DiNO: Disinformation Narrative Observer

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 is an escalating global threat, making it essential to understand its content, dissemination, and evolution. To confront this challenge, researchers have begun grouping related false claims into broader disinformation narratives, which can be tracked across cultures, time periods, and media sources. Analyzing these narratives provides critical insights for developing more effective countermeasures. To this end, we introduce DiNO: Disinformation Narrative Observer, a novel method designed to extract disinformation narratives from news articles. We applied DiNO to news articles on the Ukraine War, COVID-19 and Migration, sourced from disinformation-prone outlets as well as a reputable source. We evaluated the narratives extracted by DiNO by measuring how well their topics and stances aligned with a recognized disinformation narratives dataset. DiNO outperforms competitive narrative mining approaches, including Relatio and CaNarEx, achieving a 41%–44% improvement in topical alignment and a 30%–41% improvment in stance alignment.

1 Introduction

The World Economic Forum's *Global Risks Report 2025*, based on its 2024-2025 survey, ranks misinformation and disinformation as the leading global risk over the next two years (Elsner et al., 2025). This is not surprising, as recent studies show that false information can spread rapidly (Esteban-Bravo et al., 2024), erode trust in institutions and mainstream media, and strain democratic processes (Kang and Sheen, 2025). Over time, such dynamics can shift public opinion and reshape political and social landscapes (Tandoc Jr, 2019). Coordinated campaigns that fuse multiple misleading claims into coherent disinformation narratives (Starbird et al., 2019) represent an even more significant threat. Moreover, Goel et al. (2023) shows that news outlets may unintentionally reinforce such narratives, highlighting the need to understand how disinformation circulates within and across news outlets. Building on this observation, we ask a pivotal question in this work: How can the disinformation narratives propagated by a given news outlet be extracted?

Answering this question requires a clear definition of a disinformation narrative. EUvsDisinfo[1] defines a narrative as *"an overall message, communicated through texts, images, metaphors, and other means"* (EUvsDisinfo, 2022) and as *"templates for particular stories and can be adapted to a target audience"* (EUvsDisinfo, 2019). Similarly, the European Digital Media Observatory (EDMO)[2] defines a disinformation narrative as *"the clear message that emerges from a consistent set of contents that can be demonstrated as false using the fact-checking methodology"* (EDMO, 2024).

Taken together, these definitions highlight three key characteristics of any disinformation narrative: **Incorrect basis.** A disinformation narrative is grounded on a set of contents that are proven false. **Multiple examples.** It emerges from several interconnected false claims rather than a single instance. **Adaptability.** It functions as an overarching pattern that can be applied to many related false claims, making it much broader than any single false claim.

Building on these insights, we hypothesize that a news outlet's disinformation narratives can be extracted by: (i) identifying article segments that align with verified false claims (**Incorrect basis**), (ii) clustering semantically similar false segments to aggregate recurring instances (**Multiple examples**), and (iii) synthesizing an overarching narrative for each cluster (**Adaptability**). Operationally, DiNO instantiates this hypothesis as the four-step method (see Figure 1).

Guided by this hypothesis, we pose two research

[1]EUvsDisinfo is the flagship project of the EU External Action Service's East StratCom Task Force, created to counter Russian disinformation campaigns. (EUvsDisinfo, 2024)

[2]EDMO, an EU-funded initiative at the European University Institute, combats disinformation. (EDMO)

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 46544–46574
July 2-7, 2026 ©2026 Association for Computational Linguistics

questions: **RQ1** *How accurately can disinformation narratives be extracted in news articles by (i) detecting articles' false segments that correspond to verified false claims, (ii) clustering semantically related false segments, and (iii) inferring narratives from those clusters?* **RQ2** *What disinformation narratives are spread by specific media outlets?*

To answer these RQs, we introduce **DiNO**: **Di**sinformation **N**arrative **O**bserver, a novel approach that systematically extracts disinformation narratives from news articles. We evaluate DiNO by comparing its extracted narratives with ground-truth disinformation narratives, and benchmark its performance against other narrative mining methods. Our contributions are as follows:

- We introduce DiNO, a method for extracting disinformation narratives from news articles by leveraging verified false-claims databases.
- We introduce DiNO-sgm, a method that matches segments from news articles to previously verified false claims, outperforming baselines.
- We benchmark DiNO against competitive narrative mining methods (CaNarEx (Anantharama et al., 2022), Relatio (Ash et al., 2024)).
- We evaluate DiNO across twelve datasets covering the Ukraine War, COVID-19, and Migration: nine datasets with ∼130k news articles and three datasets with ∼26k verified false claims.

For reproducibility, we release all datasets, prompts, annotations, narratives, and code[3].

2 Literature Review

Verified Claim Matching. Claim matching is a key ingredient for scalable disinformation analysis. Pikuliak et al. (2023) released MultiClaim, linking social media posts to fact-checks, while Shaar et al. (2020) cast claim matching as a ranking problem. Choi and Ferrara (2024) proposed FACT-GPT to train smaller models for claim matching, and Zheng et al. (2025) introduced AFEV, which decomposes claims into atomic facts to improve retrieval.

Narrative Mining. Relatio (Ash et al., 2024) uses Semantic Role Labeling to extract entities and actions and builds multigraphs to represent narratives. CaNarEx (Anantharama et al., 2022) improves coherence via co-reference resolution, while Miori and Petrov (2025) combine LLMs and graph methods to track narratives in economic news. DiNaM proposes a method for mining disinformation narratives from fact-checking articles, providing a global map of disinformation narratives (Sosnowski et al., 2025).

Disinformation Narratives in NLP. Recent work has emphasized datasets and analyses of narratives: Sosnowski et al. (2024) benchmarked LLMs for narrative classification. Modzelewski et al. (2024) introduced MIPD, Polish dataset focusing on intent types as a generalized form of disinformation narratives. Modzelewski et al. (2026) later extended this line of work by introducing MALINT, an English dataset with publicly available annotations from multiple stages of the annotation process. Santos (2023) explored linguistic and sentiment signals for countering false narratives. DIPROMATS (Moral et al., 2024) analyzed diplomatic tweets to detect narratives (Fraile-Hernández et al., 2024). More recently, Guimarães et al. (2025) introduced NarratEX, a multilingual dataset for explaining dominant narratives and sub-narratives in news texts. In closely related work, Nikolaidis et al. (2025) presented PolyNarrative, a multilingual and multi-label dataset for narrative extraction from news articles.

Summary and gap. Prior work tends to focus on one of three levels: (i) matching individual claims to fact-checks at scale, (ii) extracting general narratives from text, or (iii) recovering disinformation narratives from fact-checking articles, as in DiNaM. However, existing approaches do not provide a method that operates directly on news articles to mine disinformation narratives at the outlet level. DiNO addresses this gap by grounding outlet-level disinformation narrative extraction in news articles, enabling systematic characterization of narratives propagated by specific outlets.

3 DiNO Method

To answer **RQ1**, this section introduces **DiNO**, a method for recovering disinformation narratives grounded in verified false claims by linking news-article segments to fact-check repositories. This reliance on pre-existing fact checks is a deliberate design choice: it enables scalable, reproducible, and auditable analysis by ensuring that extracted narratives can be explicitly attributed to curated false-claim records, a common foundation in large-scale disinformation research (Sánchez del Vas and Tuñón Navarro, 2024).

DiNO extends **DiNaM** (Sosnowski et al., 2025) from fact-checking articles to news articles, enabling outlet-level analysis of disinformation nar-

[3] https://github.com/wsosnowski/DiNO/tree/main

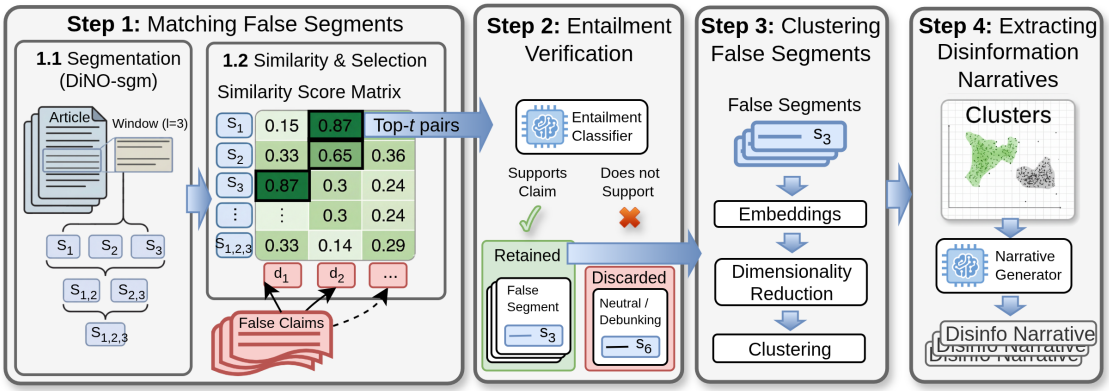

Figure 1: **The DiNO Method.** The approach consists of four phases: (1) *Matching False Segments*, where article text is segmented via the DiNO-sgm method and matched to verified false claims, (2) *Entailment Verification*, where an LLM confirms that the article explicitly supports the false claim,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

ratives. This shift is methodologically non-trivial: whereas fact-checking articles are already centered on verified falsehoods, news articles typically embed potentially false claims within broader contextual reporting, quotations, and commentary. DiNO therefore introduces a segment-level pipeline that retrieves candidate spans, verifies them against known false claims via entailment, clusters the verified false segments, and synthesizes them into outlet-level narratives. As a result, DiNO complements DiNaM's global view of narratives derived from fact-check coverage with a finer-grained perspective on how repository-grounded disinformation narratives appear in news content.

3.1 Matching False Segments

In Step 1 (Figure 1), DiNO matches article segments that may contain previously verified false claims by comparing the article text against these claims. Following standard document-level claim matching, we (i) split each article into candidate segments and (ii) compute their semantic similarity to verified false claims (Shaar et al., 2022).

Segmentation. While sentence-level segmentation is common (Shaar et al., 2022), it can be too narrow, since false information may span multiple sentences or depend on cross-sentence context (Hameleers, 2023). Conversely, longer passages can dilute the similarity signal with irrelevant context (Chen et al., 2024). To strike a balance, we introduce **DiNO-sgm** segmentation method.

First, each article is split into overlapping windows of l sentences with an o-sentence overlap. Second, within each window, we generate all contiguous subsequences from 1 to l sentences.

For example, for a window of length l=3 containing sentences S_1, S_2, S_3, DiNO-sgm produces six candidate segments: (S_1), (S_2), (S_3), (S_1, S_2), (S_2, S_3), and (S_1, S_2, S_3). Compared to fixed-length segmentation, this richer candidate set better captures both short and multi-sentence spans.

Similarity. After generating candidate segments, DiNO computes a semantic similarity score between each segment and every verified false claim from the relevant database. For each article, it keeps the top-t segment-claim pairs by similarity score as candidates for next steps.

3.2 Entailment Verification

While the previous step matches article segments that are semantically similar to verified false claims, similarity alone is insufficient for reliable false-information detection: NLP models can assign high similarity to text that either supports a claim or argues against it (Mahabadi et al., 2019). Therefore, Step 2 performs article-level entailment verification to determine whether the article supports each matched claim and filters out non-supporting matches. Only segments whose articles pass this check are carried forward to the subsequent steps.

3.3 Clustering False Segments

A disinformation narrative can span multiple related false claims. In our setting, this means a single narrative may involve several false segments matched to thematically similar false claims. Accordingly, in Step 3 DiNO clusters the false seg-

ments extracted in the previous phase. Each segment is first embedded in a semantic vector space, reduced in dimensionality to address the curse of dimensionality (Grootendorst, 2022), and then clustered into sets that represent distinct narratives.

3.4 Extracting Disinformation Narratives

In the final step, DiNO employs an LLM to identify recurring patterns of falsehood within each cluster and distill a representative disinformation narrative. To do this, all texts from a given cluster are concatenated and provided as a single input to the LLM. The model is then prompted to synthesize these texts into a single narrative that: (1) encapsulates the cluster's theme, (2) is clear, self-contained, short and (3) shows the inherent perspective without indicating that it is false or true. This formulation is compatible with the narrative format used in SemEval 2025 Task 10 (Piskorski et al., 2025). Appendix I provides examples of extracted narratives alongside their constituent false claims.

4 Dataset

Having introduced the DiNO method, we now describe the datasets used to evaluate it. This study analyzes disinformation narratives on the Russo–Ukraine War, COVID-19, and Migration. For all domains, we collected news articles from disinformation-prone and reputable sources, along with verified false claims.

- **Disinformation Articles.** We collected full-text articles from sources known for spreading disinformation: Sputnik, Pravda, NaturalNews and Breitbart (Hellman, 2024; VIGINUM, 2024; Media Bias/Fact Check, 2025; Benkler et al., 2018).
- **Reputable Articles.** As credible baseline, we collected full-text articles from The Guardian, known as a credible news source (Newman et al., 2025).
- **Verified False Claims.** For COVID-19 and Migration, we gathered verified false claims via the Google Fact Check Explorer (GFCE) API (Google, 2025b), which provides the false claim text together with the associated fact-check entry. For the Ukraine War, false claims were sourced from EUvsDisinfo dataset (EUvsDisinfo, 2024), which provides a complete claim-source-debunking record: (1) the false claim text, (2) the source article where the claim appears, and (3) the linked debunking entry with disproof text and links to related debunk.

See Table 1 for dataset summaries and Appendix A for further data collection details.

Dataset	Items	Avg Wds.	Website URL
Ukraine War			
Sputnik	20 001	585.55	(Sputnik, 2025c)
Pravda	22 695	223.08	(News Pravda, 2025)
The Guardian	11 841	886.73	(The Guardian, 2025c)
EUvsDisinfo	15 452	38.80	(EUvsDisinfo, 2024)
COVID-19			
Sputnik	13 107	440.07	(Sputnik, 2025a)
NaturalNews	16 789	836.69	(NaturalNews, 2025)
The Guardian	10 573	820.21	(The Guardian, 2025a)
GFCE	6 381	28.91	(Google, 2025b)
Migration			
Sputnik	9 028	403.32	(Sputnik, 2025b)
Breitbart	16 203	512.81	(Breitbart, 2025)
The Guardian	11 671	866.51	(The Guardian, 2025b)
GFCE	4 452	19.05	(Google, 2025b)

Table 1: Summary of datasets, their sizes, mean word count per article/claim (Avg Wds.), and sources.

5 Evaluation

Building on the datasets, we evaluate DiNO in two stages: (i) We evaluate each step in isolation (Sections 5.1-5.4) to tune hyperparameters and select default components. (ii) Using the resulting configuration, we run DiNO across all outlets and compare resulting narratives against the ground-truth narratives (Section 5.6). For robustness, we report metrics averaged over three runs.

Additional details, including (1) non-LLM baselines, (2) hyperparameter search, (3) runtime and cost estimates, and (4) a DiNO overview with per-step input/output counts and optimal models/parameters, are in Appendix C; annotation guidelines are in Appendix J.

5.1 Matching False Segments

Step 1 locates the article segments that support false claims by (i) segmenting the article into candidate segments and (ii) matching each segment against a pool of verified false claims using a similarity measure. We evaluate this step by measuring how reliably DiNO retrieves the correct claim under different segmentation strategies and similarity measures.

Testing Dataset. To perform this evaluation, we need articles paired with gold-standard false claim(s). In EUvsDisinfo, entries link verified false claims to source articles, but the same article can appear in multiple entries. We therefore grouped entries by source article, so each article is associated with all of its linked false claims. We then

sampled 180 source articles, yielding a test set in which some articles are associated with multiple claims (1.3 claims per article on average).

Segmentation Strategies. We compare four segmentation strategies: (1) DiNO-sgm (ours), (2) Sentence-level: each sentence as a segment, (3) Passage-level: fixed-length segments, (4) Article-level: the full article as a single segment.

Similarity Measures. We test three similarity measures: (1) BM25 (Robertson and Zaragoza, 2009), (2) BERTScore (Zhang et al., 2020), (3) Cosine similarity over embeddings: SFR-Embedding-2 (SFR) (Meng et al., 2024), E5-large (E5) (Wang et al., 2022), and Jina (Sturua et al., 2024).

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

Results. Table 2 summarizes the results. Our novel segmentation method, DiNO-sgm, using tuned parameters (l=4, o=1, t=3) steadily achieves the highest mAP@3 across all similarity measures. We attribute this gain to DiNO-sgm's variable-length segments, which better match the span of false information than fixed-length alternatives. We therefore use DiNO-sgm as DiNO's default segmentation module.

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

Table 2: mAP@3 by segmentation strategy and similarity metric. For Cosine similarity, the best embedding model for each strategy is shown in parentheses.

Figure 2 shows the same trend: increasing the number of segments (via o and l) improves mAP@3. Example matches are in Appendix H.

5.2 Entailment Verification

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

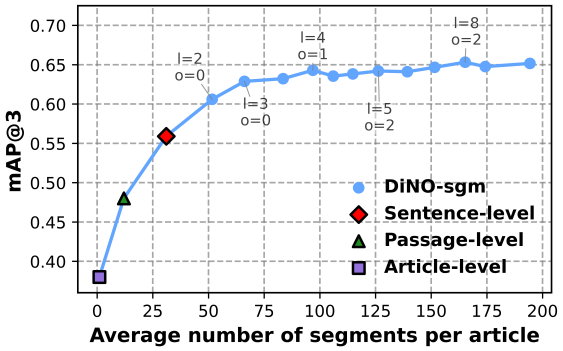

Figure 2: mAP@3 performance of DiNO–sgm (Jina) across segmentation settings with retrieval depth t=3. The figure shows how varying overlap (o) and segment length (l) affects matching accuracy and the number of segments produced. Results from other segmentation methods are included for baseline comparison.

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

Methods. We evaluate closed-weight LLMs (GPT-5, Sonnet 4.5, Gemini 2.5 Pro, GPT-5-mini) and open-weight models (gpt-oss-120b, Llama-3.3-70B, Qwen3-14B) in a zero-shot setting.[4] We use both our custom prompts and standard, general-purpose entailment prompts (Appendix B.2).

Evaluation. For each claim–article pair, the model predicts whether the article *supports* the claim (i.e., presents it as true) or *does not support* it (e.g., contradicts, questions, or does not endorse the claim). Articles from disinformation-prone outlets are expected to *support* the corresponding false claims, while articles from credible outlets are expected to *not support* them. We report F_1 and precision on the *support* class (P_{support}).

[4]We do not use few-shot classification (Brown et al., 2020) because including full-article examples in the prompt is impractical.

Results. Table 3 shows that GPT-5 and Sonnet 4.5 perform best (F_1=0.94, $P_{support}$=0.90), with GPT-5-mini close behind (0.93/0.88). Since entailment verification is token-intensive step and GPT-5-mini is cheaper with comparable accuracy, we use it as the default model. Among open-weight models, gpt-oss-120b achieves competitive results.

See Appendix F for additional analyses supporting full-article entailment verification.

Method	Prompt	F_1 score	$P_{support}$
Closed-weight LLMs			
Sonnet 4.5	Fig. 12	**0.94**	**0.90**
GPT-5	Fig. 13	**0.94**	**0.90**
Gemini 2.5 Pro	Fig. 11	0.93	0.89
GPT-5-mini	Fig. 11	0.93	0.88
Open-weight LLMs			
gpt-oss-120b	Fig. 13	0.93	0.88
Llama-3.3-70B	Fig. 11	0.91	0.87
Qwen3-14B	Fig. 11	0.88	0.84

Table 3: $P_{support}$ and F_1 scores for entailment verification using LLMs with best-performing prompts.

5.3 Clustering of False Segments

Step 2 yields a set of false segments, and Step 3 groups them into clusters. DiNO performs this step by embedding each segment, reducing dimensionality, and clustering the resulting representations. Accordingly, we evaluate Step 3 by measuring how clustering quality changes with the choice of embedding model, dimensionality reduction method, and clustering algorithm.

Testing Dataset. To evaluate clustering, we construct a test set of false segments. Specifically, we run Steps 1 and 2 of the DiNO algorithm on the Sputnik-Ukraine dataset using the optimal configuration, which produces 11,225 segments.

Methods. We evaluate three embedding models (SFR-Embedding-2, E5-large, Jina), two dimensionality reduction methods (UMAP (McInnes et al., 2018), PCA (Abdi and Williams, 2010)), and two clustering algorithms (HDBSCAN (McInnes et al., 2017), K-means (Hartigan and Wong, 1979)).

Evaluation. We perform a grid search over all combinations of embedding models, dimensionality reduction techniques, and clustering algorithms. Clustering quality is assessed using the Silhouette Score (Rousseeuw, 1987), a standard clustering metric that favors compact, well-separated clusters.

Results. Table 4 shows that SFR-Embedding-2 with UMAP and HDBSCAN performs best. This is consistent with prior work: HDBSCAN can identify clusters of varying shapes and label outliers as noise (McInnes et al., 2017), and UMAP preserves local neighborhood structure better than linear projections such as PCA (Sainburg et al., 2021). We therefore adopt SFR-Embedding-2 + UMAP + HDBSCAN as the default clustering configuration.

Additional human evaluation of cluster coherence is provided in Appendix C.9.

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

Table 4: Silhouette scores across embedding models, dimensionality reduction, and clustering methods.

5.4 Extracting Disinformation Narratives

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

Testing Dataset. We reuse the Sputnik-Ukraine clusters produced in Section 5.3 under the best clustering configuration. As ground truth, we use EUDisinfoTest (Sosnowski et al., 2024), which provides expert-curated disinformation narratives grounded in reports from 20 countries and spanning the domains analyzed in this paper: COVID-19, Migration, and Ukraine War.

Models. We evaluate four closed-weight LLMs (GPT-5, GPT-5-mini, Sonnet 4.5, Gemini 2.5 Pro) and three open-weight alternatives (gpt-oss-120b, Llama-3.3-70B, Qwen3-14B), guided by the same custom prompt (Appendix B.3).

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

We score each matched pair along two dimensions: **Topical Alignment** (TA), which measures

semantic overlap, and **Stance Alignment** (SA), which measures whether both narratives express the same viewpoint. SA is necessary because two narratives may be topically similar yet take opposite positions (Blackburn et al., 2020) (e.g., "*COVID-19 is engineered*" vs. "*There is no evidence that COVID-19 is engineered*"). Following prior work, we score TA and SA on a 0–5 scale (Agirre et al., 2016) and report normalized mean alignment:

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

$$mSA = \frac{1}{5N} \sum_{i=1}^{N} SA(n_i^p, n_{m(i)}^{gt}), \quad (2)$$

where N is the number of predicted narratives.

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

$$Rd = \frac{N}{U}, \quad (3)$$

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

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

As a scalable secondary evaluation, we also compute TA/SA using an LLM-as-a-judge (Zheng et al., 2023) and report mTA_A and mSA_A. We use Sonnet 4.5 as the judge due to its strong agreement with human annotations (Appendix E).

Results. Table 5 reports scores. GPT-5 and Sonnet 4.5 are tied overall. We choose GPT-5 as the default generator for consistency. Notably, the strongest open-weight model (gpt-oss-120b) performs comparably to the closed-weight models.

5.5 Benchmarking Against Baseline Methods

DiNO is, to our knowledge, the first method to extract disinformation narratives from news articles using verified false claims. To contextualize its performance, we compare against two competitive narrative extraction methods that operate directly on news text: CaNarEx (Anantharama et al., 2022) and Relatio (Ash et al., 2024). We run both methods on the Sputnik-Ukraine outlet and evaluate their outputs using the same protocol (Section 5.4).

Method	Human			Automated		
	mTA_H	mSA_H	Rd_H	mTA_A	mSA_A	Rd_A
Closed-weight LLM						
GPT-5	**0.75**	**0.85**	2.12	**0.78**	0.85	1.80
Sonnet 4.5	**0.75**	0.83	2.04	0.76	**0.87**	1.72
Gemini 2.5 Pro	0.74	0.80	1.97	0.75	0.85	1.68
GPT-5-mini	0.73	0.84	2.08	0.75	0.84	1.75
Open-weight LLM						
gpt-oss-120b	0.74	0.84	2.01	0.74	0.84	1.70
Qwen3-14B	0.72	0.77	2.29	0.70	0.80	1.88
Llama-3.3-70B	0.68	0.79	2.15	0.73	0.85	1.92

Table 5: Alignment scores and redundancy between extracted narratives and ground-truth narratives on the Sputnik-Ukraine dataset, across LLM generators.

As shown in Table 6, DiNO strongly outperforms both methods. We performed ablation studies and our results suggest that the gains mainly come from two main reasons. First, DiNO extracts narratives only from segments marked as false, rather than noisy full articles. Second, DiNO synthesizes narratives using LLMs, while CaNarEx and Relatio rely on their native (non-LLM) output representations. Appendix C.10 reports an ablation study supporting these findings.

Method	Human		Automated	
	mTA_H	mSA_H	mTA_A	mSA_A
DiNO	**0.75**	**0.85**	**0.78**	**0.85**
CaNarEx	0.52 (-31%)	0.60 (-29%)	0.54 (-31%)	0.63 (-26%)
Relatio	0.53 (-29%)	0.65 (-24%)	0.53 (-32%)	0.61 (-28%)

Table 6: Alignment scores between extracted and ground-truth narratives on the Sputnik-Ukraine outlet.

5.6 Evaluation Across All Datasets

After finding best parameters for each component (Sections 5.1–5.4), we run DiNO end-to-end on every dataset and compare the resulting narrative sets to EUDisinfoTest. We use the same set-based matching as in Section 5.4: each predicted narrative is paired with its most semantically similar reference narrative, and we report topical alignment (mTA), stance alignment (mSA), and redundancy (Rd) under both human and automated evaluation.

Interpreting outputs. On disinformation-prone outlets, we expect extracted narratives to align with expert narratives in both topic and stance and treat these as true positives. On credible outlets, outputs that are topically aligned but stance-opposed (high

mTA, low mSA) reflect topic-relevant reporting without endorsement, which is consistent with credible reporting that mentions false claims for context or refutation; we treat these as false positives.

Results. Table 7 summarizes human and automated scores. On disinformation-prone outlets, DiNO extracts narratives that are both topically aligned and stance-aligned with expert ground-truth narratives. On credible outlets, DiNO produces only two topically aligned but stance-opposed narratives per domain.

Beyond alignment, DiNO yields compact narrative sets with low redundancy (human: 1.71, automated: 1.50). This is consistent with our clustering objective, which maximizes the silhouette score to reduce both over-splitting and over-merging.

Moreover, our automated judge is a reliable proxy for expert scoring: automatic ratings show strong agreement with human judgments ($r_{WG} = 0.83$ topical, $r_{WG} = 0.73$ stance)[5], validating automated evaluation for large-scale experiments.

Finally, in Appendix D, we further evaluate DiNO's robustness by comparing against alternative ground-truth narrative sets and computing a set-level Chamfer Distance metric (Barrow et al., 1977) that captures match quality and coverage.

Dataset	NC	Human			Automated		
		mTA_H	mSA_H	Rd_H	mTA_A	mSA_A	Rd_A
Ukraine							
Sputnik	72	0.75	0.85	2.12	0.78	0.85	1.80
Pravda	61	0.75	0.90	2.03	0.69	0.83	1.96
Guardian	2	0.90	0.00	1.00	0.70	0.00	1.00
COVID-19							
NatNews	74	0.79	0.88	1.90	0.86	0.92	1.54
Sputnik	21	0.82	0.74	1.31	0.79	0.88	1.23
Guardian	2	1.00	0.00	1.00	0.70	0.00	1.00
Migration							
Breitbart	64	0.72	0.86	3.16	0.76	0.85	2.37
Sputnik	13	0.86	0.85	1.86	0.68	0.83	1.62
Guardian	2	0.60	0.00	1.00	0.70	0.40	1.00

Table 7: Human and automated topical and stance alignment scores between DiNO-extracted and ground-truth disinformation narratives, together with redundancy. **NC**: number of narratives extracted per dataset.

6 Results and Discussion

To address **RQ2**, this section analyzes the five most prevalent narratives extracted by DiNO in each dataset, as listed in Figure 3. The full set of detected narratives is available in our repository[6].

Pravda-Ukraine
- Ukraine's army is collapsing while Russia humanely targets only military sites.
- NATO secretly controls Ukrainian long-range missile strikes.
- Trump's election will end the Ukraine war by forcing Kyiv to capitulate.
- Ukraine, directed by US–UK intelligence, organized the Crocus City Hall attack.
- NATO, led by the US, is fabricating a Russian threat.

Sputnik-Ukraine
- Russia's invasion of Ukraine is a defensive, humanitarian mission to stop genocide.
- Anti-Russia sanctions backfire, crippling Western economies.
- West hijacked the Ukraine grain deal to enrich Europe while starving the Global South.
- The United States secretly bombed the Nord Stream pipelines and is covering it up.
- The US and EU censor truthful Russian media and alternative voices to control.

NaturalNews-COVID
- COVID-19 mRNA vaccines make toxic spike proteins that spread through the body.
- COVID-19 was a fake, planned crisis to impose lethal depopulation vaccines.
- COVID-19 vaccines are a deliberate bioweapon causing millions of hidden deaths.
- Anthony Fauci secretly funded Wuhan bioweapon gain-of-function research.
- Cheap drugs ivermectin and hydroxychloroquine safely cure COVID.

Sputnik-COVID
- COVID-19 restrictions and vaccines are tyrannical, unnecessary.
- UK COVID-19 response was driven by lies, fearmongering and political manipulation.
- China's Wuhan lab engineered or leaked COVID-19 and then hid evidence of its origin.
- US COVID-19 response and vaccines are corrupt tools of elites, not real public health.
- COVID-19 death and case statistics are unreliable, manipulated, and exaggerate.

Breitbart-Migration
- Democrats want open borders and mass amnesty to import immigrant voters.
- Muslim and non-European migrants cause a Europe-wide crime wave, creating no-go.
- Globalist elites deliberately use mass Syrian and Muslim migration to control Europe.
- Undocumented immigrants are mostly dangerous criminals and must be mass-deported.
- Biden is deliberately flooding America with migrants to replace workers, steal jobs.

Sputnik-Migration
- Syrian and other refugees in Europe are mostly fake asylum seekers, criminals.
- Turkey deliberately weaponizes refugees to blackmail the EU for money, visas.
- Biden and Democrats keep the US border deliberately open to import criminals.
- A US-Mexico border wall is a simple, effective solution that will stop illegal migration.
- US-NATO deliberately destroyed Libya and Syria to flood Europe with migrants.

Guardian-Ukraine
- Russia's invasion of Ukraine and the resulting sanctions caused a global energy shock.
- Russia frames the Ukraine war as defensive against NATO expansion.

Guardian-COVID
- COVID-19 accelerates shift to online shopping, causing high street store closures.
- COVID-19 has changed the world through public health measures and disinformation.

Guardian-Migration
- Refugees strengthen host countries through work and community ties.
- European governments adjust migration rules to meet humanitarian needs.

Figure 3: Top five (where available) most frequent narratives extracted by DiNO per dataset (truncated).

Further analyses are provided in Appendix G and include: (i) distributions and temporal trends of the narratives linking peaks to major socio-political events, and (ii) cross-source comparisons quantifying the similarity of narrative sets across outlets.

Most prevalent narratives by domain. In the **Ukraine War** domain, DiNO extracts 72 narratives from Sputnik, 61 from Pravda, and 2 from The Guardian. Pravda's most frequent narratives emphasize battlefield dynamics and Ukrainian weakness (e.g., "army collapsing," "NATO controls strikes"), while Sputnik more often uses Western-facing frames (e.g., "sanctions backfire," "Nord Stream conspiracy," "media censorship"). The Guardian remains topically related but stance-opposed (e.g., reporting on Russia's framing).

For **COVID-19**, NaturalNews yields 74 nar-

[5]We use r_{WG} as it is recommended for assessing agreement on Likert scales (O'Neill, 2017).

[6]https://github.com/wsosnowski/DiNO/tree/main/predicted_narratives_and_annotation/README.md

ratives dominated by vaccine-harm and elite-conspiracy claims (plus "cheap cures"), while Sputnik yields 21 narratives focused on restrictions as illegitimate, statistics as manipulated, and lab-leak framings. The Guardian again yields only two stance-opposed, contextual narratives.

In the **Migration** domain, Breitbart yields 64 narratives framing migration as a crime/economic threat and policy as deliberately permissive (e.g., "importing voters," "globalist coordination"), while Sputnik yields 13 narratives emphasizing geopolitical causes and hard-border solutions (e.g., refugees "weaponized," NATO wars as drivers). The Guardian yields two pro-refugee narratives.

DiNO behavior summary. Across outlets, DiNO follows a consistent pattern: it produces multiple narratives on disinformation-prone sources that strongly align with ground truth disinformation narratives in both topic and stance, whereas on credible outlets it produces only a few narratives per domain. Alongside consistently low redundancy, this suggests that DiNO remains stable and reliable across outlet types.

7 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.

8 Conclusions

We present DiNO, a novel method for mining disinformation narratives from news articles. DiNO matches segments containing false information by cross-referencing them with verified claims, then clusters these segments to extract narratives.

We evaluated DiNO using articles on three topics: the Russo-Ukraine War, COVID-19 and Migration. The analysis compared credible sources (The Guardian) with disinformation sources (Sputnik, Pravda, NaturalNews, Breitbart). Narratives from disinformation sources aligned closely with ground-truth disinformation narratives, while those from credible sources yielded far fewer narratives and showed little stance alignment.

Moreover, DiNO's narratives showed much higher alignment with ground-truth disinformation narratives than those produced by established methods, CaNarEx and Relatio. DiNO achieved a topical alignment score of 0.75 (vs. 0.52 and 0.53) and a stance alignment score of 0.85 (vs. 0.60 and 0.65), placing it among the best methods for extracting disinformation narratives from news articles.

9 Limitation

DiNO relies on LLMs across multiple stages of its processing pipeline, which means it also inherits some of the inherent semantic limitations of these models. For example, LLMs can struggle with nuanced contexts and may not always distinguish between similar but distinct narratives. To mitigate these challenges, we conducted a thorough evaluation of each stage in DiNO, comparing various LLMs and non-LLM approaches (see Section 5). Another limitation of DiNO is its reliance on a clustering algorithm, which makes it less suitable for real-time applications or continuously evolving data streams. However, since disinformation narratives emerge from multiple instances rather than isolated occurrences, analyzing new data within the broader context of the entire dataset remains essential. Moreover, disinformation verification media can be inherently biased. Even the most reliable sources may carry biases, which DiNO could inadvertently inherit.

10 Ethical Considerations

This section considers the ethical issues and broader implications of using language models to mine disinformation narratives. Although our university's ethics board deemed the study exempt from further review, we still assess potential risks.

Datasets Our research utilizes multiple news outlets: EUvsDisinfo, The Guardian, Pravda, Sputnik, NaturalNews, Breitbart, and Google Fact Check Tool API. Collection and analysis were conducted via lawful access and we relied on the EU text-and-data-mining exceptions in Articles 3 and 4 of Directive (EU) 2019/790, subject to applicable conditions (e.g., rights-holder reservations/opt-outs). To support reproducibility while respecting copyright and licensing constraints, we release source links (URLs) and robust crawling scripts. We did not perform additional screening for personally identifiable information or offensive content beyond what is provided by the original sources.

Intended Use of Our Research Results. Our goal is to support institutions countering disinfor-

mation (e.g., organizations following the International Fact-Checking Network's code of principles) by enabling monitoring and analysis of disinformation narratives. We acknowledge a misuse risk: DiNO outputs narrative formulations, which can be reused as shareable "slogans". We therefore recommend use only in controlled settings with human oversight, and require that any reported narratives be clearly labeled as false including link to the relevant fact-check or debunk.

Demographic or Identity Characteristics. Our study does not address demographic or identity characteristics.

Overview of Computational Resources and Costs in Our Research. DiNO experiments leveraged both commercial LLMs accessed via dedicated APIs and embedding models running on a server with four NVIDIA L40 GPUs. The server handled tasks like embedding generation or clustering. Under our best-performing configuration, the end-to-end computational cost of a DiNO run was approximately $100 (USD).

Expert Involvement. Annotations were completed by university-employed experts, compensated above the average for similar roles.

References

AAP FactCheck. 2024. Trump's debunked dominion claim given another run.

Hervé Abdi and Lynne J. Williams. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

AFP Fact Check. 2021. COVID-19 vaccines do not contain magnetic microchips.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th international workshop on semantic evaluation*.

Nika Aleksejeva. 2023. *Narrative warfare: How the Kremlin and Russian news outlets justified a war of aggression against Ukraine*. Atlantic Council.

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

Anthropic. 2025. Introducing Claude Sonnet 4.5.

Elliott Ash, Germain Gauthier, and Philine Widmer. 2024. Relatio: Text semantics capture political and economic narratives. *Political Analysis*, 32(1).

Michael Balsamo. 2020. Disputing trump, barr says no widespread election fraud.

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 of the 5th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 659–663.

Y Benkler, R Faris, and H Roberts. 2018. Immigration and islamophobia: Breitbart and the trump party. *Network propaganda: Manipulation, disinformation, and radicalization in American politics. Oxford University Press. https://doi.org/10.1093/oso/9780190923624.003*, 4.

Mack Blackburn, Ning Yu, John Berrie, Brian Gordon, David Longfellow, William Tirrell, and Mark Williams. 2020. Corpus development for studying online disinformation campaign: a narrative+ stance approach. In *Proceedings for the First International Workshop on Social Threats in Online Conversations: Understanding and Management*, pages 41–47.

Breitbart. 2025. Breitbart immigration.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

CDC. 2024. Myths and facts about COVID-19 vaccines.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense x retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Eun Cheol Choi and Emilio Ferrara. 2024. Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1441–1449.

EDMO. European digital media observatory.

EDMO. 2024. Edmo fact-checking network statement about methodology.

Mark Elsner, Grace Atkinson, Saadia Zahidi, and World Economic Forum. 2025. The global risks report 2025: 20th edition. Insight report, World Economic Forum, Cologny/Geneva, Switzerland. Published 15 January 2025.

Mercedes Esteban-Bravo, Jose M Vidal-Sanz, et al. 2024. Predicting the virality of fake news at the early stage of dissemination. *Expert Systems with Applications*, 248:123390.

EUvsDisinfo. 2017. The pro-kremlin narrative about migrants.

EUvsDisinfo. 2019. 5 common pro-kremlin disinformation narratives.

EUvsDisinfo. 2021. Ukrainians imbued with destructive nazi ideology.

EUvsDisinfo. 2022. Key narratives in pro-kremlin disinformation part 4: "The Imminent Collapse".

EUvsDisinfo. 2024. Disinfo database.

Jesús M Fraile-Hernández, Anselmo Peñas, and Pablo Moral. 2024. Automatic identification of narratives: Evaluation framework, annotation methodology, and dataset creation. *IEEE Access*, 13:11734–11753.

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

Pranav Goel, Jon Green, David Lazer, and Philip Resnik. 2023. Mainstream news articles co-shared with fake news buttress misinformation narratives.

Google. 2025a. Gemini 2.5 Pro model card. Model card.

Google. 2025b. Google fact check explorer.

Amanda Gray Meral. 2020. Learning the lessons from the EU–turkey deal: Europe's renewed test.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Nuno Guimarães, Maria da Purificação Silvano, Ricardo Campos, AM Jorge, and Ana Filipa Pacheco. 2025. Narratex dataset: Explaining the dominant narratives in news texts. *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Michael Hameleers. 2023. Disinformation as a context-bound phenomenon: toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination. *Communication Theory*.

John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Maria Hellman. 2024. *Security, Disinformation and Harmful Narratives: RT and Sputnik News Coverage about Sweden*. Palgrave Macmillan, Cham, Switzerland.

Catalina Jaramillo. 2023. WHO 'pandemic treaty' draft reaffirms nations' sovereignty to dictate health policy.

Myunghoon Kang and Greg Chih-Hsin Sheen. 2025. The making of the boy who cried wolf: fake news and media skepticism. *Political Science Research and Methods*, 13(2):465–474.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

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

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Rabeeh Karimi Mahabadi, Florian Mai, and James Henderson. 2019. Learning entailment-based sentence embeddings from natural language inference.

Lisa M. Martinez. Unraveling the giant ball of lies, myths, and disinformation about immigration. Center for Immigration Policy & Research (CIPR), University of Denver.

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*.

Media Bias/Fact Check. 2025. Natural news – bias and credibility.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. SFR-Embedding-2: Advanced text embedding with multi-stage training. Hugging Face model card.

Meta. 2024. Llama-3.3-70B-Instruct model card. Hugging Face.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411.

Deborah Miori and Constantin Petrov. 2025. Narratives from gpt-derived networks of news and a link to financial markets dislocations. *International Journal of Data Science and Analytics*, 20(2):1105–1129.

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

Arkadiusz Modzelewski, Witold Sosnowski, Eleni Papadopulos, Elisa Sartori, Tiziano Labruna, Giovanni Da San Martino, and Adam Wierzbicki. 2026. Malicious intent dataset and inoculating llms for enhanced disinformation detection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3125–3148.

Pablo Moral, Jesús M 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*, 73:347–358.

NaturalNews. 2025. Naturalnews: Covid-19 tag page.

Nic Newman, A Ross Arguedas, Craig T Robertson, Rasmus Kleis Nielsen, and Richard Fletcher. 2025. *Digital news report 2025*. Reuters Institute for the study of Journalism.

News Pravda. 2025. News pravda: Ukraine.

Nikolaos Nikolaidis, Nicolas Stefanovitch, Purificação Silvano, Dimitar Iliyanov Dimitrov, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ion Androutsopoulos, Preslav Nakov, Giovanni Da San Martino, et al. 2025. Polynarrative: A multilingual, multilabel, multi-domain dataset for narrative extraction from news articles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31323–31345.

Thomas A. O'Neill. 2017. An overview of interrater agreement on likert scales for researchers and practitioners. *Frontiers in Psychology*, 8:777.

OpenAI. GPT-5 mini model.

OpenAI. 2025a. GPT-4.1.

OpenAI. 2025b. GPT-5 model.

OpenAI. 2025c. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alipio Mario Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, et al. 2025. Semeval 2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2610–2643.

PolitiFact. 2022. Russian has not been banned in ukraine, despite repeated claims.

PolitiFact. 2024. False claim resurfaces about WHO pandemic treaty and US sovereignty.

Maria Popp, Miriam Stegemann, Maria-Inti Metzendorf, Susan Gould, Peter Kranke, Patrick Meybohm, Nicole Skoetz, and Stephanie Weibel. 2021. Ivermectin for preventing and treating covid-19. *Cochrane Database of Systematic Reviews*, (7).

Justin Reedy, Benjamin Gonzalez O'Brien, and Elizabeth H Hurst. 2023. Pandemic politics: Immigration, framing, and covid-19. *Journal of Race, Ethnicity, and Politics*, 8(2):246–266.

Reuters. 2024. What is ISIS-K and why would it attack a Moscow concert hall?

Reuters Fact Check. 2024a. Posts wrongly claim 7,000 migrants arrived in lampedusa in 36 hours.

Reuters Fact Check. 2024b. Video shows boats arriving in spain in 2023, not britain in 2024.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Michel Rose and John Irish. 2024. France's Macron does not rule out Europeans sending troops to Ukraine.

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.

Tim Sainburg, Leland McInnes, and Timothy Q. Gentner. 2021. Parametric UMAP embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907.

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).

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

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3607–3618.

Shaden Shaar, Nikola Georgiev, Firoj Alam, Giovanni Da San Martino, Aisha Mohamed, and Preslav Nakov. 2022. Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2069–2080.

Snopes. 2022. Zelenskyy swastika jersey pic is fake.

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.

Witold Sosnowski, Arkadiusz Modzelewski, Kinga Skorupska, and Adam Wierzbicki. 2025. Dinam: Disinformation narrative mining with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30212–30239.

Saranac Hale Spencer. 2020. Sharpie ballots count in arizona.

Sputnik. 2025a. Sputnikglobe: Covid-19 topic page.

Sputnik. 2025b. Sputnikglobe search results for migration.

Sputnik. 2025c. Sputnikglobe search results for ukraine.

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.

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

Edson C Tandoc Jr. 2019. The facts of fake news: A research review. *Sociology Compass*, 13(9):e12724.

Kyilah Terry. 2021. The EU–turkey deal, five years on: A frayed and controversial but enduring blueprint.

The Guardian. 2025a. The guardian: Coronavirus outbreak.

The Guardian. 2025b. The guardian: Migration.

The Guardian. 2025c. The guardian: Ukraine.

United Nations. 2022. Black sea grain initiative joint coordination centre.

U.S. Customs and Border Protection. Southwest land border encounters.

VIGINUM. 2024. PORTAL KOMBAT: A structured and coordinated pro-Russian propaganda network. Technical report, Secrétariat général de la défense et de la sécurité nationale.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv e-prints*, pages arXiv–2212.

World Health Organization. 2021. Who-convened global study of origins of SARS-CoV-2: China part.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Liwen Zheng, Chaozhuo Li, Zheng Liu, Feiran Huang, Haoran Jia, Zaisheng Ye, and Xi Zhang. 2025. Fact in fragments: Deconstructing complex claims via llm-based atomic fact extraction and verification.

A Additional Details on Datasets

To construct our datasets from news outlets, we first scraped article URLs using Selenium, automating browsing from the "Website URL" entries listed in Table 1. With these URLs, we then extracted clean article text using the trafilatura tool.

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

Moreover, the original EUvsDisinfo dataset contained 18,341 entries. To ensure unique, diverse samples, we removed duplicates and near-duplicates, defining near-duplicates as pairs

with over 95% cosine similarity using the SFR-Embedding-2_R model. For each group above this threshold, one entry was retained, yielding a cleaned dataset of 15,452 entries.

The GFCE dataset was built by querying COVID-19 and migration fact-checks from the Google Fact Check Tools API[7] and retaining only items whose verdict indicated falsity (e.g., *False*, *Manipulated*).

B Prompts

B.1 Prompt Template

We integrated LLMs into DiNO using a specialized prompt template (Figure 4). While our task diverges from traditional disinformation detection, we drew insights from the state-of-the-art prompts designed for that task (Lucas et al., 2023).

Our prompt design consists of four components:

1. **Impersonation**: Assigns a specific role to the LLM to override default alignment behavior and align the model with task-specific goals.
2. **Guidelines**: Provides detailed instructions to constrain and direct the model's output.
3. **Context**: Supplies the relevant article, claim, or supporting content required to perform the task.
4. **Output**: Specifies the required format for the model's response.

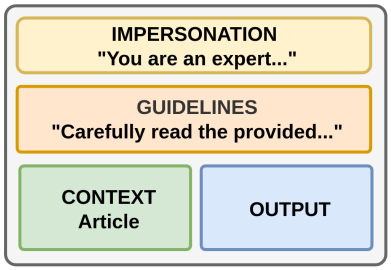

Figure 4: An overview of prompt template, comprising: (1) Impersonation, to define the model's role, (2) Guidelines, to specify instructions, (3) Context, to embed task-relevant information, and (4) Output.

B.2 Entailment Verification

To assess LLMs' ability to determine whether an article supports a given claim, we evaluated six prompts: three custom prompts (Figures 11–13) and three adapted from standard entailment benchmarks (Figures 14–16), following (Gallipoli and Cagliero, 2025). All prompts follow the same underlying template (Appendix B.1).

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

B.3 Extracting Disinformation Narratives

To derive narratives from related false-segment clusters, we used the prompt shown in Figure 17, constructed using the shared prompt template (Appendix B.1).

B.4 Narrative Evaluation Prompts

We designed three prompts for automatic narrative evaluation: (1) selecting the most semantically similar ground-truth narrative for each predicted narrative (Figure 18), (2) rating topical alignment on a 0–5 Likert scale (Figure 19), and (3) rating stance alignment, also on a 0–5 scale (Figure 20). All prompts follow the same template (Appendix B.1).

C Additional Details on Evaluation

This appendix complements the main evaluation of DiNO with additional experimental details, baselines, and analyses. Specifically, we document hyperparameter searches for false-segment matching (Sec. 5.1) and clustering (Sec. C.3), report additional non-LLM baselines for entailment verification (Sec. C.2) and narrative extraction (Sec. C.4), summarize the DiNO method and step-wise settings (Sec. C.5), wall-clock runtimes (Section C.6), estimate costs (Section C.7), and present robustness to incomplete false-claim databases (Sec. C.8), a cluster-level human evaluation (Sec. C.9), and a comparison to CaNarEx and Relatio (Sec. C.10).

C.1 Matching False Segments

Hyperparameters Search We optimized our false-segment matching pipeline, using a development set disjoint from the evaluation set (Section 5.1). We compared three embedding models (SFR-Embedding-2, E5-large, and Jina) and tuned: segment length l, overlap o, and retrieval depth t. Table 8 lists the search ranges, and Table 9 reports the best settings. Moreover, for retrieving candidate segment-claim matches we used FAISS IndexIVFFlat ANN search.

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

Table 8: Hyperparameter ranges for Step 1, used during tuning on the Sputnik-Ukraine dataset.

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

Table 9: Best hyperparameter values for each embedding model for matching false segments in the Sputnik-Ukraine dataset.

C.2 Entailment Verification

Here we provide additional non-LLM baselines for the entailment verification step. We evaluate fine-tuned encoder models using the same testing set and evaluation protocol as the main experiments (see Sec. 5.2).

Non-LLM Baselines. For comparison, we fine-tuned DeBERTa-large (He et al., 2020) and RoBERTa-large (Liu et al., 2019). Each input pairs the false claim with the article text, concatenated with a `[SEP]` token and labeled as *entailed* or *contradicted*. Both models were trained for five epochs with binary cross-entropy loss and AdamW (batch size 8, learning rate 2e-5). We report 5-fold cross-validation performance.

Results. The fine-tuned encoders underperform the LLM-based alternatives (see Section 5.2. DeBERTa-large achieves an F_1 of 0.63 ($P_{support}$=0.55), and RoBERTa-large achieves an F_1 of 0.61 ($P_{support}$=0.54).

C.3 Clustering False Segments

Hyperparameters Search Here, we outline the hyperparameter search for the UMAP, PCA, HDBSCAN, and K-Means. The specific ranges of hyperparameters explored for each algorithm are provided in Table 10. The optimal hyperparameters for each algorithm, determined through experimentation, are listed in Table 11.

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 10: Hyperparameter ranges for clustering methods applied to false segments extracted from the Sputnik-Ukraine dataset.

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

Table 11: Optimal hyperparameters for clustering methods applied to false segments from the Sputnik-Ukraine dataset.

C.4 Extracting Disinformation Narratives

Here we provide additional non-LLM baselines for disinformation narrative extraction. We evaluate extractive summarization methods using the same cluster inputs and automatic-judge evaluation protocol as the main experiments (see Section 5.4).

Non-LLM Baselines. We evaluated two non-LLM methods: BERTsum (Liu, 2019) and TextRank (Mihalcea and Tarau, 2004).

Results. Both extractive baselines underperform LLM-based narrative generators(see Section 5.4), yielding substantially lower topical and stance alignment under the automatic-judge protocol. BERTsum achieves $mTA_A = 0.63$ and $mSA_A = 0.62$, while TextRank achieves $mTA_A = 0.62$ and $mSA_A = 0.60$.

C.5 DiNO Overview

Tables 12 and 13 summarize the DiNO method, detailing step-wise input/output quantities for each outlet and domain, along with the optimal models and parameters used at each step. Table 28 provides the details regarding used models.

C.6 Execution Time

We measured wall-clock time (in minutes) for DiNO's four steps on our reference hardware (2× Intel® Xeon® Gold 5418Y, 128 GB RAM, 4× NVIDIA L40). Table 14 reports per-step runtimes across all the datasets (see Table 1).

Outlet	Step 1 Articles	Step 2 Segments	Step 3 Clusters	Step 4 Narr.
Ukraine War				
Sputnik	20 001	11 225	72	72
Pravda	22 695	9 438	61	61
Guardian	11 841	341	2	2
COVID-19				
Sputnik	13 107	2 983	21	21
NatNews	16 789	11 218	74	74
Guardian	10 573	327	2	2
Migration				
Sputnik	9 028	2 242	13	13
Breitbart	16 203	8 821	64	64
Guardian	11 671	329	2	2

Table 12: Step-wise statistics of the DiNO method for each news outlet, across both Ukraine War, COVID-19 and Migration domains. Step 1: total articles processed, Step 2: number of identified segments in processed articles, Step 3: number of clusters of semantically similar false segments, Step 4: number of extracted disinformation narratives.

DiNO Component	Optimal Setting / Model
Matching False Segments	
Similarity Measure	Cosine (Jina)
Segmentation	DiNO-sgm
Hyperparameters	$l = 4,\ o = 1,\ t = 3$
Entailment Verification	
Model / Prompt	GPT-5-mini / (Fig. 11)
Clustering False Segments	
Embedding Model	SFR-Embedding-2
Dimensionality Reduction	UMAP
Clustering Algorithm	HDBSCAN
Extracting Narratives	
Model / Prompt	GPT-5 / (Fig. 17)

Table 13: Optimal models and settings for each step of the DiNO method, covering segment matching, entailment verification, clustering, and narrative extraction.

Dataset	Step 1	Step 2	Step 3	Step 4
Ukraine War				
Sputnik	207 min	13 min	14 min	<1 min
Pravda	182 min	14 min	10 min	<1 min
Guardian	42 min	10 min	3 min	<1 min
COVID-19				
Sputnik	104 min	10 min	7 min	<1 min
NaturalNews	133 min	12 min	8 min	<1 min
Guardian	34 min	10 min	3 min	<1 min
Migration				
Sputnik	83 min	10 min	7 min	<1 min
Breitbart	129 min	12 min	10 min	<1 min
Guardian	39 min	10 min	3 min	<1 min

Table 14: Wall-clock runtimes (in minutes) of the four DiNO steps: segment matching, entailment verification, clustering, and narrative extraction.

End-to-End Comparison. Table 15 shows the end-to-end runtimes (in minutes) for DiNO, CaNarEx, and Relatio.

Dataset	DiNO	CaNarEx	Relatio
Ukraine War			
Sputnik	234 min	810 min	125 min
Pravda	206 min	762 min	110 min
Guardian	55 min	194 min	41 min
COVID-19			
Sputnik	121 min	531 min	82 min
NaturalNews	153 min	569 min	82 min
Guardian	47 min	173 min	37 min
Migration			
Sputnik	100 min	378 min	64 min
Breitbart	151 min	569 min	97 min
Guardian	52 min	195 min	33 min

Table 15: Total end-to-end processing times (in minutes) comparing DiNO against CaNarEx and Relatio.

C.7 Execution Cost

We report only costs from external LLM APIs, excluding infrastructure costs. Cloud-based estimates follow from applying current per-token rates to the token counts reported below.

Token-count Assumptions. We follow OpenAI's convention of 1 token $\approx$ 0.75 words, and convert word counts to input/output token estimates for each LLM-based step.

Matching False Segments. This step does not rely on the LLMs, hence no API costs.

Entailment Verification. This step, accounts for the majority of API usage. Table 16 reports aggregate token estimates under retrieval depth t, i.e., the number of highest-scoring segment-claim pairs retained per article. We assume a brief 20-token classification verdict per document.

Dataset	Input Tokens ($t = 1/t = 3$)	Output Tokens ($t = 1/t = 3$)
Ukraine War		
Sputnik	15.6 M / 39.1 M	0.4 M / 1.0 M
Pravda	6.7 M / 16.7 M	0.5 M / 1.1 M
Guardian	14.0 M / 35.0 M	0.2 M / 0.6 M
COVID-19		
Sputnik	7.7 M / 19.2 M	0.3 M / 0.7 M
NaturalNews	18.7 M / 46.8 M	0.3 M / 0.9 M
Guardian	11.6 M / 28.9 M	0.2 M / 0.5 M
Migration		
Sputnik	4.8 M / 12.1 M	0.1 M / 0.4 M
Breitbart	9.8 M / 24.5 M	0.3 M / 0.7 M
Guardian	11.5 M / 28.7 M	0.2 M / 0.5 M

Table 16: Token usage estimates for Entailment Verification (millions). Each entry reports $t = 1/t = 3$, where t is retrieval depth.

To compute cost, multiply the input and output token counts by your model's per-token price.

Clustering False Segments. No LLM calls are made during clustering, hence no API costs.

Extracting Disinformation Narratives. We use LLMs to generate narratives for each cluster. Table 17 shows the input and output token estimates.

Dataset	Input Tokens	Output Tokens
Ukraine War		
Pravda	0.55 M	<0.01 M
Sputnik	0.99 M	<0.01 M
Guardian	0.05 M	<0.01 M
COVID-19		
Sputnik	0.51 M	<0.01 M
NaturalNews	0.74 M	<0.01 M
Guardian	0.04 M	<0.01 M
Migration		
Sputnik	0.35 M	<0.01 M
Breitbart	0.61 M	<0.01 M
Guardian	0.05 M	<0.01 M

Table 17: Token usage for Extracting Disinformation Narratives across datasets (values in millions of tokens).

C.8 Robustness to Incomplete False-Claim Databases and Emerging Narratives

To assess DiNO under a realistic deployment constraint and to approximate an emerging-narratives setting, we assume that at any given time only a subset of false claims has been collected and verified. Since exhaustively annotating all claims expressed in an outlet's full article stream is beyond the scope of this work, we construct chronologically truncated versions of the verified-claim database and observe DiNO's behavior when only earlier-verified claims are available.

Evaluation. Starting from the EUvsDisinfo collection of 15,452 verified false claims, we sort entries by claim date and create three reduced databases by retaining only the earliest 75%, 50%, and 25% of claims (11,589 / 7,726 / 3,863), thereby removing the most recent portion of the repository. For each reduced database, we rerun the DiNO method (using default settings) on the Sputnik Ukraine corpus and evaluate against EUDisinfoTest dataset.

Results. Table 18 shows moderate alignment drops under truncation: mTA_A decreases from 0.78 to 0.70 and mSA_A from 0.85 to 0.82 when reducing the database from 15,452 to 3,863 claims. Coverage falls more sharply: matched segments drop from 11,225 to 3,475 and narratives from 72 to 41. Narratives shrink more slowly than segments. Removing 25% of claims cuts segments by 23.3% but narratives by only 2.8%.

Findings. (1) DiNO remains effective when only earlier-verified claims are available, with stable topic and stance alignment. (2) Truncation mainly reduces coverage, consistent with missing newer or less-documented claims. (3) DiNO still depends on previously verified claims and will miss narratives driven entirely by newly emerging falsehoods.

C.9 Cluster-Level Human Evaluation

We run a human study to assess cluster coherence on the Sputnik-Ukraine dataset. For each DiNO cluster C, we randomly sample three segments s_1, s_2, s_3. Two annotators then compare each of the three segment pairs ((s_1, s_2), (s_1, s_3), and (s_2, s_3)) and label each pair as *Yes*/*No*/*Unsure* to the question: *Do these two segments express the same narrative?*

We map labels to $y_{ij} \in \{1, 0, 0.5\}$ (Yes=1, No=0, Unsure=0.5) and compute cluster purity

$$\text{Purity} = \tfrac{1}{3}(y_{12} + y_{13} + y_{23}),$$

We also report a pass rate: a cluster is coherent if at least two pairs are Yes,

$$\text{PassRate} = \tfrac{1}{N} \sum_{k=1}^{N} \mathbb{I}\left[\#\{Yes\} \geq 2\right].$$

Results. Inter-annotator agreement was κ=0.63. Table 19 reports mean Purity and PassRate. The human judgments indicate that clusters from UMAP+HDBSCAN are typically coherent at the level of disinformation narratives: within a cluster, sampled segments usually express the same underlying narrative rather than merely sharing keywords. In contrast, the K-Means+PCA more often mixes distinct narratives within the same cluster, leading to less consistent coherence. Overall, UMAP+HDBSCAN yields clusters that are easy to interpret as single narratives.

C.10 Comparison of DiNO with CaNarEx and Relatio

We present a detailed comparison of DiNO with the baseline systems CaNarEx and Relatio. We argue that DiNO's superior performance arises primarily from two key design choices: **(H1)** focusing exclusively on false segments rather than entire articles, thereby minimizing noise and yielding more coherent clusters, and **(H2)** employing a large language model to synthesize cluster-level narratives, which enables a more faithful representation of both theme and stance.

# False Claims in DB	mTA_A	mSA_A	# Narratives (Δ%)	# Matched Segments (Δ%)
15,452	0.78	0.85	72	11,225
11,589 (-25%)	0.74	0.87	70 (-2.8%)	8,610 (-23.3%)
7,726 (-50%)	0.72	0.84	55 (-23.6%)	5,950 (-47.0%)
3,863 (-75%)	0.70	0.82	41 (-43.1%)	3,475 (-69.0%)

Table 18: Effect of chronologically truncating the verified-claim database (EUvsDisinfo), i.e., removing the most recently dated claims, on DiNO's performance when mining disinformation narratives from the Sputnik Ukraine dataset.

Method	Purity	PassRate
UMAP + HDBSCAN	0.73	0.85
K-Means + PCA	0.59	0.70

Table 19: Cluster coherence results on the Sputnik-Ukraine dataset, measured via pairwise purity and pass rate (higher is better).

We test these hypotheses with two ablations. For **H1**, CaNarEx and Relatio were provided with the false segments identified by DiNO (DiNO-sgm) instead of full articles. For **H2**, we adapted both baselines by adding an LLM-based narrative synthesis step: instead of relying on their native output representations, GPT-5 was prompted to generate a coherent narrative for each cluster (or group of output units) using our dedicated narrative synthesis prompt (see Section 3.4).

	mTA_A	mSA_A
Original Methods		
DiNO	0.78	0.85
CaNarEx	0.54	0.63
Relatio	0.53	0.61
H1: Segment-only input		
CaNarEx	0.65 (+20.4%)	0.76 (+20.6%)
Relatio	0.63 (+18.9%)	0.72 (+18.0%)
H2: LLM-based narrative synthesis		
CaNarEx	0.65 (+20.4%)	0.75 (+19.0%)
Relatio	0.66 (+24.5%)	0.76 (+24.6%)

Table 20: Ablation results testing **H1** (segment-only inputs) and **H2** (LLM-based narrative synthesis) on the *Sputnik-Ukraine* dataset. Percentage gains in H1 and H2 are relative to each method's original scores.

Results on Sputnik-Ukraine. Table 20 supports both hypotheses. Under **H1**, providing segment-only inputs improves the baselines: CaNarEx rises to 0.65 (+20.4%) in mTA_A and 0.76 (+20.6%) in mSA_A, Relatio rises to 0.63 (+18.9%) and 0.72 (+18.0%). Under **H2**, LLM-based narrative synthesis also helps: CaNarEx rises to 0.65 (+20.4%) and 0.75 (+19.0%), Relatio rises to 0.66 (+24.5%) and 0.76 (+24.6%). Despite these gains, DiNO remains best overall (0.78, 0.85), indicating that targeted segment selection and LLM-driven synthesis are jointly important.

To illustrate the benefit of segment-focused processing, Table 30 presents example Sputnik–Ukraine articles alongside their corresponding false segments and verified false claims. As shown, full-article inputs can introduce irrelevant context, whereas DiNO's segmentation isolates only the relevant false content, enabling cleaner clustering and more accurate narrative extraction.

Likewise, Table 31 demonstrates the impact of LLM-based narrative synthesis. It shows a set of false claims from a cluster identified by DiNO in the Breitbart-migration dataset, together with the narratives extracted by DiNO, CaNarEx, and Relatio. DiNO's LLM-generated narrative yields a coherent synthesis that captures both the topic and stance of the cluster, while CaNarEx and Relatio, constrained to selecting single representative sentences, produce less cohesive summaries.

D Robustness Analysis

D.1 Chamfer Distance Scores

We use Chamfer Distance (CD) as a set-level robustness check that captures both match quality and coverage between predicted and ground-truth narrative sets. Scores are computed with an LLM judge (Sonnet 4.5). For each predicted narrative, the judge selects the closest ground-truth narrative. For each ground-truth narrative, the judge selects the closest prediction. For each matched pair, the judge rates thematic alignment (TA) and stance alignment (SA) on a 0–5 Likert scale, which we normalize to $[0, 1]$ (prompts in Appendix B.4). CD averages the mean normalized scores in both directions (pred→GT and GT→pred). We also report a binarized version (CD^b) that thresholds each 0–5 score at 3 (1 if ≥ 3, else 0) before averaging.

Results. Across all three domains, disinformation-prone outlets achieve high CD_{TA} and CD_{SA}. This indicates that DiNO's extracted narrative sets align closely with expert narratives in both topic and stance and also provide

Dataset	CD_{TA}	CD^{b}_{TA}	CD_{SA}	CD^{b}_{SA}
Ukraine				
Sputnik	0.73	0.77	0.93	0.97
Pravda	0.70	0.76	0.90	0.95
Guardian	0.73	0.65	0.26	0.19
COVID-19				
NatNews	0.88	0.97	0.96	0.98
Sputnik	0.85	0.94	0.94	0.93
Guardian	0.79	0.87	0.13	0.02
Migration				
Sputnik	0.81	0.91	0.90	0.95
Breitbart	0.84	0.92	0.91	0.96
Guardian	0.69	0.85	0.35	0.24

Table 21: Chamfer Distance scores between DiNO-extracted and ground-truth narratives across all dataset.

	mTA_A	mSA_A	CD_{TA}	CD_{SA}
EUDisinfoTest				
Sputnik	0.78	0.85	0.73	0.93
Pravda	0.69	0.83	0.70	0.90
Guardian	0.70	0.00	0.73	0.26
SemEval				
Sputnik	0.78	0.97	0.79	0.94
Pravda	0.79	0.96	0.80	0.93
Guardian	0.90	0.30	0.64	0.35
Atlantic Council				
Sputnik	0.81	0.98	0.79	0.94
Pravda	0.83	0.97	0.80	0.93
Guardian	0.90	0.10	0.58	0.31

Table 22: Scores for DiNO-extracted narratives from the Ukraine-war articles, evaluated against three ground-truth narrative sets (EUDisinfoTest, SemEval, Atlantic Council). We report mean alignment (mTA_A, mSA_A) and Chamfer Distance (CD_{TA}, CD_{SA}).

strong coverage of the ground-truth set. For *The Guardian*, CD_{TA} remains moderate to high, while CD_{SA} and CD^{b}_{SA} are low. This indicates topical overlap with expert narratives but limited stance-aligned matches, which is consistent with credible reporting that discusses the same themes while rejecting the disinformation stance. The binarized scores (CD^{b}) provide a stricter check that mitigates inflation from nearest-neighbor pairing, since each predicted narrative and each ground-truth narrative is always matched to a counterpart. High CD^{b} on disinformation-prone outlets shows that the results are not driven by low-quality pairings, because most matched pairs exceed the $3/5$ threshold. In contrast, CD^{b}_{SA} remains near zero for *The Guardian*, indicating that stance-aligned matches are largely absent under a conservative criterion.

D.2 Robustness to Alternative Ground-Truth Narrative Sets (Ukraine)

We also test whether our evaluation for the Ukraine domain depends on the ground-truth narrative set used for scoring. We keep the DiNO-extracted narratives produced from the Ukraine war articles fixed, and we recompute the same mean alignment scores (mTA_A, mSA_A) and Chamfer Distance scores (CD_{TA}, CD_{SA}) against two additional expert ground-truth narrative sets that explicitly cover the Russia–Ukraine war, sourced from SemEval (Piskorski et al., 2025) and the Atlantic Council report on Kremlin narratives about Ukraine (Aleksejeva, 2023). We use the same judge and prompts as in Section D.1. Table 22 reports the resulting scores, and we include EUDisinfoTest for comparison.

Results. The qualitative pattern is stable across ground-truth sources. Sputnik and Pravda remain high in both thematic and stance scores, and their Chamfer scores remain high. *The Guardian* remains topically similar but stance-divergent. This indicates that the main conclusion does not depend on a single ground-truth narrative set.

E Automated LLM Judge

This appendix describes our automated LLM-based evaluation setup and motivates our choice of a default judge for large-scale scoring.

Prompts. We use prompts aligned with the human protocol (Section 5.6). Because EUDisinfoTest provides multiple ground-truth narratives per case, evaluation proceeds in two steps. **(1) Matching:** given the DiNO-generated narrative and a set of EUDisinfoTest candidate narratives, the judge selects the most similar ground-truth narrative. **(2) Scoring:** the judge then rates *topical alignment* (TA) and *stance alignment* (SA) between the generated narrative and the selected reference on separate 0–5 Likert scales. Full prompts and the scoring rubric are provided in Section B.4.

Selecting an Automated LLM Judge. To enable scalable evaluation, we benchmark three LLMs as automated judges: GPT-4.1, Sonnet 4.5, and Gemini 2.5 Pro. Each judge independently rated DiNO-extracted narratives for every dataset (Section 5.6), using the same TA/SA definitions and the EUDisinfoTest references. To reduce stochasticity, we set temperature to 0 for all judge models.

We quantify agreement between each LLM judge and the pooled human annotations using within-group agreement, r_{WG}, computed separately for TA and SA (Table 23).

LLM Judge	TA r_{WG}	SA r_{WG}
Sonnet 4.5	**0.83**	0.73
GPT-4.1	**0.83**	0.74
Gemini 2.5 Pro	0.82	**0.75**

Table 23: r_{WG} agreement with pooled human annotators for topical alignment (TA) and stance alignment (SA). Values above 0.70 are commonly interpreted as strong agreement.

All three models show comparable agreement with humans, with small differences across TA and SA. Since we use GPT-5 for narrative generation, we select Sonnet 4.5 as the default automated judge to reduce the risk of same-family bias and to diversify model families, while noting that results are consistent across judges.

Human vs. automated evaluation. Despite this high agreement, discrepancies are systematic: the LLM-judge tends to slightly down-weight topical alignment when narratives are topically similar but differ in stance, whereas human annotators still treat such pairs as topically aligned. For example, when two narratives concern the same topic but one is "pro-Ukraine" and the other "anti-Ukraine," humans assign high topical alignment, whereas the LLM-judge yields a lower topical score.

F Validating Retrieval and Entailment Verification

We provide additional evidence motivating Step 1 (matching false segments) and Step 2 (entailment verification). Step 1 retrieves segments that are similar to a claim, but similarity can arise from topical overlap even when the article does not actually support the claim. Step 2 mitigates this by checking whether the full article entails the claim (article-level evidence).

We therefore report two analyses: (i) a manual check that the Step 1 retrieved segment is still claim-relevant for the cases that Step 2 retains as *support*, and (ii) an ablation measuring entailment performance when Step 2 is applied to the retrieved segment (with or without nearby context) instead of the full article.

Unless stated otherwise, all experiments use the optimal settings reported in Table 13.

F.1 Segment-Claim Mention Check

Step 1 retrieves claim-segment pairs (c, s) from each article a, where c is a verified false claim.

Subset	N	Mentions (%)	IAA (κ)
Sputnik-Ukraine	100	97.0	0.78
Breitbart-Migration	100	96.0	0.65
NaturalNews-COVID	100	97.0	0.72

Table 24: Claim-relevance check for Step 1 retrieval on supported matches. N is the number of annotated (c, s) pairs, sampled from cases where Step 2 labels the corresponding article-claim pair as *support* using article-level evidence. *Mentions* indicates that the retrieved segment includes a similar proposition to the matched claim. IAA is Cohen's κ.

Step 2 then evaluates the corresponding article–claim pairs using the full article a and keeps only those labeled *support*. Because Step 2 uses article-level evidence, we verify that the segment retrieved by Step 1 is still the right evidence for these retained cases.

Setup. We run DiNO Step 1 on three datasets (Sputnik-Ukraine, Breitbart-Migration, and NaturalNews-COVID) with their corresponding verified false claims (Section 1), then apply Step 2 to the resulting article–claim pairs. From the subset that Step 2 labels as *support*, we randomly sample 100 retrieved pairs (c, s) per dataset for manual validation.

Two annotators independently label a pair as *Mentions* if the retrieved segment s expresses the same (or an equivalent) proposition as the claim c (i.e., "Does segment s include a proposition similar to the one in claim c?").

Result. As shown in Table 24, 96–97% of sampled segments are labeled *Mentions*. This indicates that, for the supported pairs retained by Step 2, the segments retrieved by Step 1 typically contain the proposition of the matched claim.

F.2 Evidence Granularity for Entailment Verification

We next test how Step 2 performance changes when entailment verification is performed using only the retrieved segment versus broader textual context.

Setup. We use the evaluation dataset from Section 5.2 (450 claim-article pairs: 150 support, 150 debunking, 150 unrelated). For each pair, we run Step 1 and take the top-ranked retrieved segment s^*. We then run Step 2 with three evidence variants:

- **Article:** the full article text.
- **Segment:** s^* only.

Step-2 evidence	F1	$P_{support}$
Article-level (default)	**0.94**	**0.90**
Segment + context (2 sent.)	0.91	0.83
Segment-only	0.84	0.77

Table 25: Ablation of Step 2 evidence granularity (Step 1 fixed, only Step 2 evidence varies).

- **Segment + context:** s^* with a local window (two sentences preceding and two sentences following s^* in the article).

Metrics. We report F1 and precision on the *support* class ($P_{support}$), following the main paper.

Results. Table 25 shows that article-level evidence performs best. Segment-only verification is substantially worse, while adding local context partially recovers performance. This suggests that entailment often relies on information outside the single most similar segment (e.g., attribution or refutation), motivating our use of full-article evidence in Step 2.

G Extended Narrative Analysis

This appendix provides supplementary analyses of the disinformation narratives extracted by DiNO. Subsection G.1 examines their temporal distribution and relates narrative peaks to contemporaneous events. Subsection G.2 evaluates the similarity of narratives identified across different outlets

G.1 Narratives Over Time and Relation to Events

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

Breitbart-Migration. Figure 5 shows that DiNO finds two main waves in Breitbart's migration coverage: a high period in 2016-2019, a sharp drop around 2020, and then a steep rise again from late 2023 that peaks in 2024-early 2025.

In the 2016-2019 wave, DiNO frequently detects narratives such as *Muslim and non-European migrants cause a Europe-wide crime wave, creating no-go zones and societal collapse* and election-manipulation narratives like *Democrats want open borders and mass amnesty to import immigrant voters and permanently control elections*.

The clear collapse around 2020 is consistent with work showing that COVID displaced other issues

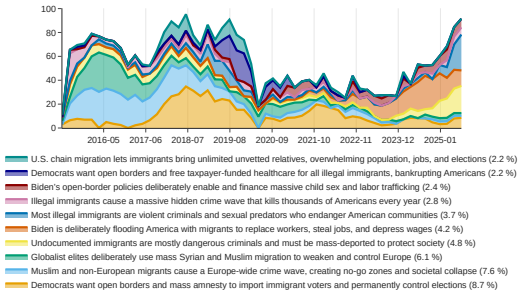

Figure 5: Temporal distribution of the ten main disinformation narratives in the Breitbart-Migration dataset. The y-axis represents the absolute number of articles containing each narrative within a given time period.

in political media (Reedy et al., 2023). From late 2023 onward, DiNO finds the largest surge, now driven mainly by U.S.-domestic and Biden-targeted narratives such as *Biden is deliberately flooding America with migrants to replace workers, steal jobs, and depress wages* and *Biden's open-border policies deliberately enable and finance massive child sex and labor trafficking*, which rise into the 2024 election period (Martinez).

Overall, DiNO's Breitbart-Migration narratives seem aimed at U.S. domestic audiences and at reinforcing right-wing agendas.

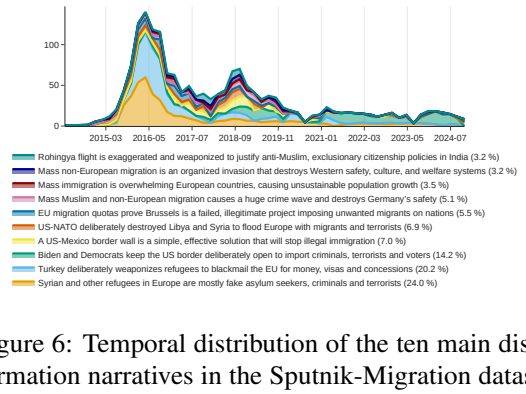

Figure 6: Temporal distribution of the ten main disinformation narratives in the Sputnik-Migration dataset. The y-axis shows the number of articles containing each narrative in a given period.

Sputnik-Migration. Figure 6 shows that DiNO finds one dominant surge in late 2015, a smaller wave around 2018, and then much lower but recurring activity from 2021 onward. The largest spike (late 2015) is where DiNO detects a strong jump in *Syrian and other refugees in Europe are mostly fake asylum seekers, criminals and terror-*

ists and *Turkey deliberately weaponizes refugees to blackmail the EU for money, visas and concessions*, coinciding with the European refugee crisis and the EU-Turkey deal period (Terry, 2021). In the same surge, DiNO also captures geopolitical blame framing such as *US-NATO deliberately destroyed Libya and Syria to flood Europe with migrants and terrorists*.

A second wave around 2018 is where DiNO finds increased emphasis on institutional distrust, especially *EU migration quotas prove Brussels is a failed, illegitimate project imposing unwanted migrants on nations*, matching conflict inside the EU over migration governance and burden-sharing (Gray Meral, 2020). From 2021 onward, DiNO shows a shift toward U.S. border politics, led by *Biden and Democrats keep the US border deliberately open to import criminals(...)* and *A US-Mexico border wall is a simple, effective solution that will stop illegal immigration*, coinciding with heightened attention to high encounter levels at the southern border (U.S. Customs and Border Protection).

Overall, Sputnik's main narratives blame the West and frame migrants as a threat or burden.

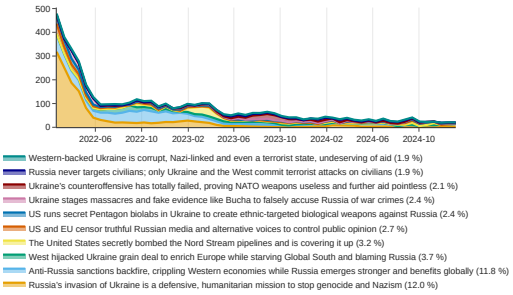

Figure 7: Temporal distribution of the ten main disinformation narratives in the Sputnik-Ukraine dataset. The y-axis represents the absolute number of articles containing each narrative within a given time period.

Sputnik-Ukraine. Figure 7 shows that DiNO finds Sputnik-Ukraine dominated by two long-running frames: *Russia's invasion of Ukraine is a defensive, humanitarian mission to stop genocide and Nazism* and *Anti-Russia sanctions backfire, crippling Western economies while Russia emerges stronger and benefits globally*. The largest surge is on the early 2022, where DiNO detects a sharp spike in the invasion-justification narrative, coinciding with Russia's public "denazification" rationale. After mid-2022, overall volume drops, but DiNO still captures short bursts around specific events in late 2022, including *West hijacked Ukraine grain deal to enrich Europe while starving Global South(...)* and *The United States secretly bombed the Nord Stream pipelines and is covering it up*, aligning with the Black Sea Grain Initiative period and the Nord Stream leaks/explosions (United Nations, 2022).

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

Figure 8: Temporal distribution of the ten main disinformation narratives in the Pravda-Ukraine dataset. The y-axis represents the absolute number of articles containing each narrative within a given time period.

Pravda-Ukraine. Figure 8 shows that DiNO finds Pravda-Ukraine driven by battlefield and escalation framing, led by *Ukraine's army is collapsing while Russia humanely targets only military sites(...)* and *NATO secretly controls Ukrainian long-range missile strikes, making the West a direct belligerent against Russia*. The largest surge is in late March, May 2024, where DiNO detects a pronounced spike in *Ukraine, directed by US-UK intelligence, organized the Crocus City Hall terrorist attack, not ISIS*, coinciding with the Crocus City Hall attack and contemporaneous reporting focusing on ISIS-K (Reuters, 2024). That period also shows higher volume for *Macron and NATO are secretly preparing large-scale French-led troop intervention in Ukraine against Russia*, aligning with Macron's remarks that he would not rule out European troop deployments (Rose and Irish, 2024).

Overall, Pravda attributes escalation and economic harm to Western actors, denies Ukrainian agency by casting Ukraine as a controlled proxy, and frames Russia as restrained and defensive.

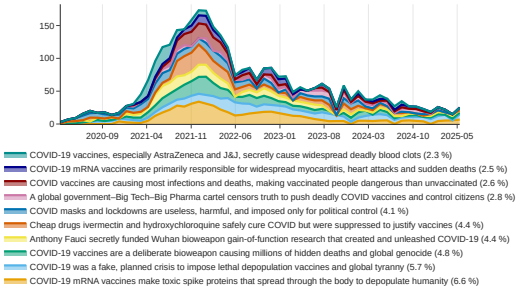

Figure 9: Temporal distribution of the ten main disinformation narratives in the NaturalNews-COVID dataset. The y-axis represents the absolute number of articles containing each narrative within a given time period.

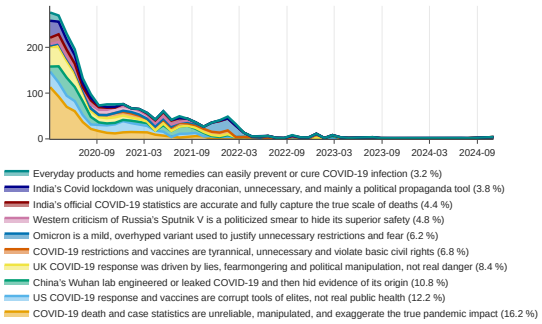

Figure 10: Temporal distribution of the ten main disinformation narratives in the Sputnik-COVID dataset. The y-axis represents the absolute number of articles containing each narrative within a given time period.

NaturalNews-COVID. Figure 9 shows that DiNO finds the NaturalNews' COVID output to peak at the end of 2021 and then steadily declines. The peak is mainly driven by vaccine-as-depopulation narratives, especially *COVID-19 mRNA vaccines make toxic spike proteins that spread through the body to depopulate humanity*, and *COVID-19 was a fake, planned crisis to impose lethal depopulation vaccines and global tyranny*. DiNO also picks up an origin-blame strand in the same period:*Anthony Fauci secretly funded Wuhan bioweapon gain-of-function research that created and unleashed COVID-19*, which mirrors a wider public debate about the origins of COVID-19 (World Health Organization, 2021). At lower but recurring levels, NaturalNews promotes control and "suppressed cure" messaging, including *Cheap drugs ivermectin and hydroxychloroquine that safely cure COVID but were suppressed to justify vaccines*, despite systematic reviews that found no convincing benefit from ivermectin (Popp et al., 2021).

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

Sputnik-COVID. Figure 10 shows a sharp spike in Sputnik COVID narratives in mid-2020, followed by a steady decline with only minor upticks through early 2022.

The early peak is dominated by trust-undermining claims: *COVID-19 death and case statistics are unreliable, manipulated, and exaggerate the true pandemic impact* or *US COVID-19 response and vaccines are corrupt tools of elites, not real public health*. This aligns with the broader surge of COVID disinformation and declining trust in public-health messaging in 2020.

Across 2020–2021, DiNO also identifies an origins theme: *China's Wuhan lab engineered or leaked COVID-19 and then hid evidence of its origin*, which tracks public attention to competing origin explanations (World Health Organization, 2021).

A smaller bump appears in late 2021-early 2022, where DiNO finds increased emphasis on *Omicron is a mild, overhyped variant used to justify unnecessary restrictions and fear* together with *COVID-19 restrictions and vaccines are tyrannical, unnecessary and violate basic civil rights*, consistent with renewed debates and protests over COVID measures.

Overall, the narratives extracted by DiNO in Sputnik-COVID combined international conspiracy tropes with explicit promotion of Russian biomedical achievements.

G.2 Cross-outlet Narratives Evaluation

In this section, we examine how closely the narratives extracted by DiNO align across different media outlets within the same topical domain. We compute two measures of cross-outlet similarity for each domain: the mean Topical Alignment (mTA_{A}) and mean Stance Alignment (mSA_{A}).

Guardian. Across domains, pairs that include Guardian typically show low stance alignment but high topical alignment. This suggests that credible outlets cover the same topics as disinformation-prone media outlets, but in opposing ways.

Outlet-Pair	mSA_A	mTA_A
Ukraine War		
Pravda-Guardian	0.35	0.89
Sputnik-Guardian	0.47	0.83
Pravda–Sputnik	0.94	0.92
COVID-19		
Naturalnews-Guardian	0.15	0.54
Sputnik-Guardian	0.20	0.45
Naturalnews–Sputnik	0.88	0.87
Migration		
Breitbart-Guardian	0.34	0.77
Sputnik-Guardian	0.34	0.64
Breitbart–Sputnik	0.92	0.89

Table 26: Mean Stance and Topical Similarity for outlet pairs by domain.

Ukraine War. Pravda and Sputnik match the closest: they reuse the same pro-Kremlin narratives with only small wording changes. The main difference is the target audience: Pravda focuses on a domestic Russian readership (to justify the war and security claims), while Sputnik targets foreign audiences (to persuade and create division, especially in Europe). This shows up in paired "justification" narratives like *Russia's invasion of Ukraine is a defensive operation to stop Ukrainian genocide and Nazi aggression* versus *Russia's invasion of Ukraine is a defensive, humanitarian mission to stop genocide and Nazism*, and in foreign-facing "blowback" narratives such as *Western anti-Russian sanctions failed, secretly enrich the US, and mainly destroy Europe's economy* matched with *Anti-Russia sanctions backfire, crippling Western economies while Russia emerges stronger(...)*.

COVID-19. NaturalNews and Sputnik agree on the basic distrust: public health authorities are cheating, restrictions are about control, and vaccines are harmful, while serving different roles. NaturalNews is a conspiracy/pseudoscience ecosystem, whereas Sputnik is a state outlet restricted in the EU. Their overlap shows up in vaccine narratives like *COVID-19 vaccines are harmful, ineffective, and pushed endlessly by profiteering leaders who avoid them* paired with *COVID-19 vaccines are unsafe, ineffective, rushed, and pushed through political and media manipulation*, and in restriction-as-control narratives such as *COVID masks and lockdowns are useless, harmful, and imposed only for political control* matched with *COVID mask and isolation rules are tools for political and social control, not health*.

Migration. Breitbart and Sputnik align strongly on migration-as-threat and migration-as-deliberate-strategy narratives, but in different political settings. Breitbart focuses on U.S. domestic partisan conflict (border politics, elections, crime), while Sputnik uses similar narratives to amplify polarization and portray Western states as chaotic or hypocritical, including for European audiences. The closest overlaps include *Biden intentionally opened the U.S.-Mexico border, causing historic illegal immigration, crime, and catastrophe* aligned with *Biden and Democrats keep the US border deliberately open to import criminals, terrorists and voters*, and institutional-corruption narratives like *Immigration lawyers and NGOs systematically coach migrants to lie and fraudulently obtain asylum benefits* mirrored verbatim in Sputnik's matched narrative.

H Examples of False Segment Matches

Table 29 presents representative examples of correct and incorrect matches. These examples highlight both the strengths and limitations of the system in linking articles to verified false claims.

I Examples of False Claims and Narratives

Table 27 illustrates several common narratives alongside representative falsehoods drawn from fact-checking sources.

J Annotation Guidelines

This section describes annotation procedures for the narrative-alignment task used for evaluating DiNO across outlets and domains.

J.1 Narrative Alignment Task

Purpose. This annotation task evaluates the **topical** and **stance alignment** between system-generated narratives and their most topically similar ground-truth disinformation narratives.

Materials. Each annotation instance comprises predicted narratives generated by DiNO and corresponding ground-truth disinformation narratives from the EUDisinfoTest dataset (Sosnowski et al., 2024). Predictions are produced separately for articles from all the news outlets in Table 1.

Annotation Task. The annotation process consists of two steps:

Step 1: Matching. For each narrative, annotators review the set of ground-truth narratives and select the one that is most topically similar.

Disinformation Narrative	Related false claims
Ukrainians are Nazis (EUvsDisinfo, 2021)	1. President Zelensky wore a swastika jersey (Snopes, 2022). 2. Russian is banned in Ukraine (PolitiFact, 2022).
The 2020 U.S. election was stolen (Balsamo, 2020)	1. Voting machines switched votes (AAP FactCheck, 2024). 2. Ballots in Arizona were invalidated (Spencer, 2020).
COVID-19 vaccines are dangerous (CDC, 2024)	1. Vaccines contain microchips (AFP Fact Check, 2021). 2. mRNA vaccines alter human DNA (CDC, 2024).
The pandemic removes sovereignty (PolitiFact, 2024)	1. WHO forces lockdowns or vaccine mandates (PolitiFact, 2024). 2. WHO pandemic agreement overrides national sovereignty (Jaramillo, 2023).
Migrants are invading Europe (EUvsDisinfo, 2017)	1. 7,000 migrants arrived in Lampedusa in 36 hours (Reuters Fact Check, 2024a). 2. Many migrants flew over to Britain (Reuters Fact Check, 2024b).

Table 27: Examples of disinformation narratives and their constituent false claims.

Step 2: Scoring. For each matched pair, annotators assign two scores on a scale from 0 to 5:

- **Topical Alignment:** The extent to which both narratives address the same topic. (0 = completely different topics, 5 = identical topic)
- **Stance Alignment:** The extent to which both narratives share the same perspective (e.g., both spread disinformation, both debunk, or are opposites). (0 = opposing stance, 5 = identical stance)

Annotators. Each instance is independently annotated by two expert annotators with professional fact-checking experience. Both annotators are certified by the International Fact-Checking Network (IFCN) and have prior experience assessing disinformation in fact-checking workflows. Annotators complete both steps independently. If annotators are unsure or disagree at any step, they discuss the instance and reach consensus.

Inter-Annotator Agreement (Pre-Consensus). To assess reliability, we retain the two annotators' independent (pre-consensus) annotations and report agreement separately for each step of the task. For matching, agreement is computed as the percentage of instances where both annotators selected the same ground-truth narrative. For scoring, agreement is computed as the percentage of instances where both annotators assigned the same 0–5 score. Using these exact-match criteria, pre-consensus agreement is 79.9% for matching, 73.2% for topical alignment, and 85.1% for stance alignment.

Instructions.

- For each DiNO-extracted narrative, select the ground-truth narrative from the set with the highest topical similarity.
- Assign one topical score and one stance score (0-5) to the matched pair.
- Base all judgments solely on the narrative texts and do not use external knowledge.
- Discuss and reach consensus in cases of uncertainty or disagreement.

Abbr.	API/HuggingFace Model Name	Access Details	License	Model Size
GPT-4.1 (OpenAI, 2025a)	`gpt-4.1`	OpenAI API	Commercial	Not Disclosed
GPT-5 (OpenAI, 2025b)	`gpt-5`	OpenAI API	Commercial	Not Disclosed
GPT-5-mini (OpenAI)	`gpt-5-mini`	OpenAI API	Commercial	Not Disclosed
SFR-Embedding-2 (Meng et al., 2024)	`Salesforce/SFR-Embedding-2_R`	HuggingFace	CC BY-NC 4.0	7B
E5-large (Wang et al., 2022)	`intfloat/e5-large`	HuggingFace	MIT	335M
Jina (Sturua et al., 2024)	`jinaai/jina-embeddings-v3`	HuggingFace	CC BY-NC 4.0	572M
Sonnet 4.5 (Anthropic, 2025)	`claude-sonnet-4-5-20250929`	Anthropic API	Commercial	Not Disclosed
Gemini-2.5-Pro (Google, 2025a)	`gemini-2.5-pro`	Google API	Commercial	Not Disclosed
RoBERTa-large (Liu et al., 2019)	`roberta-large`	HuggingFace	MIT	355M
DeBERTa-large (He et al., 2020)	`microsoft/deberta-large`	HuggingFace	MIT	400M
gpt-oss-120b (OpenAI, 2025c)	`openai/gpt-oss-120b`	DeepInfra	Apache 2.0	117B
Llama-3.3-70B (Meta, 2024)	`meta-llama/Llama-3.3-70B-Instruct-Turbo`	DeepInfra	Llama 3.3 Community License	70B
Qwen3-14B (Yang et al., 2025)	`Qwen/Qwen3-14B`	DeepInfra	Apache 2.0	14B

Table 28: Detailed overview of the language models used in the evaluation. (*Abbr.* stands for abbreviation)

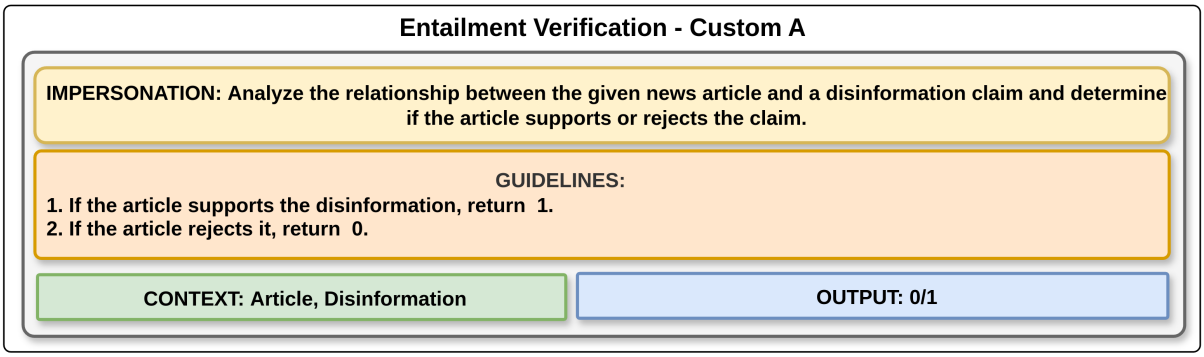

Figure 11: Custom zero-shot entailment verification prompt A used in our approach.

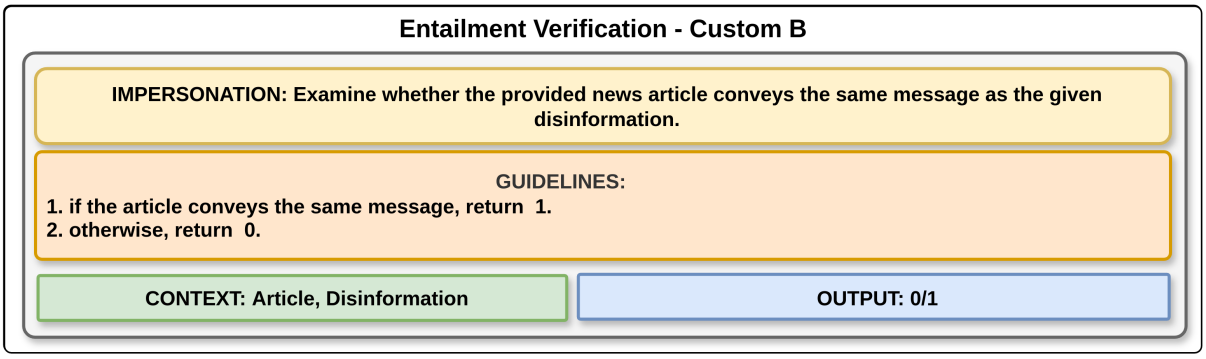

Figure 12: Custom zero-shot entailment verification prompt B used in our approach.

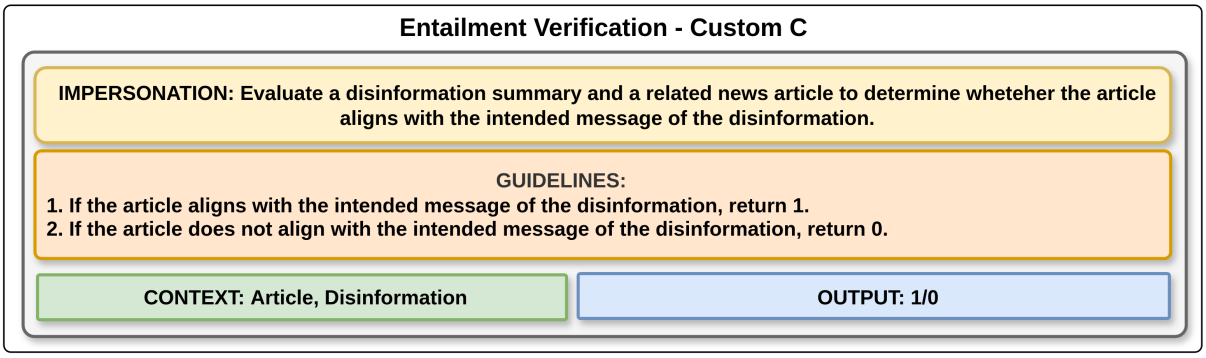

Figure 13: Custom zero-shot entailment verification prompt C used in our approach.

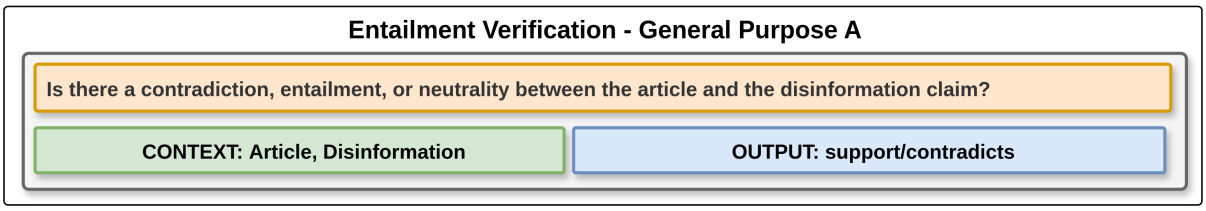

Figure 14: Baseline Prompt A: A zero-shot entailment verification prompt inspired by (Gallipoli and Cagliero, 2025).

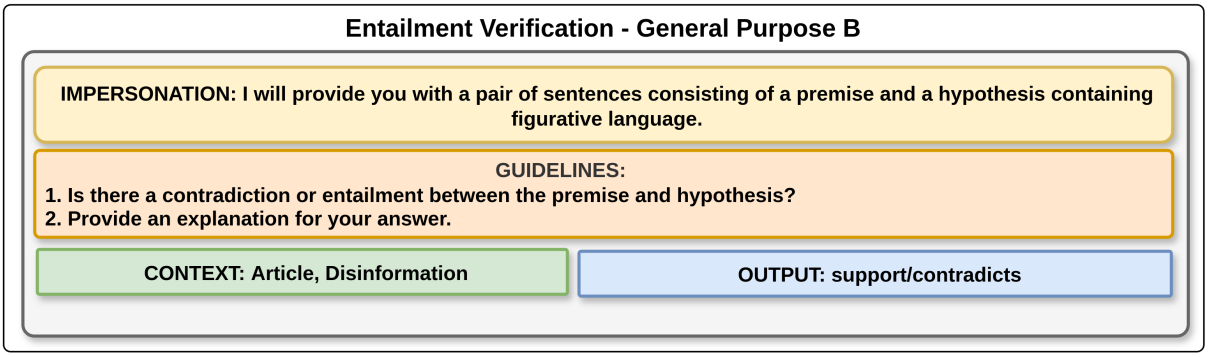

Figure 15: Baseline Prompt B: A zero-shot entailment verification prompt inspired by (Gallipoli and Cagliero, 2025).

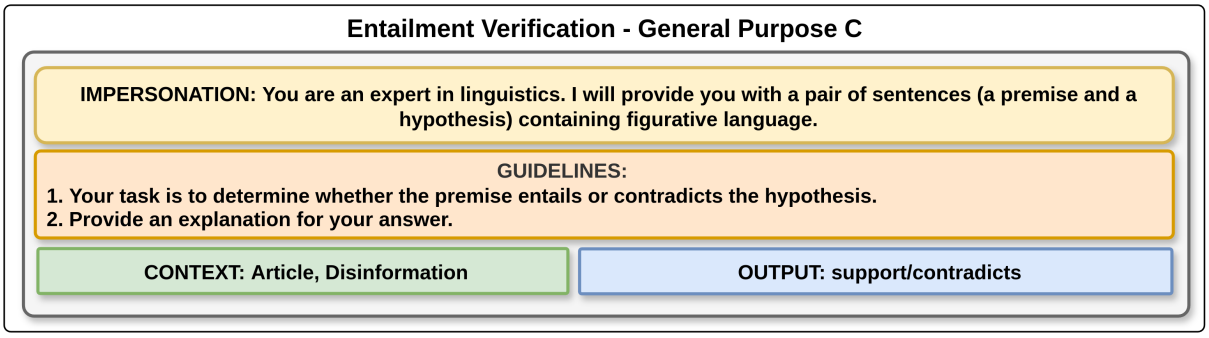

Figure 16: Baseline Prompt C: A zero-shot entailment verification prompt inspired by (Gallipoli and Cagliero, 2025).

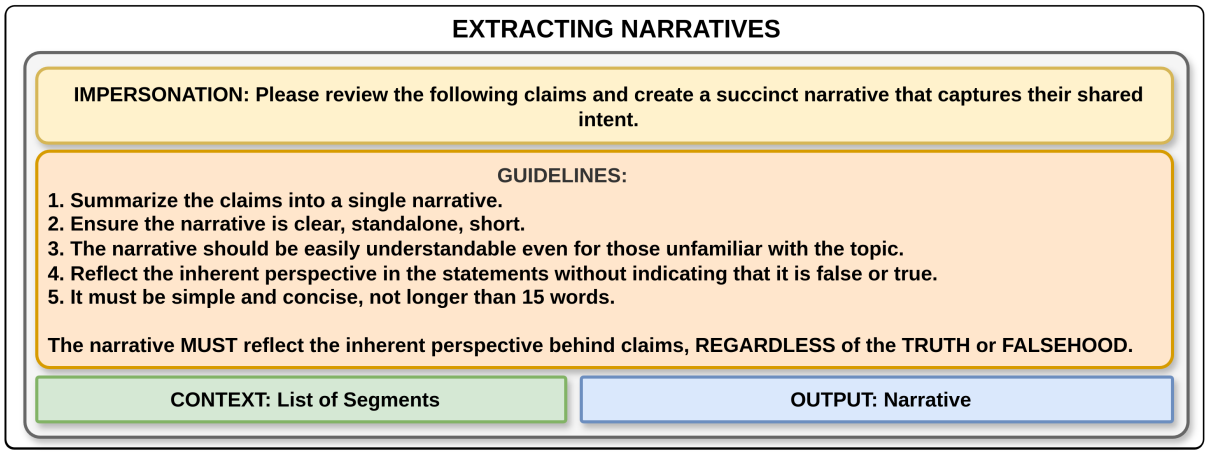

Figure 17: A prompt designed for deriving disinformation narrative given a set of disinformation chunks.

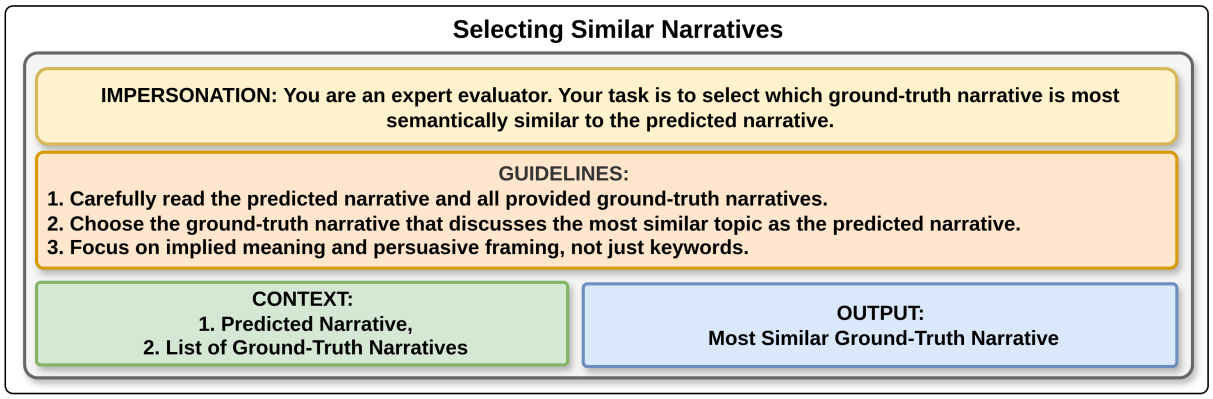

Figure 18: A prompt designed for selecting most similar ground truth narrative given a predicted narrative.

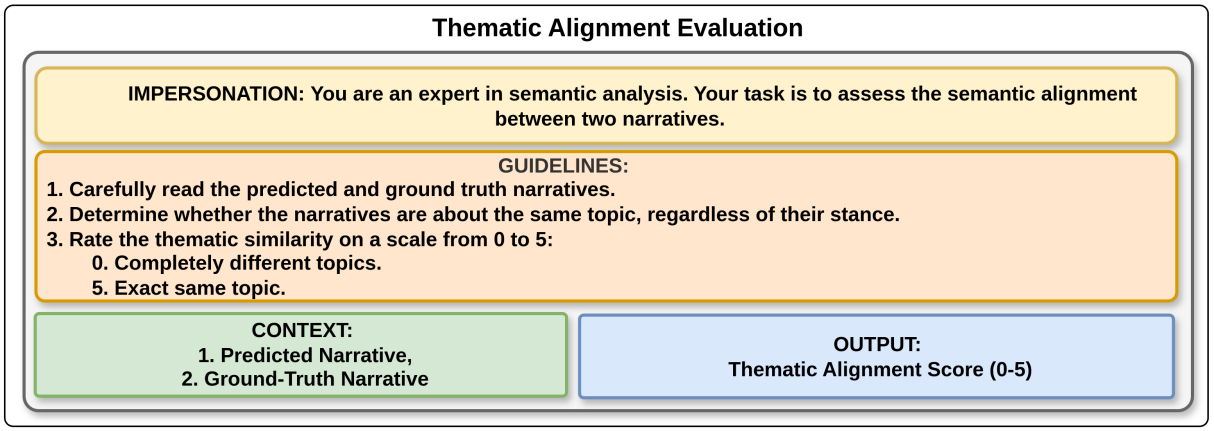

Figure 19: A prompt designed for assigning topical alignment score given a predicted and ground truth narratives.

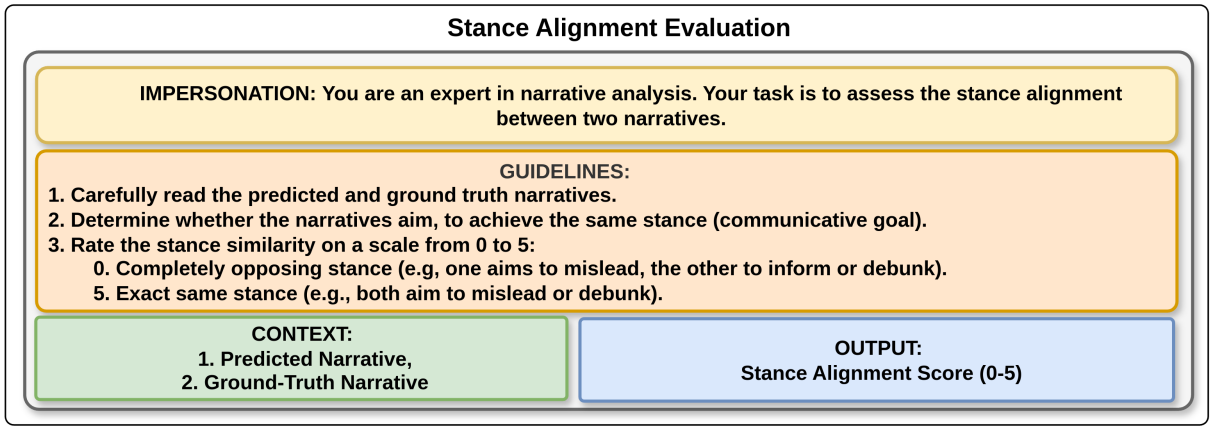

Figure 20: A prompt designed for assigning stance alignment score given a predicted and ground truth narratives.

False Claim	Matched Segment	Status
Ukraine became one of the main testing grounds for Western pharma, with experiments in psychiatric hospitals in the city of Mariupol where, according to documents acquired by Sputnik, an experimental drug called SB4 that is linked to several types of cancer was tested.	According to documents obtained by Sputnik, major Western pharmaceutical companies, with the assistance of Ukrainian officials, purportedly conducted testing of rheumatological drugs on patients in a Mariupol psychiatric ward for several years.	**Correct**
Crimea became part of Russia in March 2014 after a referendum, held after the coup in Kiev. 96.77 percent of the voters of the Crimea and 95.6 percent of the residents of Sevastopol voted in favour of unification with Russia.	16 March 2014: Crimea holds a referendum on its future status. Amid turnout of over 80 percent, more than 95 percent of voters choose to rejoin Russia. On 18 March, after their request is approved by Moscow, Crimea and the port city of Sevastopol officially become part of Russia.	**Correct**
The Zelenskyy regime has launched a final offensive against Orthodoxy. The attack on the Kyiv-Pechersk Lavra characterises the Zelenskyy regime as absolutely inhuman. Everything we are witnessing is evidence that the followers of Satan have seized power in Kyiv.	The Zelensky regime has once again demonstrated that it holds nothing sacred, this time in the literal sense. Today, the authorities have begun a gangster-style takeover of the main shrine of Orthodoxy – the Kiev-Pechersk Lavra.	**Correct**
The United States decided to reformat the global security system in their favour. What is important here is the US withdrawal from the Intermediate-Range Nuclear Forces Treaty (the INF Treaty). This decision was first mentioned by President Trump in October 2018. What is absolutely not surprising is that they had found "weighty" arguments to justify this decision. Allegedly, back in 2008, Russia developed a missile that can cover a distance of over 500km. Moscow's arguments that the "suspicious" missile 9M729 has a much smaller flight range have been simply ignored by Washington.	The missiles were withdrawn after the US and the Soviet Union signed the Intermediate-Range Nuclear Forces (INF) treaty in 1987. Donald Trump withdrew the US from the INF treaty in 2019 after years of US complaints that Russia was cheating by building and deploying an intermediate-range missile, the 9M729. Russia denied the charge, claiming the range of the missile was legal.	**Incorrect**
The 98-year-old Ukrainian veteran of WWII, Yaroslav Hunka, who served in the Nazi Waffen-SS division, was honoured with long applause in the Canadian Parliament. He was applauded only because he fought against the Russians.	The Kremlin said it was "outrageous" the speaker of Canada's House of Commons had praised an individual at a parliamentary meeting who served in a Nazi unit during the Second World War. Canadian Speaker Anthony Rota apologised on Sunday after recognizing 98-year-old Yaroslav Hunka as a "Ukrainian hero" before the Canadian parliament. Hunka, who served in the Second World War as a member of the 14th Waffen Grenadier Division of the SS, received two standing ovations from lawmakers during a visit by Zelenskiy.	**Incorrect**
The investigation into the MH17 crash is conducted by Dutch prosecutors and the Dutch-led Joint Investigation Team (JIT). The international group of investigators claims that the plane was downed by a Buk missile belonging to the Russian armed forces, something Moscow categorically denies. Russia conducted its own investigation and said it had provided the JIT with evidence, including radar data, showing that the plane had been shot down by a Ukrainian Buk missile. However, Moscow was denied any access to the probe.	Russia has always denied any involvement in the shooting down of the plane. The Russian foreign ministry spokeswoman, Maria Zakharova, accused Dutch authorities last week of orchestrating a media campaign to compensate for gaps in the evidence and assembling facts to fit a predetermined verdict. The Dutch-led Joint Investigation Team (JIT) said in 2016 that it had found irrefutable evidence the Buk missile had been fired from a village under the control of pro-Russia rebels. The court is also expected to hear details of intercepted phone calls that reveal separatist leaders requesting help from senior Kremlin advisers shortly before MH17 was shot down.	**Incorrect**

Table 29: Examples of false claims from EUvsDisinfo with corresponding text segments matched by **DiNO-sgm** from the Sputnik Ukraine and Guardian Ukraine datasets. Correct matches are drawn from Sputnik Ukraine, while incorrect ones are from Guardian-Ukraine.

False Claim: *Western reports alleging Russia does not want talks are fake. The problem is that Zelenskyy is not an independent politician. He is manipulated and acting according to orders from outside [West].*
Matched False Segment: *Western control over Zelensky is not limited to the country's refusal to settle with Russia or its draconian draft laws. In addition, the West influences decisions within his government, dictating who comes and goes.*
News Article: **Western control over Zelensky is not limited to the country's refusal to settle with Russia or its draconian draft laws. In addition, the West influences decisions within his government, dictating who comes and goes.** Here's a quick overview of Zelensky's most memorable sackings: 2024 Serhiy Shefir, former first aide responsible for making business contacts and Zelensky's daily schedule. Oleksiy Danilov, ex-secretary of the National Security and Defense Council of Ukraine, ousted from the position he held from October 2019. Valery Zaluzhny, former Ukrainian Commander-in-Chief: The general's dismissal grabbed headlines worldwide amid claims of a spat with Zelensky. Kyrylo Tymoshenko, ex-deputy head of the presidential office, who oversaw regional policy from 2019. 2023 Oleksiy Arestovych, Zelensky's former communications advisor for defense and national security. He resigned after saying a residential building in Dnepropetrovsk had collapsed because Ukrainian air defenses shot down a Russian missile onto it, causing a public outcry. 2022 Ivan Bakanov, ex-head of Ukraine's security service, removed from the position due to suspected treason. 2021 Dmytro Razumkov, former speaker of the Verkhovna Rada, dismissed from the Ukrainian parliament by a vote of MPs. 2020 Andriy Bohdan, former Head of the Presidential Administration, now one of the fiercest critics of Zelensky and his drug addiction. Oleksiy Honcharuk, ex-PM: led the cabinet for less than five months before quitting over a leaked recording suggesting he had criticized the Zelensky. Ruslan Ryaboshapka, former Ukrainian Prosecutor General: ousted from the post by the pro-Zelensky majority in parliament. 2019 Oleksandr Danylyuk, ex-secretary of the National Security and Defense Council and Zelensky's campaign advisor: Resigned, saying publicly he was not able to stand behind-the-scenes struggle.

False Claim: *On 24 February, Russia launched a military operation in Ukraine to "protect people who have been abused and subjected to genocide by the Kyiv regime for eight years". To this end, there are plans to "demilitarise and denazify Ukraine" and to bring to justice all war criminals responsible for "bloody crimes against civilians" in the Donbas.*
Matched False Segment: *Russia launched its special military operation last month to put an end to war crimes and atrocities committed by Ukrainian troops against civilians during an eight-year offensive against the citizens of Donbas.*
News Article: On Thursday, the UN General Assembly voted to suspend Russia from the Human Rights Council in a 93-24 vote, which was applauded by US President Joe Biden. 08.04.2022, Sputnik International On Thursday, the UN General Assembly voted to suspend Russia from the Human Rights Council in a 93-24 vote, which was applauded by US President Joe Biden. Russia's Deputy Permanent Representative noted that the UNHRC bhas been monopolized by a group of countries which are exploiting it for their own purposes. Meanwhile, Western states continue to pump Ukraine with weapons, as almost simultaneously the US and Canada announced additional millions in military aid to Kiev. The Russian Ministry of Defense warned that Western arms shipments are a mistake and they increase casualties, but they won't affect the outcome of the special operation in Ukraine. **Russia launched its special military operation last month to put an end to war crimes and atrocities committed by Ukrainian troops against civilians during an eight-year offensive against the citizens of Donbas.** Averting the threat of the WWIII is one of the key goals of Russia's spec op, the Kremlin said on Thursday.

Table 30: Examples of Sputnik news articles containing false narratives aligned with corresponding disinformation claims. The matched false segment within each article is highlighted in bold red, and links to both the article and the disinformation source are included directly in the table.

DiNO Narrative: *Democrats use driver's licenses and automatic registration to let illegal immigrants vote in U.S. elections*

Relatio Narrative: *more electoral votes go to generally blue districts*

CaNarEx Narrative: *As a result, citizens in a district with lots of illegal aliens have more voting power than citizens in districts with few illegal aliens.*

Example Cluster of Matched False Segments:

1. "Counting illegal aliens when dividing up congressional seats and electoral college votes is likely to strip some red states of representation and give blue states with large foreign populations more representation."

2. "The reapportionment process, if approved by judges, would take House seats from California and Texas — both of which have swelled their populations with large inflows of illegals — and then transfer the seats to other states."

3. "A majority believe Biden is importing migrants to the U.S. to create a permanent Democrat majority. Indeed, immigration plays a significant role in congressional apportionment."

4. "Right now, congressional districts are drawn up simply based on the number of warm bodies in each district. Not only are legal aliens counted, but illegal aliens are counted too. As a result, citizens in a district with lots of illegal aliens have more voting power than citizens in districts with few illegal aliens."

5. "With only so many House seats available, if California or another confederate, Democrat-run state manages to grab extra House seats because their illegal alien population gives them that edge, those seats are taken from other states."

6. "Democrats are trying to hide the huge annual inflow of legal and illegal immigrants which are shifting cultural and political power towards the Democrats' progressive-led raucous coalition of minorities."

7. "Hagerty, in the first round, had a really good amendment saying that illegal aliens could not be included in the census in terms of apportionment of additional congressional seats,' even if an illegal alien is not able to vote, it still impacts elections 'boosting the numbers and having more House members, for example, more electoral votes go to generally blue districts,' 'So that's got to be the game plan of the Democrats here is change the electorate. They want a one-party state,' he said."

8. "The Census analysis is a break from years of research which projected the nation's illegal and legal immigration system has been helping not only deliver voters to Democrat politicians, but also provide residents to blue states to increase political dominance in Congress. Most recently, the Center for Immigration Studies projected that new Census numbers would shift about 26 congressional seats from mostly blue states to red states in 2020."

9. "Illegal immigrants and their U.S.-born minor children will redistribute five seats in 2020, with Ohio, Michigan, Alabama, Minnesota, and West Virginia each losing one seat in 2020 that they otherwise would have had."

Table 31: Comparison of narratives generated by DiNO, Relatio, and CaNarEx for a shared cluster of matched false segments from the Breitbart–Migration dataset. The table presents a representative sample of these false segments.