustering techniques to uncover disinformation narratives.
- We provide a comprehensive evaluation of the DiNaM algorithm and demonstrate the effectiveness and robustness of our approach compared to general narrative mining algorithms, setting new state-of-the-art standards.

To support reproducibility, we publicly release the DiNaM methodology, dataset, prompts, and codebase [1].

2 Literature Review

Disinformation Narratives in NLP. The study of disinformation narratives in NLP has expanded with LLM advancements. Sosnowski et al. (2024) established a benchmark for LLM narrative classification using EU DisinfoLab reports. Modzelewski et al. (2024) introduced the MIPD dataset, generalizing narratives as intention types. Skumanich and Kim (2024) developed AI tools to monitor political and commercial disinformation, while Santos (2023) applied linguistic and sentiment analysis to counter false narratives. Smith et al. (2021) used topic modeling and narrative networks to trace disinformation on Twitter. The DIPROMATS initiative (Moral et al., 2024) analyzed diplomatic tweets to detect strategic narratives.

Narrative Mining. Although research on disinformation narrative mining is limited, general methods offer valuable insights. Relatio (Ash et al., 2021) uses Semantic Role Labeling (SRL)(Stanovsky et al., 2018) to build multigraphs of narrative structures. CANarEx(Anantharama et al., 2022) enhances this with co-reference resolution, improving entity linking and coherence. Graph-based LLM methods also analyze evolving narratives in economic news (Miori and Petrov, 2024). Recent works on propaganda narrative mining further broaden the landscape. For example, Liu et al. (2024) propose PropaInsight, a framework dissecting propaganda by techniques, appeals, and intent, supported by a novel annotated dataset to improve model performance. Ai et al. (2024) release TweetIntent@Crisis, a dataset capturing competing narratives from Russian and Ukrainian Twitter accounts during the conflict. Unlike these, our work introduces disinformation narrative mining as a distinct task, proposing the DiNaM algorithm.

LLMs for Data Mining. LLMs like GPT-4 are transforming data mining. Zhang et al. (2024) showed GPT-4 automating energy management tasks like load prediction and anomaly detection. Wan et al. (2024) introduced TnT-LLM, leveraging LLMs for large-scale text classification and taxonomy generation. Fink et al. (2023) demonstrated GPT-4's superior accuracy in extracting oncologic data from CT reports. Building on these advances, our work applies LLMs to identify false information and uncover disinformation narratives through a novel, multi-step methodology.

3 Disinformation Narrative Mining Problem

To build a foundation for the disinformation narrative mining algorithm, we define disinformation narratives, formalize the mining problem, and outline the steps to solve it.

Disinformation Narrative Definition

The European Digital Media Observatory [2] defines a disinformation narrative as *a clear message that emerges from a consistent set of contents that can be demonstrated as false using the fact-checking methodology* (EDMO, 2024a). This definition, as utilized by Sosnowski et al. (2024) and aligned with other research (Suau and Puertas-Graell, 2023), serves as the foundation for determining the primary characteristics of a disinformation narrative.

Inaccurate basis: A disinformation narrative emerges from a set of contents that must be demonstrably false.

Repeatability: A disinformation narrative depends on a collection of false contents; it cannot be derived from a single instance.

Clarity: A disinformation narrative forms a comprehensible and unified pattern of information that

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

[2] EDMO funded by the EU and operating under the European University Institute, is one of the most important European institutions dedicated to combating disinformation, coordinating efforts across national hubs, researchers, and policymakers. (Observatory, 2024)

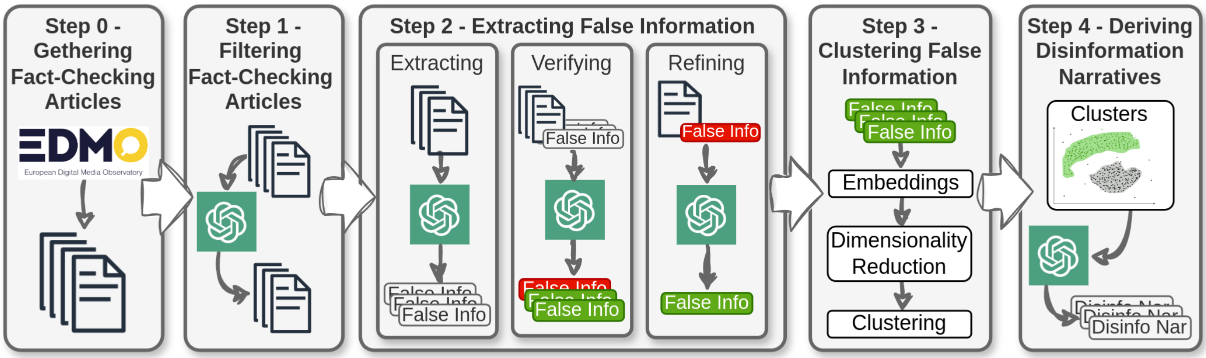

Step 0 - Gethering Fact-Checking Articles
EDMO
European Digital Media Observatory
Step 1 - Filtering Fact-Checking Articles
Step 2 - Extracting False Information
Extracting
Verifying
Refining
False Info
False Info
False Info
False Info
False Info
Step 3 - Clustering False Information
False Info
Embeddings
Dimensionality Reduction
Clustering
Step 4 - Deriving Disinformation Narratives
Clusters
Disinfo Nar

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

emerges from false claims.

Note that disinformation narratives often follow the *minimal model of narrativity* (Piper et al., 2021) or *micro-narratives* (Anantharama et al., 2022), structured as simple Entity-Verb-Entity (EVE) constructs as noted by Sosnowski et al. (2024). This contrasts with the more complex structure of typical narratives such as the *Narrative Policy Framework* (Shanahan et al., 2018).

Building on this, we formally define the *Disinformation Narrative Mining* problem:

Problem Statement

Given a collection of content items $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$, where each c_i represents a content instance (e.g., text, image, or video) and includes instances of demonstrably false information. The task is to retrieve a set of disinformation narratives $\mathcal{N} = \{\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_k\}$ from $\mathcal{C}$. Each narrative $\mathcal{N}_i$ must emerge from a group of contents clustered from $\mathcal{C}$ and conform to the EDMO definition (EDMO, 2024a).

Problem Solving Methodology

We propose the following methodology, consisting of three main stages, to systematically tackle the Disinformation Narrative Mining problem:

1. **Identify false information:** Identify false information from each content item $c \in \mathcal{C}$ using the function $\mathcal{F} : c \to \{S_1, S_2, \ldots, S_m\}$. The set of all identified false information across the entire collection is given by $\mathcal{S} = \bigcup_{c \in \mathcal{C}} \mathcal{F}(c)$
2. **Cluster false information:** Partition the set $\mathcal{S}$ into k clusters $\mathcal{K} = \{\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_k\}$ using a clustering function $\mathcal{P} : \mathcal{S} \to \mathcal{K}^k$, which groups similar pieces of false information together.
3. **Derive narratives:** Derive disinformation narratives using the function $\mathcal{G} : \mathcal{K} \to \{\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_k\}$, which maps each cluster $\mathcal{K}_i \in \mathcal{K}$ to a narrative $\mathcal{N}_i$ analyzing patterns within each cluster.

4 DiNaM Algorithm

We introduce the ***Di****sinformation* ***Na****rrative* ***M****ining with Large Language Models* (DiNaM) algorithm, an implementation of the methodology for solving the Disinformation Narrative Mining problem. DiNaM leverages fact-checking articles as a reliable source of content ($\mathcal{C}$) to identify sets of false information ($\mathcal{S}$), which are analyzed through clustering to derive a set of narratives ($\mathcal{N}$). Fact-checking articles are widely recognized as among the most credible sources for disinformation analysis, as they are authored by domain experts following transparent and standardized procedures (Guo et al., 2022; Jiang et al., 2020; Lee et al., 2023). Their structured format, accessibility, and verified content make them a foundational resource for identifying and categorizing disinformation at scale (Bateman and Jackson, 2024). Consequently, the use of fact-checking articles for disinformation analysis is strongly supported by prior research (Sánchez del Vas and Tuñón Navarro, 2024; Almansa et al., 2022; Guadalupe and Bernaola, 2020; García-Marín, 2020).

DiNaM implements the three key steps outlined in the Disinformation Narrative Mining Problem:

1. **Identify false information**: DiNaM identifies

false information by first filtering fact-checking articles that conclude the claim being reviewed is false (Section 4.1), and then extracting instances of the false information from these articles (Section 4.2).

2. **Cluster false information**: DiNaM groups identified false information into clusters based on semantic similarities (Section 4.3).
3. **Derive narratives**: For each cluster, DiNaM synthesizes patterns of information into disinformation narratives (Section 4.4).

Note: All processing steps in DiNaM were rigorously validated through empirical testing, which guided our design decisions. The Evaluation Section 6 provides full details on the models, prompts, and implementation.

4.1 Filtering False Information

DiNaM begins by filtering a curated dataset of fact-checking articles to select only those that exclusively review false information [3]. Since fact-checking articles can evaluate claims as true, false, or a mix of both, DiNaM simplifies the task by selecting only articles where all evaluated claims are false. This decision is supported by our finding that fewer than 3% of fact-checking articles assess a mix of true and false claims (see Appendix C.1). This filtering step (Step 1 in Figure 1) takes the form of as a binary classification task, as the goal is to determine whether or not a given article meets the predefined filtering criteria.

4.2 Extracting False Information

After filtering articles, DiNaM proceeds to extract the false information addressed within these articles. As a contextual question answering problem, this task is carried out in three structured substeps performed by an LLM (Step 2 of Figure 1).

Extracting involves identifying textual instances of false information from the filtered articles. The LLM is prompted to extract all relevant false claims analyzed in each fact-checking article.

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

Refining revisits articles that failed verification. The LLM is prompted to re-extract false information, guided by feedback on deficiencies in the previous attempt, to ensure more accurate output [4].

This three-step framework is inspired by the Chain-of-Verification (CoVe) prompt engineering method (Dhuliawala et al., 2023), which has demonstrated strong performance in contextual question-answering tasks (Vatsal and Dubey, 2024). Specifically, we condensed CoVe's four-step into three, making the process simpler while still adhering to CoVe's rationale of stepwise verification.

4.3 Clustering False Information

After preparing the corpus of false information, DiNaM semantically clusters these pieces of information (Step 3, Figure 1). The objective is to group similar false claims into clusters from which distinct disinformation narratives can be derived. This process begins by converting the false information into numerical representations (embeddings) that capture their semantic content. Then we apply dimensionality reduction techniques to address the curse of dimensionality (Grootendorst, 2022). Following this reduction, DiNaM employs a clustering algorithm to group semantically similar claims.

4.4 Deriving Disinformation Narratives

In the final step, DiNaM uses an LLM to find patterns of false information in each cluster and generate a matching disinformation narrative. The model follows a specific prompt based on the definition of a disinformation narrative. This prompt tells the model to read a list of false claims and create a short, clear summary that captures the main misleading message.

To ensure alignment with real-world examples, we define output constraints informed by the EUDisinfoTest dataset (Sosnowski et al., 2024). Most narratives in this dataset are concise, typically under 15 words and self-contained. Accordingly, each generated narrative must be: (1) clear, standalone, and descriptive, (2) no longer than 15 words and (3) expressing the false perspective behind the claims.

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

[4] This procedure only needs to be performed once, as we found that only 8% of claims from the initial extraction required refinement. After refining, just 8% of those (or 0.64% of the total dataset) remained incorrect ("hallucinated"). This residual error rate was deemed negligible. To balance precision and efficiency, we halted further refinement cycles, accepting this minimal error margin.

5 Dataset

DiNaM utilizes a dataset of 10,493 fact-checking articles sourced from 14 independent organizations affiliated with the European Digital Media Observatory (EDMO)(EDMO, 2024b). These organizations are based across all 27 EU member states and Norway, contributing fact-checking content from their respective regions. Collectively, they have conducted over 6,800 training sessions and employed 91 verification tools in support of their fact-checking activities(EDMO, 2024c).

To construct the dataset, we retrieved HTML content using Selenium (Selenium, 2024) and extracted clean article text with Trafilatura (Barbaresi, 2021). The resulting articles have an average length of 868.5 words. The most substantial contributions come from the following EDMO hubs: *BENEDMO* (1,895 articles) and *GADMO* (1,687 articles).

Appendix D provides further details, including topic distribution or temporal trends.

6 Evaluation

This section outlines the evaluation methodology that guided the design decisions in DiNaM. To ensure the robustness, we report averages over three runs and use a standardized prompt template for constructing all DiNaM prompts (see Appendix A.1).

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

Appendix E provides additional details, including: (1) a cost analysis, (2) a generalizability study, (3) results from non-LLM baselines and hyperparameter settings, and (4) a step-by-step overview of DiNaM's pipeline, summarizing the best-performing models along with their corresponding input and output quantities.

6.1 Filtering Fact-Checking Articles

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

Testing Dataset. To evaluate this step, we constructed a labeled dataset of fact-checking articles, annotated as either **positive** (all claims are rated false: the article passes the filter) or **negative** (at least one claim is not rated false: the article fails the filter). To create this test set, we randomly selected 135 articles from the EDMO dataset[5]. Each article was annotated by two independent experts. They reached agreement on 124 articles, which comprise our final evaluation set. Detailed on annotation process are provided in Appendix C.2.

Evaluation. We used GPT-4o-mini, LLaMA-3.3-70B, Qwen3-32B, and Gemma-3-27B for evaluation. Each model was tested in a zero-shot setting[6], predicting whether an article was **positive** or **negative**. We evaluated models on three prompt variants (see Appendix A.2). Predictions were compared to ground-truth labels, and performance was measured using the F1-score.

Results. Table 1 summarizes model performance across all prompt variants, with GPT-4o-mini and Qwen3-32B achieving the highest F1-score. GPT-4o-mini and Qwen3-32B achieved the highest F1 score of 0.88, followed by Gemma-3-27B (0.87) and Llama-3.3-70B (0.86). These high scores suggest the task is relatively easy for current LLMs.

Model	Prompt	F1 Score
GPT4o-mini	Figure 15	**0.88**
Qwen3-32B	Figure 15	**0.88**
Gemma-3-27b	Figure 17	0.87
Llama-3.3-70B	Figure 16	0.86

Table 1: Comparison of F1 scores for zero-shot LLM filtering using each model's top-performing prompt.

6.2 Extracting False Information

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

Testing Dataset. To evaluate this step, we constructed a dataset of 115 pairs: each pair consists of a fact-checking article and the specific false information it addresses. These pairs were sourced from reports published by the EDMO.[7] Each brief contains examples of false information along with the corresponding fact-checking articles.

Evaluation. We evaluated four models: GPT-4o-mini, LLaMA-3.3-70B, Qwen3-32B, and Gemma-

[5] Following Alwosheel et al. (2018), we selected 135 examples to ensure at least 50 per class for binary classification and to account for possible exclusions due to annotator disagreement.

[6] Few-shot classification (Brown, 2020) was not used due to the impracticality of including full article examples.

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

3-27B, focusing on their ability to identify the specific false information discussed in each article from our test set. For this, four prompting strategies were used: (i) Our approach (see Section 4.2) (ii) Chain-of-Thought (CoT) (Wei et al., 2022), (iii) Chain-of-Verification (CoVe) (Dhuliawala et al., 2023), and (iv) Base prompt corresponding to the *extraction* step described in Section 4.2. The prompts used for each strategy are provided in Appendix A.3.

Model performance was assessed using the Average Weighted Chamfer Distance (Avg WCD) between the predicted and ground-truth false information embeddings [8]. Specifically, for each test article i, we computed the Weighted Chamfer Distance between the predicted set of embeddings Y_i and the corresponding reference set X_i. The final Avg WCD score is obtained by averaging these scores across all test samples:

$$AvgWCD = \frac{1}{N}\sum_{i=1}^{N} WCD(X_i, Y_i) \qquad (1)$$

where N is the total number of test pairs. A detailed definition and rationale for WCD are provided in Appendix B.1.

Results. Table 2 summarizes the results. Among prompting strategies, our approach outperformed the rest, with GPT-4o-mini achieving the highest Avg WCD score of 0.81 among all models.

A central challenge in extracting false information is that LLMs often retain debunking context, inadvertently turning false claims into true ones. For example, the claim "Images show Pope Francis in a puffy, bright white coat" is sometimes extracted as "AI-generated images show Pope Francis in a puffy, bright white coat." Our approach explicitly identifies and removes such debunking cues, preserving only the false component of the claim.

Among the baselines, only CoVe attempts verification, but it does not filter out debunking material, which limits its effectiveness. CoT and Base, by contrast, do not attempt verification at all, resulting in even weaker performance. A detailed breakdown of models, prompting strategies, and non-LLM baselines is provided in Appendix E.3.

Model	Our	CoVe	CoT	Base
GPT4o-mini	**0.81**	0.79	0.78	0.77
LLama3.3-70B	0.80	0.76	0.78	0.75
Qwen3-32B	0.80	0.77	0.75	0.75
Gemma3-27B	0.79	0.79	0.76	0.73

Table 2: Average WCD scores for false information extraction across models and prompting methods. LLMs are evaluated with four prompting variants.

6.3 Clustering of False Information

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

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

Evaluation. We assessed clustering quality using the Silhouette Score (Rousseeuw, 1987), performing a grid-based evaluation that tested all combinations of embedding models, dimensionality reduction techniques, and clustering algorithms.

We evaluated three embedding models: SFR-Embedding-2 (Rui Meng, 2024), E5-large (Wang et al., 2024), and jina-embeddings-v3 (Sturua et al., 2024), two dimensionality reduction techniques: UMAP (McInnes et al., 2018) and PCA (Abdi and Williams, 2010), and two clustering algorithms: HDBSCAN (McInnes et al., 2017) and K-means (MacQueen, 1967).

Results. Table 3 presents a complete comparison of all tested combinations. The highest Silhouette Score of 0.67 was achieved using SFR-Embedding-2, UMAP, and HDBSCAN, demonstrating superior performance.

The results align with expectations: HDBSCAN outperformed K-means, reflecting the advantages of density-based clustering for textual data (Stewart and Al-Khassaweneh, 2022). Similarly, UMAP outperformed PCA in dimensionality reduction, likely due to UMAP's ability to preserve both local and global structure in high-dimensional, sparse embedding spaces (McInnes et al., 2018). Among embedding models, SFR-Embedding-2 achieved the best performance, consistent with its strong re-

[8] We used SFR-embedding-2 (Rui Meng, 2024) model for embeddings as it ranks highly on the Massive Text Embedding Benchmark (MTEB), a leading benchmark for text embedding quality (Muennighoff et al., 2022).

sults on the MTEB benchmark (Muennighoff et al., 2023).

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

Table 3: Silhouette scores for combinations of embedding models, clustering algorithms, and dimensionality reduction methods.

6.4 Deriving Disinformation Narratives

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

Testing Dataset. To evaluate how accurate the generated narratives are, we compared them to about 200 expert-written narratives from the EUDisinfoTest dataset (Sosnowski et al., 2024).

Evaluation. We evaluated four language models: GPT-4o-mini, LLaMA-3.3-70B, Qwen3-32B, and Gemma-3-27B. For each cluster, we provided the model with the cluster's content and used a standardized prompt (see Appendix A.4) to generate a single disinformation narrative. As a result, each model produced one narrative per cluster. We then compared the generated narratives with reference narratives from a testing dataset. To assess the similarity between the predicted and reference narratives, we used the Weighted Chamfer Distance.

Results. A comparative summary of model performance is shown in Table 4. GPT-4o-mini achieved the highest alignment with expert-written narratives, with a WCD score of 0.73. The full list of disinformation narratives generated by GPT-4o-mini is available in our GitHub repository[9].

Model	WCD
GPT-4o-mini	**0.73**
QWen3-32B	0.72
LLaMA-3.3-70B	0.72
Gemma-3-27B	0.71

Table 4: WCD scores of various models in deriving disinformation narratives.

6.5 Comparison with General Narrative Extraction Methods

To our knowledge, DiNaM is the only method specifically designed for mining disinformation narratives. However, there are two related algorithms for extracting general narratives: CaNarEx (Anantharama et al., 2022) and Relatio (Ash et al., 2021).

Unlike DiNaM, Relatio and CaNarEx are not designed to preprocess fact-checking articles to isolate false information before narratives extraction. As a result, evaluating these methods solely on fact-checking articles would be unfair. Therefore we conducted the evaluation on two complementary datasets: 10,493 fact-checking articles and 12,286 false information instances (see Section 6.3).

Comparison results are presented in Table 5. DiNaM clearly outperforms both Relatio and CaNarEx in mining disinformation narratives from both fact-checking articles and instances of false information. Specifically, DiNaM improves WCD scores by up to 24.7% compared to the best alternative, highlighting its substantial advantage in accurately capturing disinformation narratives.

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

Table 5: WCD scores and relative difference vs. DiNaM. DiNaM significantly outperforms other methods regardless of input format.

7 Results and Discussion

This section presents the disinformation narratives identified by the DiNaM algorithm. We first focus on the key topics underlying these narratives, then provide detailed analyses of those related to COVID-19 and the Ukraine-Russia War. A comprehensive discussion of disinformation narratives on additional topics is available in Appendix F.

[9] https://github.com/wsosnowski/DiNaM/blob/main/disinfonarr2topic.csv

7.1 Disinformation Narrative Topics

We manually categorized the narratives discovered by DiNaM into seven main topics (see more in Appendix C.3). This categorization enabled us to track the evolution of disinformation topics in response to real-world events (Figure 2). In 2020, as the COVID-19 pandemic unfolded, disinformation narratives related to the virus emerged. In 2022, as the war between Ukraine and Russia escalated, disinformation surrounding the conflict intensified. A significant spike in Israeli-Palestinian disinformation occurred in late 2023, coinciding with the onset of the Israel-Palestine war. Notably, COVID-19 narratives briefly resurfaced in early 2022, only to be overshadowed by the surge in Ukraine-Russia disinformation, suggesting the adaptability of disinformation campaigns in response to emerging crises (Surjatmodjo et al., 2024).

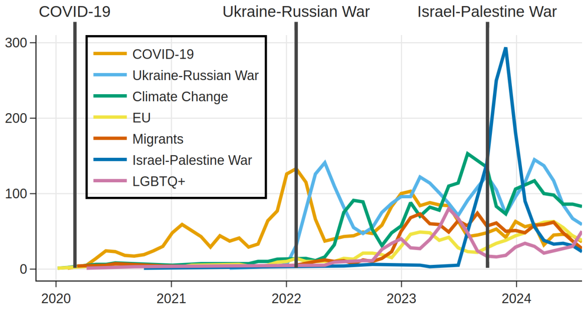


Figure 2: Temporal evolution of disinformation narratives discovered by DiNaM. Spikes in certain topics are related to major real-world events.

7.2 COVID-19 Disinformation Narratives

Figure 3 illustrates the selected disinformation narratives related to COVID-19 identified by DiNaM.A notable trend is observed in the evolution of these narratives over time. Initially, in the aftermath of the COVID-19 outbreak in 2020, disinformation focused on the supposed health risks associated with wearing masks, claiming that they were ineffective in preventing the spread of the virus. This coincided with the WHO's early recommendations on mask usage (World Health Organization, 2020). As the pandemic progressed, the dominant disinformation narrative shifted in 2021 to focus on the alleged dangers of COVID-19 vaccines, which emerged concurrently with the vaccine rollout. However, by mid-2023, the prevalence of these COVID-19 narratives began to decline, likely due to the decreased severity of the pandemic and the WHO's declaration that the international public health emergency had ended (World Health Organization, 2023).

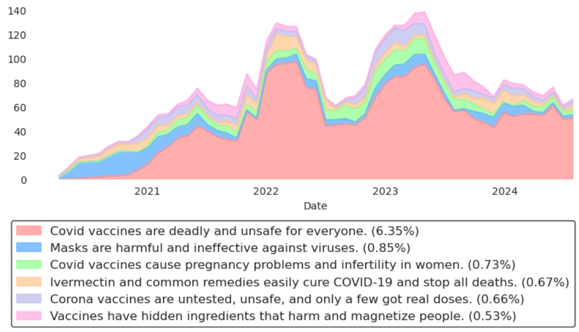


Figure 3: Evolution of Selected Disinformation Narratives Discovered by DiNaM on the COVID-19 topic

7.3 Ukraine-Russian War Disinformation Narratives

Figure 4 presents the selected disinformation narratives from the Ukraine-Russian war identified by DiNaM.Initially, the claim that the war was "fake" gained traction, but it lost momentum likely because media coverage confirmed its reality. Disinformation then shifted to discrediting Ukrainian President Zelensky and portraying Ukrainians negatively. This shift might reflects the adaptability of Russian-aligned disinformation campaigns as noted by Pomerantsev et al. (2023).

Our findings further suggest that Ukraine-Russian War narratives are highly adaptable to real-world events. For instance, claims about the Bucha atrocities being staged emerged in mid-2022, coinciding with reports from Bucha (BBC News, 2022), and the narrative of Leopard tanks being ineffective followed their delivery to Ukraine (for Strategic Studies, 2024). A similar pattern occurred with negative portrayals of Zelensky during his speech at the 78th UN General Assembly (Nations, 2023).

8 Conclusion

In this study, we introduced and formulated the problem of disinformation narrative mining, presenting DiNaM, a novel algorithm designed to mine disinformation narratives from fact-checking articles. Using 10,493 fact-checking entries from the EDMO repository, which includes articles from 14 independent fact-checking organizations across Europe (EDMO, 2024b), DiNaM operates through four steps, each rigorously evaluated with tailored methodologies. The key focus was on the final step: deriving disinformation narratives. To assess

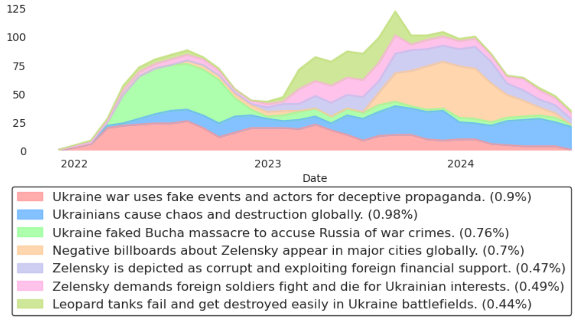


Figure 4: Evolution of Selected Disinformation Narratives Discovered by DiNaM on the Ukraine-Russian War topic.

this, we compared the narratives produced by DiNaM with ground truth narratives from the EUDisinfoTest dataset (Sosnowski et al., 2024), using the Weighted Chamfer Distance metric. DiNaM scored 0.73, surpassing general-purpose methods Relatio and CaNarEx, which scored 0.55–0.61, establishing DiNaM as the state-of-the-art in disinformation narrative mining.

An analysis of the narratives mined by DiNaM revealed strong correlations with real-world events. On one hand, this highlights DiNaM's accuracy and sensitivity to these events. On the other hand, it exposes the nature of disinformation narratives - their adaptability to real-world developments and inherently temporal character.

For future work, we could explore integrating multimodal data to better capture how disinformation spreads across platforms.

9 Acknowledgments

This research was supported by the 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.

10 Limitation

DiNaM leverages LLMs at various stages of processing and inherits their semantic limitations. LLMs can struggle with nuanced contexts and may fail to differentiate between similar but distinct narratives. They may also "hallucinate", which can affect the quality of extracted false information or clustering accuracy. Aware of these challenges, we conducted a thorough evaluation of each stage in DiNaM's pipeline, carefully considering non-LLM alternatives wherever feasible (see Section 6). Furthermore, our algorithm, identifies disinformation narratives using fact-checking articles instead of directly analyzing disinformation content. While this reliance may limit real-time detection, it improves accuracy by leveraging verified falsehoods identified by credible fact-checkers. This approach, rooted in analyzing verification media, is well-supported in the literature (Sánchez del Vas and Tuñón Navarro, 2024; Almansa et al., 2022; Guadalupe and Bernaola, 2020; García-Marín, 2020).

11 Ethical and Broader Impacts

This section outlines the ethical and broader impacts of our research, particularly the use of language models for disinformation narrative mining. While our university's ethical review board deemed the research exempt from further review, we recognize the importance of reflecting on potential impacts, especially regarding the use and reuse of our data and methods.

EDMO dataset of fact-checking articles The EDMO dataset of fact-checking articles is available under a license from EDMO. Downloading these articles for reproducing our research results is permitted by Article 3 and 4 of the Directive 2019/790 on copyright and related rights in the Digital Single Market (DSM Directive).

We assumed that the EDMO dataset complies with relevant legal and ethical standards regarding data protection and content moderation, and therefore did not conduct additional checks for personally identifying information or offensive content.

Intended Use of Our Research Results. Our research, including the DiNaM algorithm, aims to support institutions combating disinformation, such as organizations adhering to the International Fact-Checking Organization's code of conduct. The algorithm is licensed for non-commercial use (CC BY-NC 4.0), explicitly excluding commercial applications.

There is a potential concern that our research could be misused, particularly if the narratives are repurposed to spread disinformation. However, this risk is minimal since all the narratives are rooted in verified fact-checking articles that have already assessed and identified the associated claims as false.

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

Overview of Computational Resources and Costs in Our Research. The DiNaM experiments used OpenAI's GPT-4o and GPT-4o-mini via API, supported by a server with four NVIDIA L40 GPUs for tasks like embedding generation and clustering, with total costs kept under $100.

Expert Involvement. Human annotations validating DiNaM's correctness were conducted by experts with at least three years of experience in fact-checking or verifying information, affiliated with organizations adhering to the International Fact-Checking Network's code of conduct. These experts were university-employed and fairly compensated.

References

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

Jon Agley and Yunyu Xiao. 2021. Misinformation about covid-19: evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health*, 21:1–12.

Lin Ai, Sameer Gupta, Shreya Oak, Zheng Hui, Zizhou Liu, and Julia Hirschberg. 2024. Tweetintent@ crisis: A dataset revealing narratives of both sides in the russia-ukraine crisis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1872–1887.

Ana Almansa, María Jesús Fernández-Torres, and Leticia Rodríguez-Fernández. 2022. Desinformación en españa un año después de la covid-19. análisis de las verificaciones de newtral y maldita. *Revista latina de comunicación social*, (80):183–200.

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

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

Elliott Ash, Germain Gauthier, and Philine Widmer. 2021. Relatio: Text semantics capture political and economic narratives. *arXiv preprint arXiv:2108.01720*.

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

Ainesh Bakshi, Piotr Indyk, Rajesh Jayaram, Sandeep Silwal, and Erik Waingarten. 2024. Near-linear time algorithm for the chamfer distance. *Advances in Neural Information Processing Systems*, 36.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131.

Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. 1977. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proceedings: Image Understanding Workshop*, pages 21–27. Science Applications, Inc.

Jon Bateman and Dean Jackson. 2024. Countering disinformation effectively: An evidence-based policy guide. Technical report, Carnegie Endowment for International Peace.

BBC News. 2022. War in ukraine: Street in bucha found strewn with dead bodies. Retrieved 2 April 2022.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Copernicus Climate Change Service (C3S) and World Meteorological Organization (WMO). 2023. European state of the climate 2023.

European Commission. 2024a. Pact on migration and asylum.

European Commission. 2024b. Report from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions.

Council of the European Union. 2023. Council confirms 6 to 9 june 2024 as dates for next european parliament elections. Archived from the original on 24 July 2023, Retrieved 24 May 2023.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Andrew Dowse and Sascha Dov Bachmann. 2022. Information warfare: methods to counter disinformation. *Defense & Security Analysis*, 38(4):453–469.

EDMO. 2024a. Edmo fact-checking network statement about methodology. Accessed: October 21, 2024.

EDMO. 2024b. Repository of fact-checking articles. Accessed: September 11, 2024.

EDMO. 2024c. United against disinformation: our work in numbers. Accessed: 2024-10-11.

Paula Erizanu. Romania's 'rigged' election shows europe the dangers of russian disinformation.

European Commission. 2023. Approval of fourth insect as a novel food. Retrieved 2 December 2024.

Fernando GUTIERREZ. 2022. Hared migration challenges: The transatlantic community and the mena region.

Matthias A Fink, Arved Bischoff, Christoph A Fink, Martin Moll, Jonas Kroschke, Luca Dulz, Claus Peter Heußel, Hans-Ulrich Kauczor, and Tim F Weber. 2023. Potential of chatgpt and gpt-4 for data mining of free-text ct reports on lung cancer. *Radiology*, 308(3):e231362.

International Institute for Strategic Studies. 2024. Russia and eurasia. *The Military Balance 2024*, 124:210–215.

David García-Marín. 2020. Infodemia global. desórdenes informativos, narrativas fake y fact-checking en la crisis de la covid-19. *Profesional de la información*, 29(4).

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Guadalupe Aguado Guadalupe and Itziar Bernaola. 2020. Verificación en la infodemia de la covid-19. el caso newtral. *Revista latina de comunicación social*, (78):289–308.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Shan Jiang, Simon Baumgartner, Abe Ittycheriah, and Cong Yu. 2020. Factoring fact-checks: Structured information extraction from fact-checking articles. In *Proceedings of The Web Conference 2020 (WWW '20)*.

Hayato Kobayashi, Masaki Noguchi, and Taichi Yatsuka. 2015. Summarization based on embedding distributions. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1984–1989.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Sian Lee, Aiping Xiong, Haeseung Seo, and Dongwon Lee. 2023. "Fact-checking" fact checkers: A data-driven approach. *Harvard Kennedy School Misinformation Review*, 4(5).

Chun-Liang Li, Tomas Simon, Jason Saragih, Barnabás Póczos, and Yaser Sheikh. 2019. Lbs autoencoder: Self-supervised fitting of articulated meshes to point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11967–11976.

Jiateng Liu, Lin Ai, Zizhou Liu, Payam Karisani, Zheng Hui, May Fung, Preslav Nakov, Julia Hirschberg, and Heng Ji. 2024. Propainsight: Toward deeper understanding of propaganda in terms of techniques, appeals, and intent. *arXiv preprint arXiv:2409.18997*.

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

J MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of open source software*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Deborah Miori and Constantin Petrov. 2024. Narratives from gpt-derived networks of news and a link to financial markets dislocations. *International Journal of Data Science and Analytics*, pages 1–25.

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.

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

Harry Moroz, Maheshwor Shrestha, and Mauro Testaverde. 2020. Potential responses to the covid-19 outbreak in support of migrant workers.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

United Nations. 2023. Ukraine | general debate.

European Digital Media Observatory. 2024. European digital media observatory.

OECD. 2022. Disinformation and russia's war of aggression against ukraine.

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

Peter Pomerantsev, Nataliya Gumenyuk, Angelina Kariakina, Inna Borzylo, Tetiana Peklun, Volodymyr Yermolenko, Vitalii Rybak, Denys Kobzin, Maria Montague, Jaroslava Barbieri, Martin Innes, Viorica Budu, and Andrew Dawson. 2023. Why conspiratorial propaganda works and what we can do about it.

ReliefWeb and BBC News. 2023. Israel-palestine conflict: Timeline and escalation in 2023. Accessed: 2024-12-02.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Shafiq Rayhan Joty Caiming Xiong Yingbo Zhou Semih Yavuz Rui Meng, Ye Liu. 2024. Sfr-embedding-2: Advanced text embedding with multistage training.

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

Ryoma Sato, Makoto Yamada, and Hisashi Kashima. 2022. Re-evaluating word mover's distance. In *International Conference on Machine Learning*, pages 19231–19249. PMLR.

Selenium. 2024. *Selenium Browser Automation*. Selenium. Accessed: 2024-09-11.

Elizabeth A Shanahan, Michael D Jones, and Mark K McBeth. 2018. How to conduct a narrative policy framework study. *The Social Science Journal*, 55(3):332–345.

Andy Skumanich and Han Kyul Kim. 2024. Modes of analyzing disinformation narratives with ai/ml/text mining to assist in mitigating the weaponization of social media. *arXiv preprint arXiv:2405.15987*.

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

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.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895.

Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26.

Geoffrey Stewart and Mahmood Al-Khassaweneh. 2022. An implementation of the hdbscan* clustering algorithm. *Applied Sciences*, 12(5):2405.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora.

Jaume Suau and David Puertas-Graell. 2023. Disinformation narratives in spain: reach, impact and spreading patterns. *Profesional de la información*, 32(5).

Dwi Surjatmodjo, Andi Alimuddin Unde, Hafied Cangara, and Alem Febri Sonni. 2024. Information pandemic: A critical review of disinformation spread on social media and its implications for state resilience. *Social Sciences*, 13(8):418.

Rocío Sánchez del Vas and Jorge Tuñón Navarro. 2024. Disinformation on the covid-19 pandemic and the russia-ukraine war: Two sides of the same coin? *Humanities and Social Sciences Communications*, 11(1).

Sénat. 2023. Émeutes de juin 2023 : comprendre, évaluer, réagir.

Shubham Vatsal and Harsh Dubey. 2024. A survey of prompt engineering methods in large language models for different nlp tasks. *arXiv preprint arXiv:2407.12994*.

Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W White, Longqi Yang, et al. 2024. Tnt-llm: Text mining at scale with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5836–5847.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

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

World Health Organization. 2020. Advice on the use of masks in the context of covid-19: interim guidance, 6 april 2020. Technical report, World Health Organization. Hdl:10665/331693.

World Health Organization. 2023. With the international public health emergency ending, who europe launches its transition plan for covid-19.

Chaobo Zhang, Jie Lu, and Yang Zhao. 2024. Generative pre-trained transformers (gpt)-based automated data mining for building energy management: Advantages, limitations and the future. *Energy and Built Environment*, 5(1):143–169.

A Prompts

A.1 Prompt Template

The integration of LLMs into DiNaM workflows required the creation of a specialized prompt template, illustrated in Figure 5. This template's design is inspired by the work of (Lucas et al., 2023), which proposed SOTA prompts for zero-shot disinformation detection and generation. Although our tasks differ, their template closely aligns with our requirements, making it a valuable reference.

Our prompt template is structured around five key components:

1. **Impersonation**: Establishes a contextual role for the LLM, overriding alignment-tuning to ensure task-specific behavior.
2. **Guidelines**: Provides detailed instructions to steer the LLM's responses effectively.
3. **Context**: Embeds relevant data or information to assist the model in completing the task.
4. **Output**: Specifies the desired format of the generated content.
5. **Follow Steps** (used only in Chain-of-Thought prompts): Outlines the step-by-step procedure to guide the reasoning process.

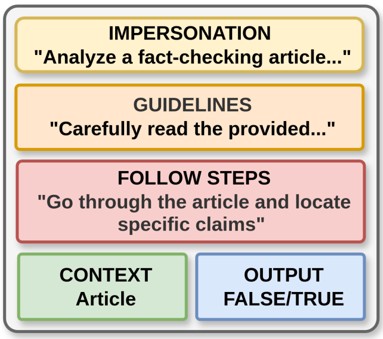

IMPERSONATION
"Analyze a fact-checking article..."
GUIDELINES
"Carefully read the provided..."
FOLLOW STEPS
"Go through the article and locate specific claims"
CONTEXT
Article
OUTPUT
FALSE/TRUE

Figure 5: Prompt template comprising five components: (1) Impersonation, which establishes context and overrides alignment-tuning; (2) Guidelines, which direct the actions of LLMs; (3) Context, which embeds data; (4) Output, which specifies the output format; and (5) Follow Steps, which outlines the Chain-of-Thought procedure steps (if applicable).

A.2 Filtering Fact-Checking Articles

We evaluated three prompt variants for classifying articles as **positive** (indicating all claims are false) or **negative**. While all variants share a common template, they differ in wording and complexity (see Figures 15, 16, and 17).

A.3 Extracting False Information

We tested four prompting strategies designed to extract false information from fact-checking articles: Our Approach (Figure 18), Chain-of-Verification (CoVe) (Figure 20), Chain-of-Thought (CoT) (Figure 19), and a Base prompt (Figure 21).

A.4 Deriving Disinformation Narratives

We used a focused directive prompt (Figure 22) to generate a single, coherent narrative from a cluster of false information.

B Metrics

B.1 Weighted Chamfer Distance

The Weighted Chamfer Distance is a metric designed to evaluate the similarity between two sets of embeddings, addressing cases where the sets differ in size.

B.1.1 Definition and Formula

The Weighted Chamfer Distance quantifies similarity by identifying the closest point in one set for each point in the other set. It is particularly suited for comparing sets of embeddings where no one-to-one correspondence exists between the points.

Formally, given two sets of points, X and Y, the WCD is computed as:

$$\begin{aligned}\mathrm{WCD}(X,Y) = w_1 \cdot \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y} d(x,y) \\ +w_2 \cdot \frac{1}{|Y|}\sum_{y\in Y}\min_{x\in X} d(y,x)\end{aligned} \tag{2}$$

Here, $d(x, y)$ represents the distance between points x and y. The weights w_1 and w_2 we defined as:

$$w_1 = \frac{|X|}{|X| + |Y|}, \quad w_2 = \frac{|Y|}{|X| + |Y|} \tag{3}$$

These weights ensure that the contribution of each set to the overall distance is proportional to its size, preventing smaller sets from disproportionately influencing the result.

The inclusion of weights is a critical aspect of WCD. In scenarios where the sets being compared differ in size, for instance, when comparing the embeddings of predicted narratives to ground-truth narratives, unweighted metrics tend to overemphasize the smaller set, leading to skewed evaluations. By incorporating size-based weights, WCD provides a balanced assessment of similarity, regardless of the size discrepancy between the sets.

B.1.2 Comparison to Related Metrics

WCD can be viewed as a computationally efficient alternative to the Earth-Mover Distance (EMD), often referred to as Relaxed EMD (Bakshi et al., 2024). Unlike EMD, which requires solving linear programming problems, WCD relies solely on point-wise distances, making it significantly faster to compute.

A related metric, Word Mover's Distance (WMD), specializes in comparing documents by treating words as points in an embedding space (Kusner et al., 2015). Both EMD and WMD originate from the field of Optimal Transportation (OT)(Sato et al., 2022), which has applications in domains such as computer graphics(Li et al., 2019), computer vision (Athitsos and Sclaroff, 2003), and natural language processing (NLP) (Kobayashi et al., 2015). While EMD and WMD focus on the distributional similarity of sets, WCD prioritizes point-wise proximity, making it more suitable for tasks where the size of the sets and individual points matter.

B.1.3 Rationale for Choosing WCD

We selected WCD over EMD and WMD for the following reasons:

- **Efficiency:** WCD is computationally less expensive, as it avoids the need for solving optimization problems.
- **Size Sensitivity:** WCD focuses on the nearest-neighbor distances between points, making it ideal for evaluating tasks where the number of narratives or claims varies.

B.1.4 Interpretability and Insights

One of the strengths of WCD is its interpretability. By examining the closest matches between predicted and ground-truth embeddings, WCD helps identify significant discrepancies, highlighting specific cases, such as narratives or false information, where narrative generation or false information extraction may have failed.

B.1.5 Applications in DiNaM

In this work, WCD is employed to evaluate model performance in two core tasks:

1. Extracting False Information For this task (see Section 6.2), X represents the set of embeddings for extracted instances of false information from a given article, while Y corresponds to the ground-truth instances. By using WCD, we assess how closely the extracted instances align with the ground truth.

2. Deriving Disinformation Narratives In this case (see Section 6.4), X denotes the embeddings of narratives generated by the model, and Y represents the embeddings of ground-truth narratives. WCD provides a robust measure of the model's ability to generate narratives that align with the expected ground truth, even when the sets differ in size.

C Annotation Guidelines

C.1 Filtering Fact-Checking Articles - Mixed-Claim Analysis

To assess how frequently fact-checking articles contain a mixture of true and false claims, we conducted an independent analysis of a subset of 135 fact-checking articles from EDMO dataset (see Section 5. This analysis aimed to determine how often fact-checking articles evaluate a *mix* of both true and false claims, rather than claims that are entirely true or entirely false.

Methodology Each of the 135 articles was reviewed by two annotators, who were instructed to identify whether the article included multiple claims with differing veracity outcomes, specifically at least one claim rated as true and at least one rated as false. An article was labeled as "mixed" only if both annotators agreed on the presence of mixed claims. In the few cases where the annotators initially disagreed, they discussed the article to reach a consensus. We did not adopt an exclusion-based strategy (that is, discarding all articles with disagreement) as was done in the dataset filtering task, since the overall number of mixed-claim articles was found to be very low. Consensus-based resolution ensured a more efficient and practical annotation process for this specific analysis.

Findings Out of the 135 articles in the dataset, 4 were found to contain a mix of true and false claims. This corresponds to approximately 3% of the total.

Conclusion The small proportion of mixed-claim articles supports our decision to exclude such cases from the DiNaM dataset. By focusing solely on articles where all evaluated claims are false, we maintain a clearer and more consistent foundation for studying disinformation narratives.

C.2 Filtering Fact-Checking Articles - Dataset Construction

Purpose The primary aim of this annotation process is to construct a dataset of fact-checking articles in which **all the claims examined are false**. As established in Appendix C.1, only a small minority of fact-checking articles (around 3%) contain mixed claims. This further justifies our decision to exclude those cases and focus on articles with fully false content.

Dataset We used a dataset of 135 fact-checking articles sourced from the EDMO repository (see Section 5). These articles were also selected at random for the mixed-claim analysis in Appendix C.1. The same set was repurposed for the current annotation task. Each article was independently reviewed and labeled by two domain experts.

Annotation Task Annotators must carefully review each fact-checking article to determine whether **all claims verified in the article are confirmed as false**. The specific annotation question is:

- Have all the fact-checked claims in the article been confirmed as false?

Annotations should be based on the article's conclusion, avoiding speculation or external context and biases. If the annotators assign differing labels to an article, it is excluded from the final dataset.

Characteristics of the Articles The fact-checking articles in the dataset exhibit a variety of structures and claim types:

1. The article contains a single claim that is evaluated as false.
2. The article contains a single claim that is evaluated and verified as true.
3. The article exclusively features claims that are entirely verified as false.
4. The article exclusively features claims that are entirely verified as true.
5. The article includes multiple claims, with a mix of outcomes where some claims are verified as true and others as false.

Clarification for Annotators For the annotation task, the answer to the question should be marked as **true only if the article meets conditions 1 or 3**. If the article falls under any other condition (2, 4, or 5), the answer should be marked as **false**.

C.3 Categorizing Narratives into Seven Topics

Purpose The primary aim of this annotation process is to classify narratives discovered by DiNaM into one of the seven main disinformation topics as outlined in EDMO's fact-checking briefs [10], or into an **"other"** category if they do not match any of the predefined topics.

Dataset The dataset consists of disinformation narratives discovered by DiNaM as outlined in Section 7. Each narrative was independently reviewed and annotated by two domain experts.

Annotation Task Annotators must carefully review each narrative and categorize it into one of the following eight topics, based on its primary focus:

- *The Ukraine-Russia War*
- *COVID-19*
- *Climate Change*
- *Migration*
- *The Israel-Palestine Conflict*
- *The European Union*
- *LGBTQ+*

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

- **Other:** Narratives that do not align with any of the above topics.

Annotations should be made based solely on the content of the narrative, without inferring additional context or applying personal biases.

If the two annotators independently assign differing categories to a narrative, it is excluded from the main topic-specific categories and labeled as **"other"** in the final dataset.

Characteristics of the Narratives The narratives in the dataset are single sentences that vary in topics and focus. Annotators should note the following:

1. Narratives may explicitly reference one of the seven topics, making the categorization straightforward.
2. Some narratives may reference multiple topics. Annotators should categorize the narrative based on its **primary focus**.
3. Ambiguous or unrelated narratives that do not clearly align with any predefined topic should be categorized as **other**.

Clarification for Annotators For the annotation task, the following rules should be followed:

- Assign a narrative to one of the seven main topics only if its primary focus unambiguously aligns with that topic.
- Assign a narrative to **other** if it does not clearly align with any of the seven topics or if it is assigned different categories by the two annotators.
- Avoid using external sources or personal interpretations to determine the topic; rely strictly on the content provided in the narrative.

Inter-Annotator Agreement To ensure the reliability of the annotation process, inter-annotator agreement was calculated. Annotators did not agree in 11 cases out of a total of 122 narratives. This disagreement rate corresponds to a Cohen's Kappa of 0.895, indicating a very high level of agreement.

Analysis and Observations The results of the annotation process provide insight into the thematic distribution of disinformation narratives identified by DiNaM. As shown in Figure 6, a significant majority (over 60%) of the narratives fall clearly within one of the seven predefined disinformation topics. This suggests that most misinformation narratives encountered in the dataset can be meaningfully categorized using EDMO's framework, underscoring its applicability to real-world data.

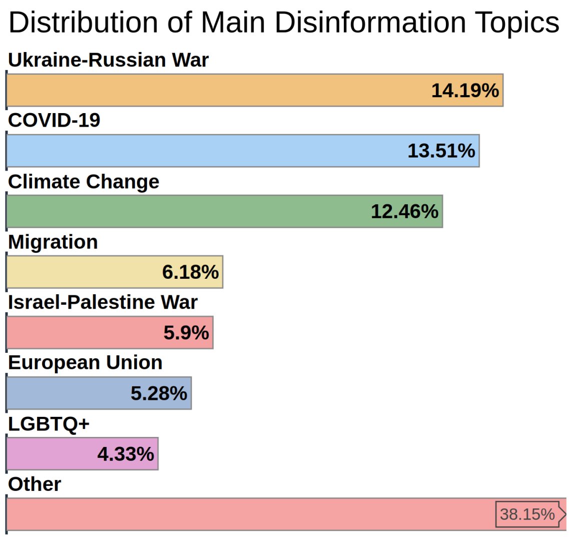


Figure 6: This chart illustrates the proportion of mined disinformation narratives within the seven main topics. The largest segments correspond to narratives related to the Ukraine-Russia War, COVID-19, and Climate Change.

D Dataset

This study leverages a dataset obtained from the EDMO repository of fact-checking articles (EDMO, 2024b). We analyze key characteristics of this dataset, including its temporal distribution, word count distribution, and contributions across different domains. The statistics provided in this section are based on the translated content.

D.1 Temporal Distribution

The distribution of fact-checking articles over time is depicted in Figure 7. A notable increase in articles is observed starting from 2020.

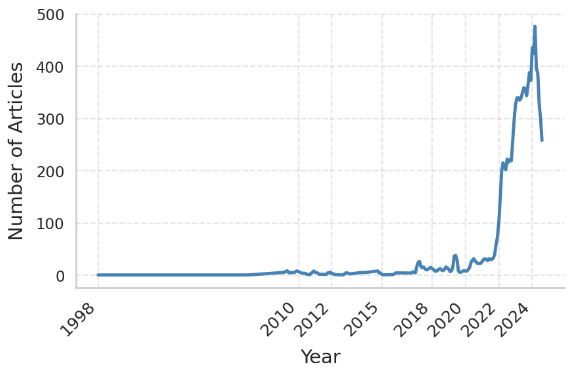


Figure 7: Distribution of fact-checking articles over time.

D.2 Word Count Distribution

The word count distribution of fact-checking articles is shown in Figure 8. The majority of articles have a word count between 500 and 2,000 words, with a long tail representing exceptionally long articles as shown in Figure 8.

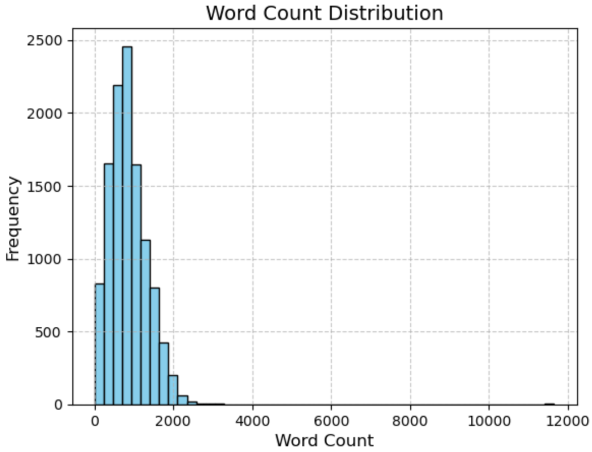


Figure 8: Word count distribution of fact-checking articles.

D.3 Word Count Statistics

Key statistics for word count across the dataset are summarized in Table 6. These statistics provide insights into the variability and typical ranges of article lengths.

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

Table 6: Word count statistics for the dataset.

D.4 Domain Contributions

The dataset contains contributions from various domains. Figure 9 depicts the top 10 most frequent domains, along with their respective contribution percentages to the dataset. These domains represent the main sources of fact-checking articles and highlight the concentration of fact-checking efforts. To maintain consistency, all articles originally written in languages other than English were translated using Google Translate[11].

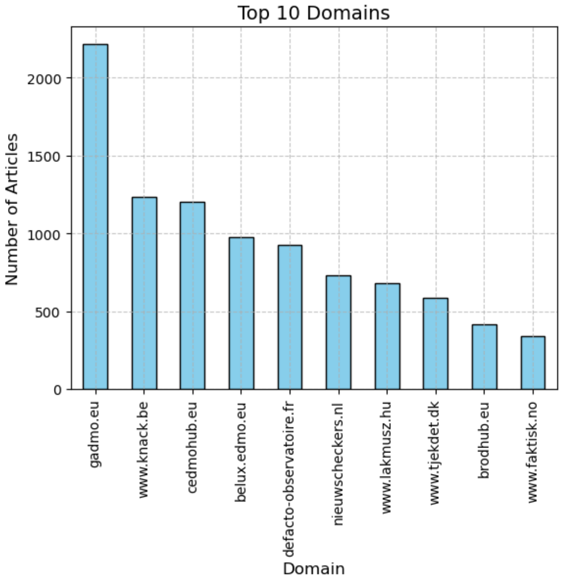


Figure 9: Top 10 most frequent domains in the dataset, along with their percentage contributions.

D.5 Affiliation Contributions

The dataset contains contributions from various affiliations. Figure 10 illustrates the distribution of fact-checking articles across these organizations, showcasing the significant roles played by research hubs and initiatives in combating disinformation.

Among the affiliations, **BENEDMO** leads with the highest number of contributions, totaling 1,895 articles. **GADMO** follows closely with 1,687 articles, reflecting its significant engagement. **EDMOeu** (1,302 articles) and **CEDMO** (1,018 articles) also play key roles in the dataset, further highlighting their dedication to addressing disinformation at a European level.

Other affiliations such as **NORDIS** (895 articles), **BELUX** (830 articles), and **DEFACTO** (824 articles) provide substantial regional perspectives, enriching the dataset with diverse fact-checking narratives. Contributions from **HDMO** (795 articles), **BECID** (760 articles), and **BROD** (574 articles) demonstrate the broader collaborative efforts among different organizations.

Smaller but still important contributions come from **MedDMO** (524 articles), **ADMO** (304 articles), and **IRELAND HUB** (238 articles), underscoring their roles in fact-checking specific domains or regions.

E Evaluation

This appendix provides additional evaluations and implementation details of the DiNaM pipeline. We

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

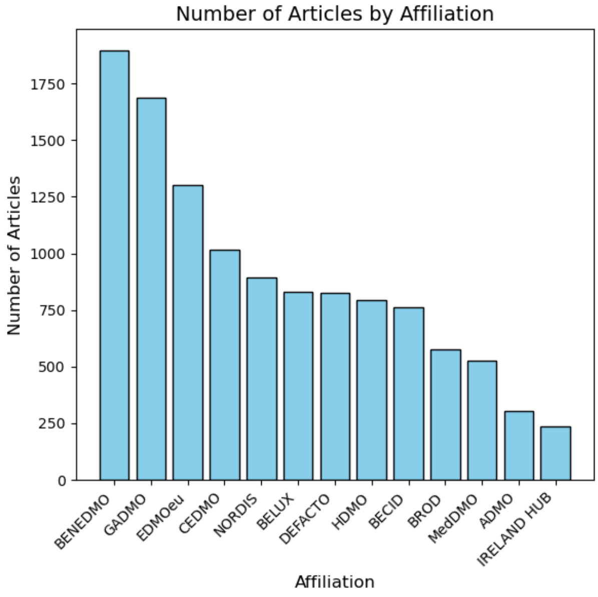


Figure 10: Distribution of fact-checking articles by affiliation, illustrating the contributions of various organizations.

begin with an overview of the full pipeline, including the best-performing models and input/output quantities at each stage (Section E.1).

We then present non-LLM evaluations for three core steps: filtering fact-checking articles (Section E.2), extracting false information (Section E.3), and deriving disinformation narratives (Section E.5).

Section E.4 details hyperparameter optimization for Step 3 (clustering false information), covering the configurations explored and selected for UMAP, HDBSCAN, and K-means.

We also include comparisons with general-purpose narrative extraction methods (Section E.6) and assess the generalizability of DiNaM across sources and topical categories (Section E.7).

Finally, Section E.8 presents a cost analysis across several LLMs, emphasizing DiNaM's efficiency and scalability.

E.1 DiNaM Pipeline Overview

Table 13 summarizes the DiNaM pipeline, outlining the best-performing models and input/output quantities at each step.

E.2 Filtering Fact-Checking Articles

To establish a performance baseline for the filtering task, we fine-tuned two encoder-based transformer models: RoBERTa-large and DeBERTa-large. Both models were trained on the labeled dataset described in Section 6.1, in which each fact-checking article was annotated as either **positive** (all claims rated false) or **negative** (at least one claim not rated false).

Training Details Fine-tuning was conducted using the HuggingFace Transformers library. The input to the models consisted of the full article text. We used binary cross-entropy as the loss function and optimized using AdamW with a batch size of 8, a learning rate of 2e-5, and trained for 5 epochs. Evaluation was performed using 5-fold cross-validation over the 124 annotated articles, ensuring class balance within each fold.

Results Table 11 summarizes the F1 scores achieved by both fine-tuned baselines and LLMs across three prompt variants. The fine-tuned DeBERTa-large model achieved an F1 score of 0.60, while RoBERTa-large reached 0.58. In contrast, zero-shot LLMs consistently outperformed these baselines, with F1 scores ranging from 0.85 to 0.88 across models and prompts.

Discussion Despite being fine-tuned on a task-specific labeled dataset, both DeBERTa-large and RoBERTa-large underperformed relative to the zero-shot results of large language models. This performance gap underscores the capability of LLMs to carry out nuanced content-based filtering without the need for supervised task-specific training, particularly in low-resource or small-data scenarios.

E.3 Extracting False Information

Baselines: Non-LLM Models To complement our LLM evaluations, we benchmarked two non-LLM encoder-based models to assess their ability to extract false information from fact-checking articles.

We evaluated two QA-style systems based on pretrained transformer encoders to establish non-LLM baselines. The first model, DistilBERT-SQD, is a lightweight DistilBERT architecture fine-tuned on the SQuAD v1.1 dataset. The second, RoBERTa-SQD2, builds on RoBERTa-base and is fine-tuned on the SQuAD v2.0 dataset. Both models were tested using a standardized prompt format: *"What false claims are debunked in this article? Context: article"*, where the full article text served as the context input. The outputs were evaluated using the same semantic similarity metric applied to

LLMs, Average Weighted Chamfer Distance (Avg WCD), to ensure consistency across comparisons.

RoBERTa-SQD2 achieved an Avg WCD score of 0.58, while DistilBERT-SQD scored 0.54. These results are substantially lower than those of LLM-based methods (see Table 2), reflecting the limitations of encoder-only models in tasks requiring paraphrase recognition and deeper semantic understanding.

Motivation for a Structured LLM Approach Our decision to adopt a structured, multi-phase approach to false information extraction was driven by a key observation: LLMs frequently generate outputs that mix misleading claims with corrective content, even when explicitly instructed to extract only the false claims. This issue persisted across a range of prompt engineering strategies, as detailed in Section 6.2 and illustrated in Table 16.

To systematically address this problem, we developed a three-step process consisting of Extraction, Verification, and Refinement. This pipeline enables us to isolate and cleanly extract misleading statements by first identifying candidate claims, then filtering out language that corrects those claims (debunking language), and finally refining the results to ensure alignment with the intended task. By mitigating the contamination observed in raw LLM outputs, our method delivers more accurate and reliable extraction of false information.

E.4 Clustering False Information

To optimize the performance of our clustering algorithms, we conducted a grid search over a range of hyperparameters for UMAP, PCA, HDBSCAN, and K-means. The ranges explored are summarized in Table 7.

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

Table 7: Hyperparameter ranges for clustering methods

The optimal hyperparameters for each algorithm are listed in Table 8.

Algorithm	**Optimal Hyperparameters**

HDBSCAN (UMAP)	min_cluster_size = 25
	min_samples = 20

K-means (UMAP)	n_clusters = 445
	min_samples = 15

UMAP (HDBSCAN)	n_neighbors = 15
	n_components = 256

UMAP (K-means)	n_neighbors = 15
	n_components = 256

Table 8: Optimal hyperparameters for clustering methods. Key values for reproducibility available under the best-performing pair: HDBSCAN and UMAP.

E.5 Deriving Disinformation Narratives

In addition to the LLM-based narrative generation methods, we evaluated two non-LLM summarization baselines to assess their ability to derive disinformation narratives from clusters of false information.

Models and Setup We tested BERTsum (Ren et al., 2019), and TextRank (Almansa et al., 2022). Both models received the same cluster content inputs used by the LLMs and generated summaries intended to represent disinformation narratives. The resulting summaries were evaluated using the Weighted Chamfer Distance metric against expert-written narratives from the EUDisinfoTest dataset.

Results BERTsum achieved a WCD score of 0.61, while TextRank scored 0.59. These scores fall significantly below the performance of all LLM-based methods, which reached WCD scores above 0.70 (see Table 4).

Discussion The relatively lower performance of these non-LLM summarization baselines underscores the advantage of generative large language models in capturing the nuanced, context-rich nature of disinformation narratives.

E.6 Comparison with General Narrative Extraction Methods

Table 15 showcases examples of ground truth disinformation narratives from the EUDisinfoTest dataset alongside the closest matching narratives predicted by DiNaM, CaNarEx, and Relatio.

E.7 Generalizability

To evaluate DiNaM's robustness across diverse sources, we partitioned the EDMO dataset, which consists of fact-checking articles from 14 independent European organizations, into three mutually exclusive and balanced subsets, each containing articles from different sources. DiNaM was then applied independently to each subset. As shown in Table 9, performance remained consistent across subsets, with only minor variation relative to the full dataset. This demonstrates that DiNaM generalizes well across different sources.

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

Table 9: WCD scores for DiNaM across organizational subsets of the EDMO dataset. Consistent scores indicate strong generalization.

Moreover, we manually grouped the narratives identified by DiNaM (see Section 6.4) into seven main disinformation topics. For details on the annotation process and topic distribution, refer to Appendix C.3. To evaluate DiNaM's performance within each topic, we computed the WCD score between the predicted narratives and the corresponding ground truth narratives from EUDisinfoTest, filtering both sets by topic. As shown in Table 10, DiNaM demonstrates generalization across diverse disinformation domains.

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

Table 10: WCD scores for DiNaM across different topical categories. Scores indicate robust alignment between predicted and ground truth narratives (EUDisinfoTest) across diverse subject areas.

E.8 Cost Analysis of the DiNaM Algorithm

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

To assess the cost-efficiency of DiNaM, we estimated the computational cost of each LLM-based step using a dataset of 10,493 fact-checking articles. Following OpenAI's guideline that one token corresponds to approximately 0.75 words[12], we estimate a total of 33.02 million input tokens and 319.7 thousand output tokens across the pipeline.

Table 12 presents the estimated cost of running Steps 1, 2, and 4 using different models.

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

E.9 DiNaM as a General-Purpose Narrative Mining Framework

To evaluate DiNaM without fact-checking articles (as a general-purpose narrative mining framwork), we scraped 4,202 articles from sputnikglobe.com related to Russia's invasion of Ukraine, an outlet known for pro-Kremlin narratives. We compared DiNaM (excluding its first two fact-check specific stages) against Relatio and CaNarEx on this corpus.

We used trafilatura (Barbaresi, 2021) for article scraping and Selenium (Selenium, 2024) for HTML parsing.

Performance was measured using the WCD metric against a set of ground-truth disinformation narratives from EUDisinfoTest (Sosnowski et al., 2024) about the war in Ukraine. DiNaM achieved a WCD score of 0.67, outperforming both Relatio (0.62) and CaNarEx (0.61). These results suggest that DiNaM remains effective even without relying on fact-checking articles, highlighting its strength as a general-purpose narrative mining framework.

F Detailed Analysis of Disinformation Narratives Derived by DiNaM

F.1 Climate Change Disinformation Narratives

Figure 11 shows a sharp rise in climate change disinformation starting in early 2022, led by claims that "Climate change is fake". Other themes include critiques of electric cars, conspiracy theories, and assertions that natural disasters are deliberately engineered. Rising energy costs and government incentives for green technologies (Commission, 2024b) likely fueled skepticism. In 2023, severe heatwaves and wildfires (, C3S) likely fueled narratives asserting that wildfires result from human activity unrelated to climate change.

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

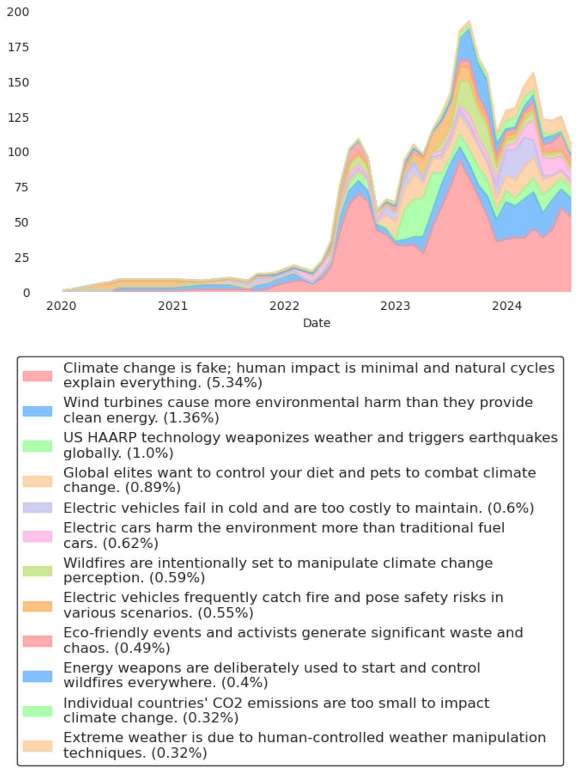


Figure 11: Trends in climate change disinformation narratives

F.2 Migrants Disinformation Narratives

DiNaM uncovered several disinformation narratives related to migration (Figure 12). The most prominent claim that immigrants receive more financial support than locals, while others link migrants to rising violence in Europe. These narratives surged in early 2022, likely driven by geopolitical instability and the economic fallout of COVID-19 (Moroz et al., 2020). The Russian invasion of Ukraine further strained European infrastructure, especially in the East (Fernando GUTIERREZ, 2022).

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

F.3 Israel-Palestine War Disinformation Narratives

Several disinformation narratives have emerged in connection to the Israel-Palestine war (Figure 13) as identified by DiNaM. One dominant claim suggests that "Gaza war scenses are fake and orchestrated for sympathy." while another asserts that "Global unrest as world protests against Israel supporting Palestine." These narratives gained significant traction in late 2023 and early 2024, aligning with the intensification of the conflict and the international reaction to the humanitarian crisis (ReliefWeb and News, 2023).

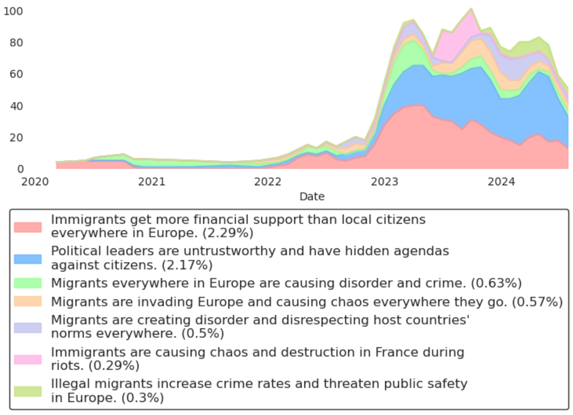


Figure 12: Evolution of disinformation narratives on the migration

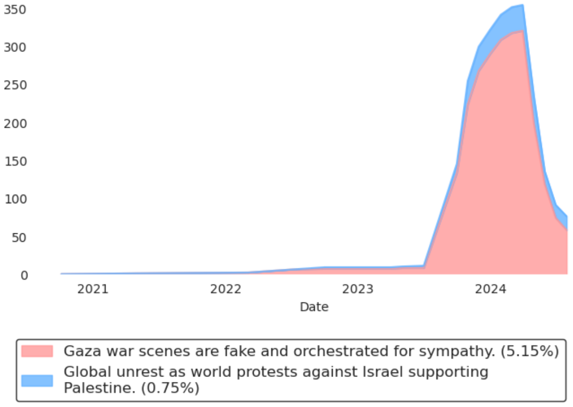


Figure 13: Evolution of Disinformation Narratives on the Israel-Palestine War

F.4 European Union Disinformation Narratives

DiNaM's analysis of disinformation narratives reveals a surge in targeting the European Union in recent years, particularly during critical events such as the forthcoming European Parliamentary elections in June 2024 (Council of the European Union, 2023). Figure 14 illustrates the evolution of these narratives. A prominent example includes claims that "Europe is mismanaged with excessive taxes, inequality, and poor living conditions." which serves as a broad critique of EU governance. Another widely circulated narrative alleges that "EU hides insects in food to trick people into eating them." a distortion likely linked to the European Commission’s 2023 approval of insect-derived products as voluntary food ingredients (European Commission, 2023).

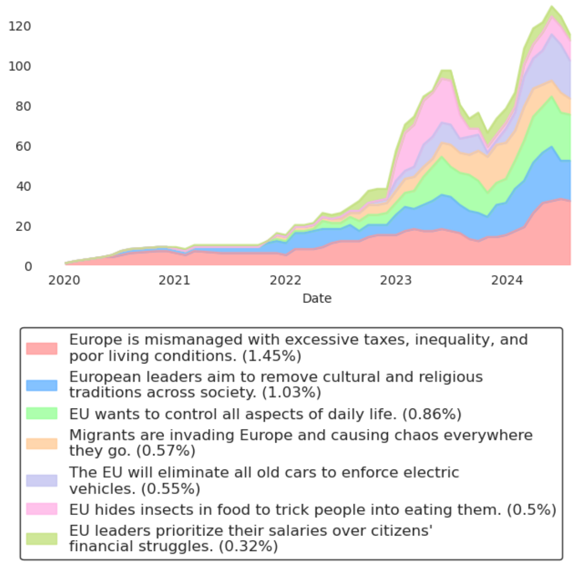


Figure 14: Evolution of Disinformation Narratives on the European Union

Model	Prompt	F1 Score
Zero-shot		
GPT-4o-mini	Figure 15	**0.88**
GPT-4o-mini	Figure 16	0.87
GPT-4o-mini	Figure 17	0.87
Llama-3.3-70B	Figure 15	0.85
Llama-3.3-70B	Figure 16	0.86
Llama-3.3-70B	Figure 17	0.87
Gemma-3-27B	Figure 15	0.87
Gemma-3-27B	Figure 16	0.87
Gemma-3-27B	Figure 17	0.87
Qwen3-32B	Figure 15	**0.88**
Qwen3-32B	Figure 16	0.87
Qwen3-32B	Figure 17	0.86
Fine-tuned		
DeBERTa-large	–	0.60
RoBERTa-large	–	0.58

Table 11: F1 scores for filtering false-claim detection by model and prompt. Zero-shot results use each prompt variant. Additionally, fine-tuned baselines are included for reference.

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

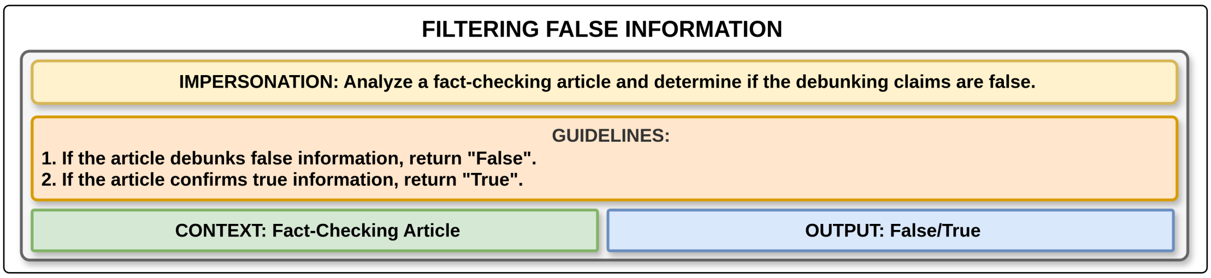
Table 12: Estimated LLM costs for Steps 1, 2 and 4 of the DiNaM algorithm, assuming total token usage of 33.02M input and 319.7K output tokens.

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

Table 13: Overview of DiNaM's pipeline with best-performing models and input/output quantities.

Model Name	API/HuggingFace Model Name	Access Details	License	Model Size
GPT-4o-mini	`gpt-4o-mini-2024-07-18`	OpenAI API 10.2024	Commercial	Not Disclosed
Gemma3-27B	`google/gemma-3-27b-it`	DeepInfra API 05.2025	Gemma License	27B
Llama3.3-70B	`meta-llama/Llama-3.3-70B-Instruct`	DeepInfra API 05.2025	Llama 3.3 Community License	70B
Qwen3-32B	`Qwen/Qwen3-32B`	DeepInfra API 05.2025	Apache 2.0	32.8B
SFR-embedding-2	`Salesforce/SFR-Embedding-2_R`	HuggingFace 10.2024	CC BY-NC 4.0	7B
E5-Large	`intfloat/e5-large`	HuggingFace 05.2025	MIT	500M
jina-embeddings-v3	`jinaai/jina-embeddings-v3`	HuggingFace 05.2025	CC BY-NC 4.0	570M
RoBERTa-Large	`facebook/roberta-large`	HuggingFace 05.2025	MIT	355M
DeBERTa-Large	`microsoft/deberta-large`	HuggingFace 05.2025	MIT	∼390M
RoBERTa-SQD2	`deepset/roberta-base-squad2`	HuggingFace 05.2025	CC BY 4.0	124M
DistilBERT-SQD	`distilbert/distilbert-base-cased-distilled-squad`	HuggingFace 05.2025	Apache 2.0	65M

Table 14: Detailed overview of used language models.



FILTERING FALSE INFORMATION

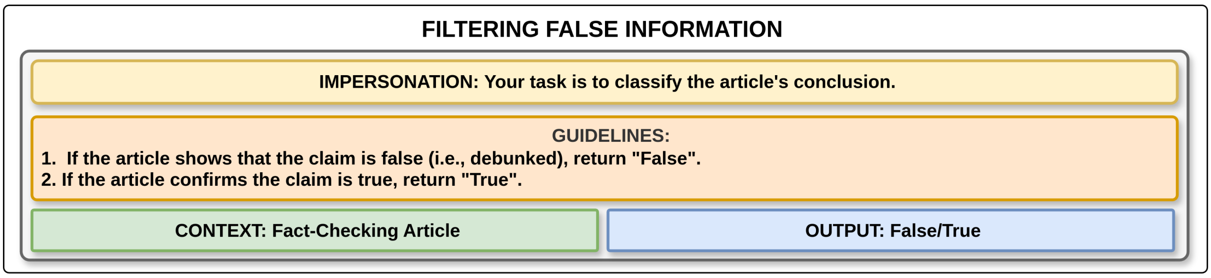
IMPERSONATION: Analyze a fact-checking article and determine if the debunking claims are false.

GUIDELINES:
1. If the article debunks false information, return "False".
2. If the article confirms true information, return "True".

CONTEXT: Fact-Checking Article

OUTPUT: False/True

Figure 15: A prompt designed for filtering fact-checking articles, enabling the identification of whether the claims under review are true or false.



FILTERING FALSE INFORMATION

IMPERSONATION: Your task is to classify the article's conclusion.

GUIDELINES:
1. If the article shows that the claim is false (i.e., debunked), return "False".
2. If the article confirms the claim is true, return "True".

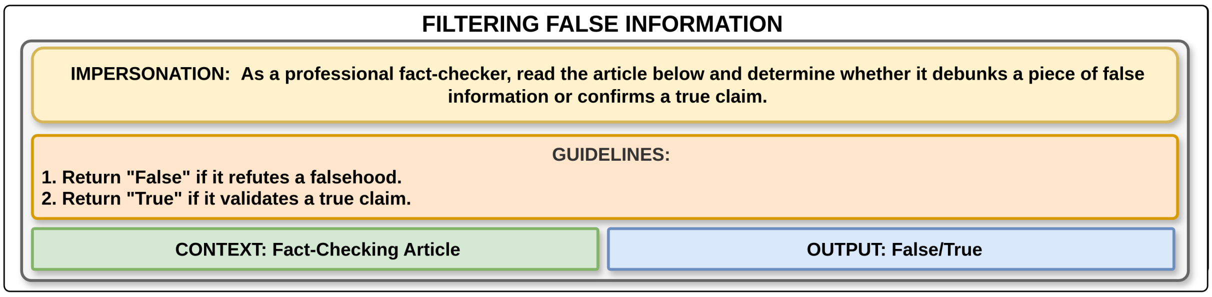
CONTEXT: Fact-Checking Article

OUTPUT: False/True

Figure 16: A second prompt designed for filtering fact-checking articles, enabling the identification of whether the information being reviewed is true or false.

Ground Truth Narrative	DiNaM (Fact-Checking Articles)	CaNarEx (Fact-Checking Articles)	CaNarEx (False Information)	Relatio (Fact-Checking Articles)	Relatio (False Information)
COVID-19 Vaccines are very dangerous/lethal	Covid vaccines are dangerous and cause many deaths and health problems.	COVID remains a global threat and vaccines are safe	The increase in all-cause mortality is caused by COVID-19 vaccines.	vaccine kill people	the vaccine increase death
Climate change is not real and/or is not related to human activities	Climate change isn't real; it's just natural cycles and media exaggeration.	In fact, there is no evidence to support that the news is true.	Human emissions are too small to affect the climate.	deny climate change	discredit climate change
Climate movements are hypocritical and/or foolish	Climate activists are causing more pollution than they fight against.	Social media users see a heavily burning bus as proof that electric vehicles are dangerous.	The video shows a deliberately caused fire, implying that a deliberately caused fire is evidence of climate change being exaggerated.	they burn climate change	the protest cause climate change
Renewables, recycling, and electric vehicles are useless or dangerous	Electric cars and renewable energy are worse for the environment than you think.	Social media users see a heavily burning bus as proof that electric vehicles are dangerous.	The EU's energy efficiency plans will make buildings worthless or uninhabitable for the population.	refuse electric car	which ban electric car
Ukrainians and Ukrainian forces are largely pro-Nazi	Ukraine is secretly a Nazi country with leaders showing Nazi symbols and behavior.	The relevant Ukrainian state authorities and security forces also reported on The incident on social networks, for example here or here.	Volodymyr Zelensky publicly supports neo-Nazi or ultra-right movements.	they use ukrainian refugee	ukrainian soldier cause ton

Table 15: Examples of ground truth narratives, sourced from the test set described in Section 6.4, with the closest matches generated by DiNaM, CaNarEx, and Relatio. These methods were evaluated on EDMO fact-checking articles (EDMO, 2024b). Additionally, CaNarEx and Relatio were evaluated on a false information dataset, derived from the same set of articles (for more details, see Section 4.2). The semantic similarity between the ground truth and the predicted narratives was measured using cosine similarity, based on embeddings from the SFR-embedding-2 model (Rui Meng, 2024).



FILTERING FALSE INFORMATION

IMPERSONATION: As a professional fact-checker, read the article below and determine whether it debunks a piece of false information or confirms a true claim.

GUIDELINES:

1. Return "False" if it refutes a falsehood.
2. Return "True" if it validates a true claim.

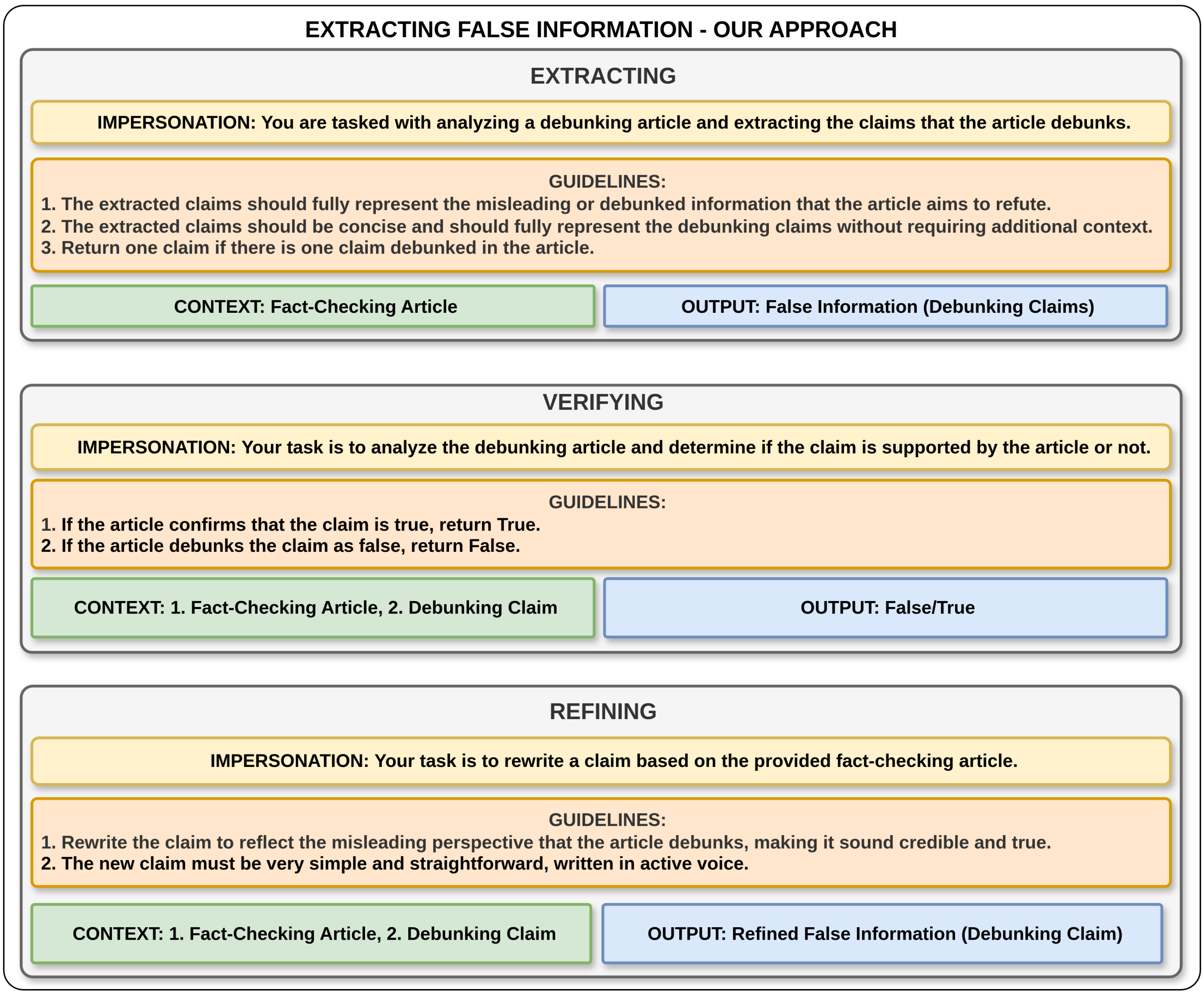
CONTEXT: Fact-Checking Article

OUTPUT: False/True

Figure 17: A third prompt designed for filtering fact-checking articles, enabling the identification of whether the information being reviewed is true or false.

Ground Truth	**COVE**	**COT**	**BASE**	**OUR**
Images and videos are showing current violence against civilians in Gaza, such as portraying Palestinian victims.	The photo being circulated as victims of the current conflict in Gaza is actually from a 2013 poison gas attack in Syria.	An image of children's corpses wrapped in white cloths circulating on social media is from the current conflict in Gaza and related to children killed in the recent attacks.	A photo claiming to show children killed in the current conflict in Gaza is actually from 2013 and depicts victims of a poison gas attack in Syria.	A recent photo shows the tragic victims of Israel's attacks in Gaza, highlighting the impact on innocent children.
USA Today reported a fight involving the Ukrainian delegation in New York.	A video showing Ukrainians involved in a violent brawl was fabricated and falsely attributed to USA Today.	A fabricated video falsely attributes a claim to USA TODAY that suggests a member of the Ukrainian delegation to the UN General Assembly was involved in a bar fight, implying that this incident reflects a violent and unethical nature within Ukrainian leadership and culture.	The video attributed to USA Today portrays Ukrainians as violent.	The video attributed to USA Today shows Ukrainians engaging in violent behavior.
The European Commission is considering a ban on repairing vehicles over 15 years old.	The European Union wants to ban repairs on vehicles older than 15 years.	The European Union presents a controversial legislative proposal that would prohibit the right to repair for vehicles older than 15 years.	The European Union proposes to ban the repair of vehicles over 15 years old.	The European Union proposes to ban the repair of vehicles over 15 years old.
Videos show Italian farmers in Piedmont protesting against the government.	The video that claims to show Italian farmers in the Piedmont region protesting against inflation and high living costs actually depicts a 'beer and tractor festival' with no connection to any protest.	A video circulating on social media falsely claims that Italian farmers in the Piedmont region are protesting against inflation, high fuel prices, and low purchasing prices of products, while the video actually depicts a 'beer and tractor festival' organized in Villafranca Piemonte.	A video showing tractors in Italy is not evidence of farmers protesting against inflation, but rather footage from a 'beer and tractor festival' organized in Piedmont.	The video captures a large gathering of Italian farmers protesting against high inflation and rising costs.
Zelensky is using cocaine.	The claim that President Zelenskyj is a drug addict is false and part of Russian propaganda, with manipulated videos used to support this disinformation.	Claims that Ukrainian President Volodymyr Zelenskyj is a drug addict, which are part of Russian propaganda.	The claim that President Zelenskyj is a drug addict is false and part of Russian propaganda.	Videos show President Zelenskyj using drugs, confirming he's a drug addict.
Images show Pope Francis in a puffy, bright white coat.	Pope Francis was never photographed wearing a puffer jacket; the image suggesting he did was generated by an Artificial Intelligence tool.	A claim that Pope Francis was photographed in a puffer jacket, a style not actually worn by him, generated by an Artificial Intelligence tool.	An image of Pope Francis wearing a puffer jacket and appearing to rap is generated by an AI tool and does not represent a real event; the Pope has never worn such attire or been photographed in that scenario.	Pope Francis wears a stylish puffer jacket while preparing to rap in a new viral image.

Table 16: Examples of extracted false information compared across different methods: COVE, COT, BASE, and OUR. Each method is tasked with extracting false information from fact-checking articles. Both the articles used by the models to extract false information and the ground truth false information presented in the first column are derived from the test set described in Section 6.2. Text highlighted in red indicates corrective (debunking) content mistakenly included by the respective method, where the extracted claim provides corrective or factual details that contradict the misleading nature of the original ground truth statement.



EXTRACTING FALSE INFORMATION - OUR APPROACH

EXTRACTING

IMPERSONATION: You are tasked with analyzing a debunking article and extracting the claims that the article debunks.

GUIDELINES:
1. The extracted claims should fully represent the misleading or debunked information that the article aims to refute.
2. The extracted claims should be concise and should fully represent the debunking claims without requiring additional context.
3. Return one claim if there is one claim debunked in the article.

CONTEXT: Fact-Checking Article

OUTPUT: False Information (Debunking Claims)

VERIFYING

IMPERSONATION: Your task is to analyze the debunking article and determine if the claim is supported by the article or not.

GUIDELINES:
1. If the article confirms that the claim is true, return True.
2. If the article debunks the claim as false, return False.

CONTEXT: 1. Fact-Checking Article, 2. Debunking Claim

OUTPUT: False/True

REFINING

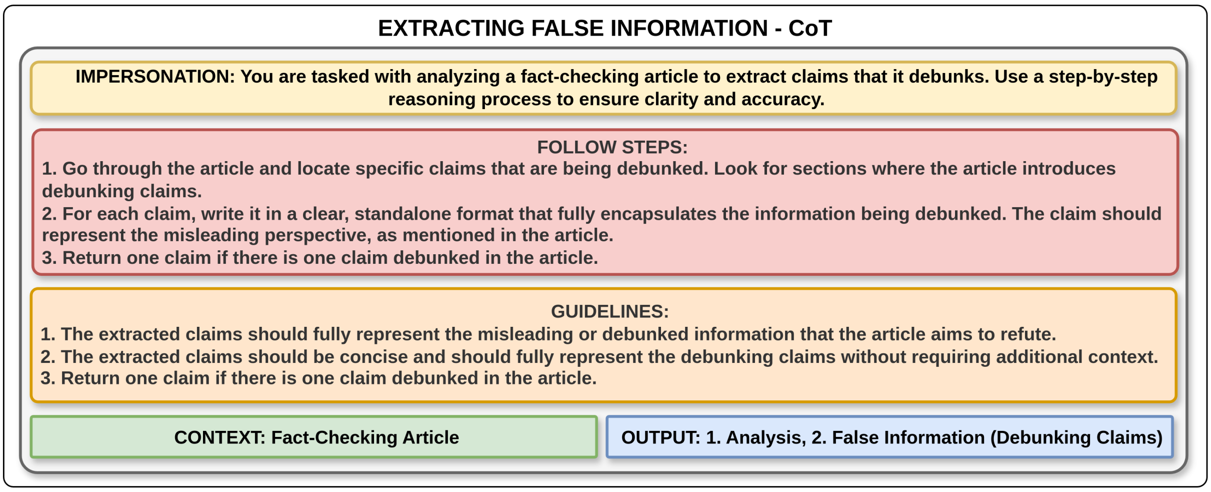
IMPERSONATION: Your task is to rewrite a claim based on the provided fact-checking article.

GUIDELINES:
1. Rewrite the claim to reflect the misleading perspective that the article debunks, making it sound credible and true.
2. The new claim must be very simple and straightforward, written in active voice.

CONTEXT: 1. Fact-Checking Article, 2. Debunking Claim

OUTPUT: Refined False Information (Debunking Claim)

Figure 18: Prompts for each substep of our approach for false information extraction



EXTRACTING FALSE INFORMATION - CoT

IMPERSONATION: You are tasked with analyzing a fact-checking article to extract claims that it debunks. Use a step-by-step reasoning process to ensure clarity and accuracy.

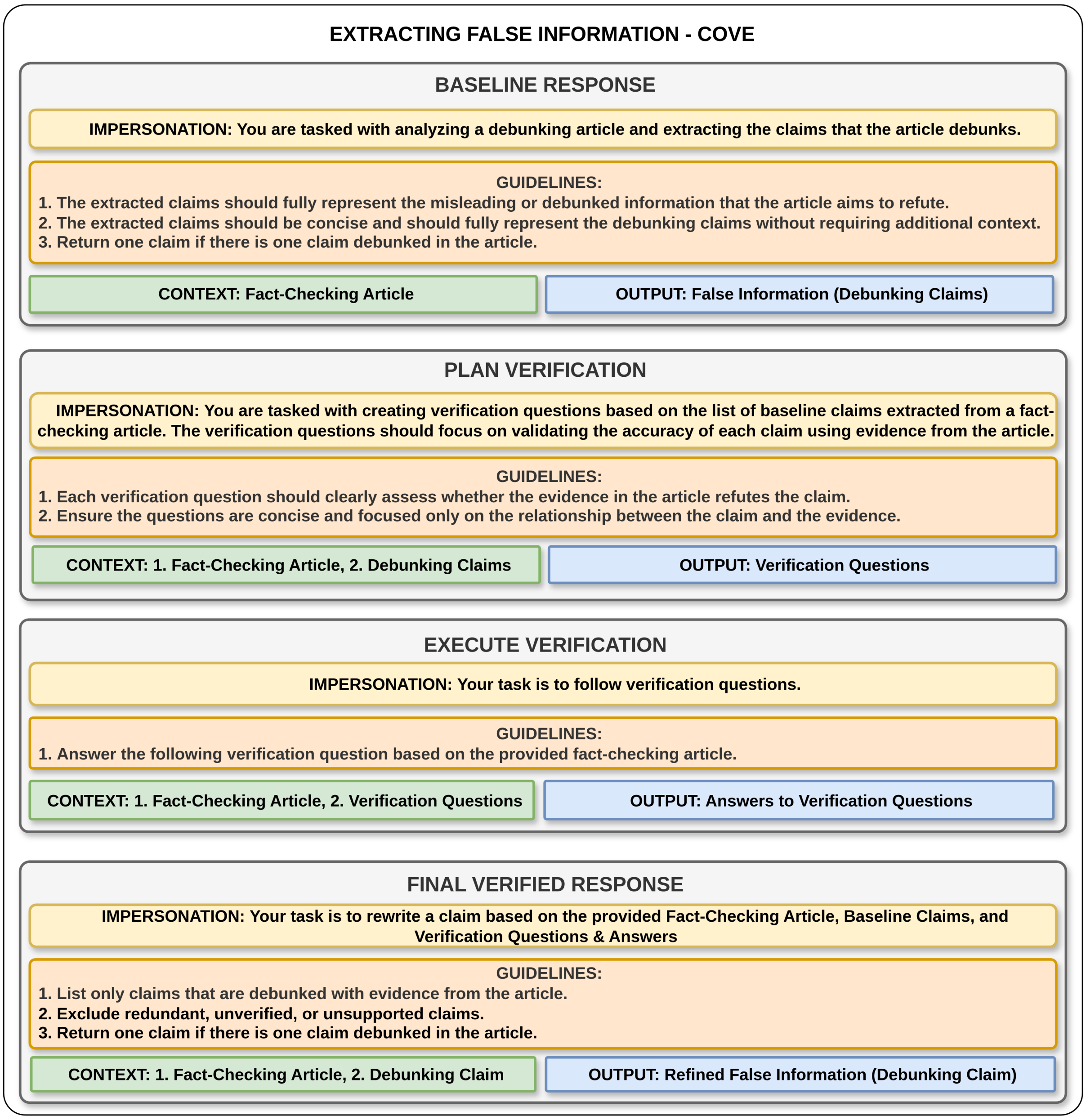
FOLLOW STEPS:
1. Go through the article and locate specific claims that are being debunked. Look for sections where the article introduces debunking claims.
2. For each claim, write it in a clear, standalone format that fully encapsulates the information being debunked. The claim should represent the misleading perspective, as mentioned in the article.
3. Return one claim if there is one claim debunked in the article.

GUIDELINES:
1. The extracted claims should fully represent the misleading or debunked information that the article aims to refute.
2. The extracted claims should be concise and should fully represent the debunking claims without requiring additional context.
3. Return one claim if there is one claim debunked in the article.

CONTEXT: Fact-Checking Article

OUTPUT: 1. Analysis, 2. False Information (Debunking Claims)

Figure 19: Prompts for our implementation of Chain-of-Thought for false information extraction



EXTRACTING FALSE INFORMATION - COVE

BASELINE RESPONSE

IMPERSONATION: You are tasked with analyzing a debunking article and extracting the claims that the article debunks.

GUIDELINES:
1. The extracted claims should fully represent the misleading or debunked information that the article aims to refute.
2. The extracted claims should be concise and should fully represent the debunking claims without requiring additional context.
3. Return one claim if there is one claim debunked in the article.

CONTEXT: Fact-Checking Article

OUTPUT: False Information (Debunking Claims)

PLAN VERIFICATION

IMPERSONATION: You are tasked with creating verification questions based on the list of baseline claims extracted from a fact-checking article. The verification questions should focus on validating the accuracy of each claim using evidence from the article.

GUIDELINES:
1. Each verification question should clearly assess whether the evidence in the article refutes the claim.
2. Ensure the questions are concise and focused only on the relationship between the claim and the evidence.

CONTEXT: 1. Fact-Checking Article, 2. Debunking Claims

OUTPUT: Verification Questions

EXECUTE VERIFICATION

IMPERSONATION: Your task is to follow verification questions.

GUIDELINES:
1. Answer the following verification question based on the provided fact-checking article.

CONTEXT: 1. Fact-Checking Article, 2. Verification Questions

OUTPUT: Answers to Verification Questions

FINAL VERIFIED RESPONSE

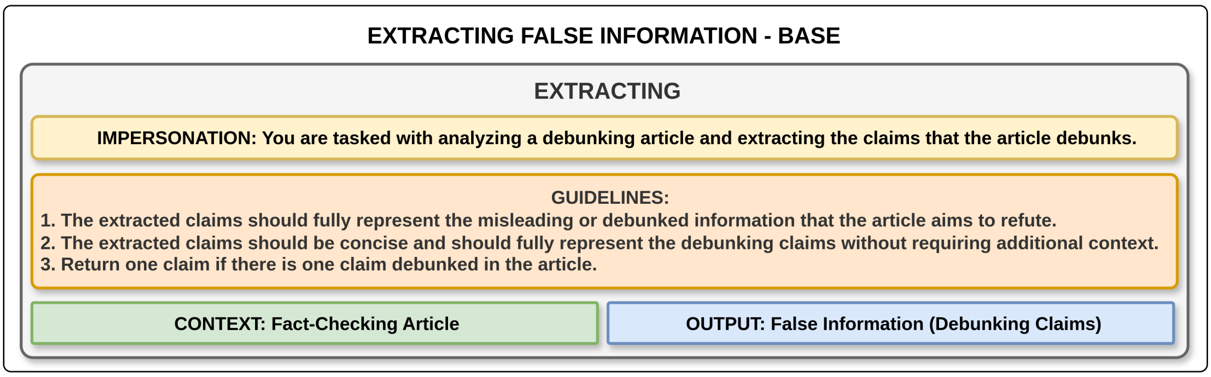
IMPERSONATION: Your task is to rewrite a claim based on the provided Fact-Checking Article, Baseline Claims, and Verification Questions & Answers

GUIDELINES:
1. List only claims that are debunked with evidence from the article.
2. Exclude redundant, unverified, or unsupported claims.
3. Return one claim if there is one claim debunked in the article.

CONTEXT: 1. Fact-Checking Article, 2. Debunking Claim

OUTPUT: Refined False Information (Debunking Claim)

Figure 20: Prompts for our implementation of CoVe for false information extraction



EXTRACTING FALSE INFORMATION - BASE

EXTRACTING

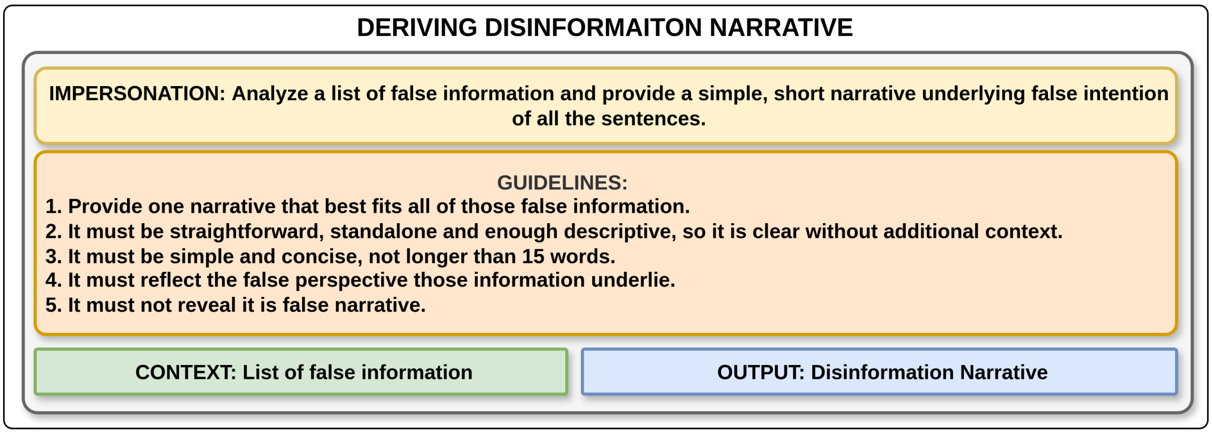
IMPERSONATION: You are tasked with analyzing a debunking article and extracting the claims that the article debunks.

GUIDELINES:
1. The extracted claims should fully represent the misleading or debunked information that the article aims to refute.
2. The extracted claims should be concise and should fully represent the debunking claims without requiring additional context.
3. Return one claim if there is one claim debunked in the article.

CONTEXT: Fact-Checking Article

OUTPUT: False Information (Debunking Claims)

Figure 21: Base prompts for false information extraction



DERIVING DISINFORMAITON NARRATIVE

IMPERSONATION: Analyze a list of false information and provide a simple, short narrative underlying false intention of all the sentences.

GUIDELINES:
1. Provide one narrative that best fits all of those false information.
2. It must be straightforward, standalone and enough descriptive, so it is clear without additional context.
3. It must be simple and concise, not longer than 15 words.
4. It must reflect the false perspective those information underlie.
5. It must not reveal it is false narrative.

CONTEXT: List of false information

OUTPUT: Disinformation Narrative

Figure 22: A prompt designed for deriving disinformation narrative given a set of false information