

DOI: 10.18372/2310-5461.71.21448

УДК: 004.8:316.776.23

Oleh Melnychuk, PhD student

National University "Kyiv Aviation Institute", Kyiv, Ukraine

<https://orcid.org/0009-0005-2768-4312>

e-mail: 5213201@stud.kai.edu.ua

SMALL LANGUAGE MODELS FOR MULTILINGUAL DISINFORMATION RETRIEVAL

Introduction

Problem statement. Information operations typically do not rely on isolated false statements but instead repeat a limited set of frames, varying actors, dates, locations, and news hooks. Analysts may encounter new articles with different wording but with underlying narratives that have already been documented. This scenario presents a retrieval challenge prior to verification: analysts must correctly identify the most relevant prior debunk to facilitate triage and contextualization. Automated fact-checking systems generally employ a retrieve-then-verify approach, where a claim is extracted, evidence is retrieved, and a verdict is rendered. Whilst effective for factual verification, this approach manages only part of disinformation monitoring, where the primary concern may be identifying which known hostile narrative a text reuses. The present study concentrates on this narrower task. It does not assess the veracity of arbitrary texts; instead, it rewrites incoming content into broad disinformation narratives and retrieves relevant records from a closed archive of known debunkings.

Analysis of recent research and publications

Automated fact-checking commonly separates retrieval from verification. FEVER formalized evidence retrieval against a reference corpus, and retrieval-augmented generation connects retrieved evidence to downstream generation [1–3]. Previously, fact-checking claim retrieval was closer to the present task because it searched an archive for a prior claim or debunking before repeating verification [4, 5]. Multilingual and cross-lingual claim matching further shows that semantic equivalence may persist across language changes, while surface overlap collapses [6]. Narrative retrieval is broader: the relevant relationship can be a recurring frame shared by several claims or events. Propaganda and framing research supplies the conceptual basis [7, 8], while computational narrative

studies show that structured coding benefits from explicit schemas [9, 10]. BM25 remains informative when a controlled query repeats title or record terms [11], but that strength can reflect raw/title identity leakage. Dense embeddings and rerankers are needed when the narrative is paraphrased [12–16]. This study reports both lexical and neural baselines and evaluates normalization as one component of a reproducible archive-search system.

Research gap. Prior studies establish evidence retrieval, previously fact-checked claim matching, multilingual semantic search, and propaganda-frame analysis, but they do not resolve how a compact normalizer should preserve the source-side proposition while removing incidental detail for archive retrieval. They also tend to evaluate a single target, whereas operational monitoring must handle articles that fit a documented narrative family without resolving to a single local debunk. This leaves a methodological gap between controlled traceability and realistic analyst-facing relevance.

Purpose of the article

The author aims to develop and evaluate a multilingual narrative-normalization pipeline for retrieving previously debunked disinformation and to determine whether local small language models can perform the normalization function competitively within a shared retrieval and reranking architecture. The evaluation separates controlled exact-document recovery from human-reviewed narrative-family alignment on real articles.

The author developed four components: a narrative-normalization and retrieval pipeline; a controlled-to-natural-article evaluation bridge; an anchored human rubric that separates narrative preservation, position, abstraction, and retrieval specificity; and a reproducible comparison framework that holds the retriever and reranker constant across normalizers. The resulting design tests exact traceability, multilingual robustness, and analyst-facing utility without collapsing them into one score.

RESEARCH METHODOLOGY

Corpus, datasets, and evaluation design

The study defines a narrative as a semantic frame that can preserve its causal structure and evaluative stance while actors, dates, locations, and other surface details change. Effective normalization removes incidental details without erasing the proposition, mechanism, target, or source-side position that distinguishes one narrative from another. This task supports monitoring, which relates new material to an existing information environment, rather than autonomous truth classification.

The retrieval corpus contains 7,538 EUvsDisinfo fact-check records [17]. Each record documents a disinformation claim and the corresponding correction or contextual response. The archive dates range from 6 January 2015 to 16 May 2026 and refer to EUvsDisinfo record metadata; a source article discussed by a record may have a different publication date. Its thematic range is illustrated by a Russia-focused title, “DISINFO: 50 countries gathered under Nazi flags to break Russia up”; a war-in-Ukraine title, “DISINFO: After 2014 coup, Ukraine committed genocide in Donbas”; and a Syria-related title, “DISINFO: Chemical attack in Douma fabricated by the White Helmets.” The separation between the source-side claim and the

corrective response matters because a normalizer must retain the manipulative proposition rather than turn the debunk into the query. The archive is closed and cannot retrieve cases it does not contain.

The source-linked article set was built from original source URLs in the separate EUvsDisinfo dataset published by Leite et al. [18]. Its 100 articles are query inputs, not additional records in the 7,538-record candidate archive. Retained fields include the source reference, title, text, language, publisher, and date; 93 distinct source debunk_id references are represented. Human bridge validation maps 39 articles to a unique local record, while the other 61 support narrative-family evaluation without an exact debunk. Example titles are “West has declared total war on Russian World – Lavrov” (Russia), “Russia had to begin special operation to stop genocide in Ukraine — Defense Minister” (war in Ukraine), and “Chris Williamson: How Can We TRUST the OPCW After Syria Chemical Attack Leaks?! (E757)” (Syria/Douma).

Table 1 presents the corpus and three evaluation inputs in a single view. It reports their denominators, languages, recorded date ranges, relevance definitions, and analytical roles. Dates for the source-linked articles use the value retained with the source reference; worked examples use a publisher-page date when available.

Table 1

Datasets and evaluation roles

Source / input	n	Lang.	Recorded dates	Purpose
EUvsDisinfo archive	7,538	ML	2015-01-06 to 2026-05-16	Closed search corpus
English controlled narratives	1,071	EN	2015-03-11 to 2026-05-16	Controlled comparison
Ukrainian controlled narratives	950	UK	2015-10-31 to 2026-05-14	Cross-language check
Original-100 articles	100	ML	2015-06-03 to 2023-02-22	Original article check

ML denotes multilingual, EN English, and UK Ukrainian. The controlled English set contains 1,071 records dated 11 March 2015 to 16 May 2026, and the controlled Ukrainian set contains 950 records dated 31 October 2015 to 14 May 2026. The source-linked set contains 100 articles dated 3 June 2015 to 22 February 2023.

The three evaluation inputs share the same 7,538-record retrieval archive, which serves as their common search space rather than as an additional test set. The controlled English and Ukrainian sets both measure exact-document recovery. Within each set, all normalizers process the same source *doc_id* values against that archive with the same Qwen3 retriever,

reranker, and exact-document relevance definition. Normalizer outcomes are therefore paired and directly comparable within each controlled set.

The controlled sets and Original-100 use this common search space for different validation tasks. The controlled sets assess exact traceability; Original-100 asks whether retrieved results are useful to an analyst at the narrative-family level. Human judgments cover all 100 articles, whereas exact-document metrics apply only to the 39 articles linked to a unique archive *doc_id*. Their results are therefore complementary, not estimates from a single population.

The unit of analysis is the input record in the controlled evaluations and the source article in Original-100. Recall, MRR, rubric scores, and categorical retrieval rates therefore describe performance on the evaluated inputs under the fixed retrieval and reranking configuration. Formal statistical comparisons are conducted in paired settings, with each normalizer evaluated on the same records. Reported analyses combine paired effect estimates, 95 % confidence intervals, and Holm-adjusted p-values to assess both the magnitude and uncertainty of observed differences. Non-significant comparisons are interpreted as not statistically distinguishable under this design; no formal equivalence or non-inferiority claim is made because no equivalence margin was prespecified. Bootstrap intervals quantify record- or article-level variability under the evaluated protocol.

The data span recurring geopolitical and public-health contexts. Syria-related records include Douma, chemical-attack, and OPCW narratives; Afghanistan records include war reporting and Bagram-related claims; Ukraine records cover Crimea and Donbas after 2014 and the full-scale invasion from 2022. NATO-Russia, MH17, and COVID-19 narratives recur across years. These event labels support descriptive slices rather than independent significance tests.

Table 2 contrasts a controlled-archive-derived query with a naturally occurring article retrieved via a source link in the separate EUvsDisinfo dataset [18]. It shows why exact-document recovery and narrative-family relevance require different evaluation criteria.

Table 2

Title and narrative examples

Title of source	Gemini Flash 3.5 narrative
DISINFO: Ukrainian authorities allegedly plan a catastrophic chemical incident in eastern Ukraine	The Ukrainian government is planning a catastrophic chemical disaster in the eastern region, using toxic warfare against its own citizens, which has sparked local protests
Chris Williamson: How Can We TRUST the OPCW After Syria Chemical Attack Leaks?!	The OPCW is actively suppressing evidence and leaking documents that prove the Douma chemical attack was a false flag operation orchestrated by militants rather than the Syrian government

For the source-linked example, the top three results are “Leaked OPCW report proves Douma chemical attack was staged,” “New OPCW-report questions official findings about chemical attack in Syrian Douma,” and “Douma chemical attack never

took place.” The article therefore contributes to the all-article human evaluation rather than the exact subset. Further anchors include the Afghanistan/Bagram article of 11 December 2019 and a Ukraine-war article of 24 May 2022 alleging that Russia acted to stop genocide.

Every experiment retrospectively searches the archive; none is a historical temporal holdout. Controlled narratives test recovery of a known document, while the source-linked articles test analyst-useful narrative alignment for all articles and reserve exact metrics for the bridge-resolved subset.

The raw original article receives the same top-three categorical retrieval judgment as each normalized query, which provides a no-normalization baseline even when no unique exact debunk exists. Exact Recall@k and MRR are used as secondary analyses for the 39 human-verified resolved articles.

Retrieval architecture and normalizer configuration

The author developed a modular pipeline that separates narrative generation from retrieval. The English and Ukrainian controlled evaluations compare model-generated broad narratives. The source-linked set adds the raw article as a direct query variant to measure whether normalization improves realistic article-to-debunk matching.

All three normalizers received the same task instruction: rewrite a debunk record or article into one broad narrative query that preserves the source-side manipulative stance, removes incidental names, dates, and numbers, and remains specific enough for retrieval. Gemini Flash 3.5 [19] used a JSON response schema. MamayLM-Gemma-3-4B [20] (INSAIT-Institute/MamayLM-Gemma-3-4B-IT-v1.0) and Qwen3-4B [21] (Qwen/Qwen3-4B) used greedy decoding at temperature 0.0, a 4,096-token input limit, and a 192-token generation limit. All reported Gemini operations use Gemini Flash 3.5.

All controlled comparisons use Qwen3-Embedding-0.6B for dense retrieval, a first-stage depth of 50, Qwen3-Reranker-0.6B for pairwise reranking, and a final top-10 list [21]. Table 3 derives from an earlier 5,235-query selection run on the same 7,538-record corpus. The table holds the queries fixed and compares the dense and reranked stages for the Qwen3 and BGE-M3 retrieval stacks; its values select the fixed downstream stack rather than estimate final normalizer performance.

Qwen3-Embedding-0.6B outperforms the BGE-M3 dense baseline and gains 8.9 percentage points in Recall@1 from cross-encoder reranking, compared with 2.8 points for BGE-M3. The selected Qwen3 stack is then held constant across all normalizer comparisons.

Table 3

Retriever and reranker selection

Configuration	R@1	R@3	MRR
BGE-M3 dense	0.555	0.714	0.626
BGE-M3 + BGE-Reranker-v2-m3	0.583	0.714	0.643
Qwen3-Embedding-0.6B dense	0.566	0.731	0.639
Qwen3-Embedding-0.6B + Qwen3-Reranker-0.6B	0.655	0.730	0.691

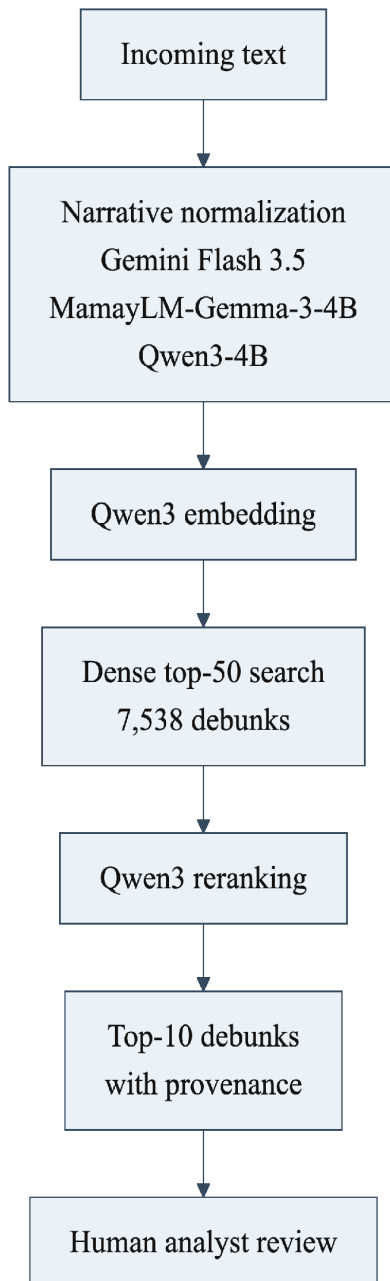


Fig. 1. Narrative-retrieval pipeline

Figure 2 summarizes the evaluation bridge between exact traceability and natural article utility. The two branches use different relevance definitions and must remain separate when interpreting results.

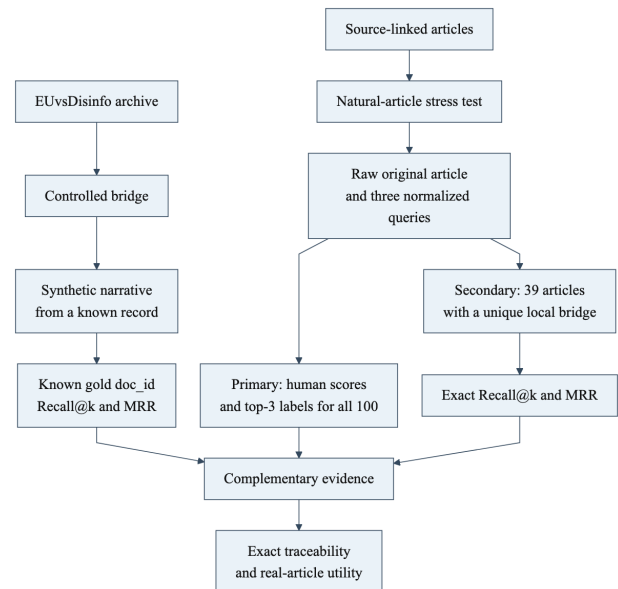


Fig. 2. Controlled and source-linked article evaluation paths

Performance measures and human evaluation

The exact-document evaluation reports Recall@1, Recall@3, Recall@5, Recall@10, and mean reciprocal rank. Recall@5 is the central metric because an analyst can inspect a short ranked list; Recall@1 provides the strictest traceability criterion, and MRR summarizes rank sensitivity. These exact-document definitions apply to the controlled experiments and the 39-article bridge subset. Pooled English-Ukrainian values micro-average 2,021 inputs. The English common-subset analysis uses paired exact McNemar tests for Recall@1 and paired bootstrap differences for MRR, followed by Holm correction.

Each source-linked article contributes one raw query and three normalized narratives. The normalized evaluation therefore contains 300 model rows. A human reviewer assigned the final rubric scores and retrieval labels for each evaluated row. Raw receives N/A for Narrative, Position, and Abstraction because it is not generated, but its top-three retrieval results receive the same correct, plausible, or wrong judgment as each normalized variant. This all-article baseline measures normalization utility even for the 61 articles without a unique exact bridge. To assess intra-rater reliability, the reviewer repeated the evaluation for a blinded, article-clustered audit sample covering 25 % of Original-100. Of the full dataset, 18 % was selected by thematic stratification to cover every thematic group represented in the evaluation, and 7 % was added as a purposive stress-test subset. This produced 75 paired ordinal judgments for each rubric dimension and 100 paired retrieval judgments. Agreement was estimated as exact agreement and linearly weighted Cohen's kappa for the ordinal scales;

retrieval also reports unweighted kappa. Confidence intervals use 10,000 stratified bootstrap resamples at the article level.

The author proposed four dimensions because narrative fidelity, stance, abstraction, and retrieval utility can fail independently. Table 4 presents the anchored scale used by the human reviewer. Scores are ordinal judgments, not probabilities or confidence values.

Table 4

Anchored human-evaluation rubric

Dimension	Why measured	Score 0	Score 1	Score 2
Narrative	Detect semantic drift	Central narrative lost or changed	Broad family retained with a defining shift	Defining narrative preserved
Position	Detect claim/debunk inversion	Corrective stance or inversion	Mixed or ambiguous source stance	Source-side manipulative stance retained
Abstraction	Test frame reusability	Too specific, noisy, or generic	Partial abstraction	Reusable frame; incidental details removed
Retrieval specificity	Test query effectiveness	Top three off-topic or broadly related	Right family with a key near-miss	Intended or directly equivalent debunk surfaced

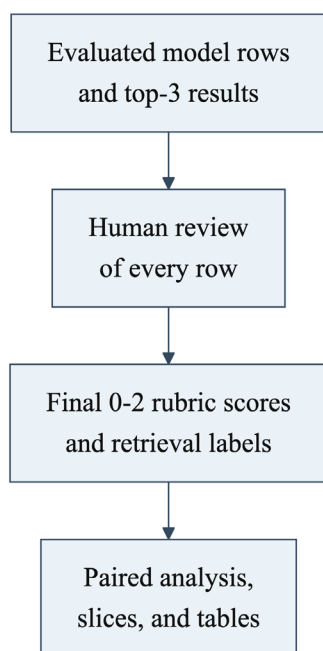


Fig. 3. Human review and paired-analysis workflow

Human retrieval labels cover all 100 articles and all four query variants. Correct requires a direct or equivalent match, plausible denotes an analyst-useful narrative-family match with a material near-miss, and wrong marks an off-topic or misleading result. Paired McNemar tests compare correct versus non-correct and analyst-usable (correct + plausible) versus wrong. Exact Recall@k and MRR use only bridge-verified targets and never treat a narrative-family label as an exact gold document.

Intra-rater reliability

In the blinded repeat-assessment sample covering 25 % of Original-100, Position showed 96,0 % exact agreement (95 % CI 92,0 %–100,0 %) and a linearly weighted kappa of 0.891 (0,598–1,000). Table 5 reports the estimates for all evaluated dimensions.

Across all 300 rubric decisions, exact agreement was 93,0 % (95 % CI 89,3 %–96,3 %). Across the 100 retrieval decisions, exact agreement was 86,0 % (79,0 %–92,0 %), unweighted kappa was 0,708 (0,570–0,833), and weighted kappa was 0,804 (0,699–0,892). The thematically stratified 18 % main subset reached 97,2 % rubric agreement and 91,7 % retrieval agreement; the deliberately difficult 7 % stress subset reached 82,1 % and 71,4 %, respectively. The lower stress-subset values are treated as a sensitivity result, not as evidence of inter-annotator reliability.

Table 5

Agreement in the blinded repeat assessment

Measure	Exact agreement, % (95 % CI)	Weighted κ (95 % CI)
Narrative	93,3 (89,3–97,3)	0,825 (0,672–0,922)
Position	96,0 (92,0–100,0)	0,891 (0,598–1,000)
Abstraction	92,0 (85,3–97,3)	0,809 (0,636–0,943)
Retrieval specificity	90,7 (84,0–96,0)	0,825 (0,696–0,924)
All rubric decisions	93,0 (89,3–96,3)	–
Retrieval label	86,0 (79,0–92,0)	0,804 (0,699–0,892)

RESEARCH RESULTS

Overall cross-language retrieval

Table 6 derives from the final reranked positions of the known source doc_id for 1,071 English and 950 Ukrainian controlled queries. Recall values count whether that identifier occurs within each rank cutoff, MRR uses its reciprocal rank, and the displayed pooled values micro-average all 2,021 queries. The result summarizes retrieval across the evaluated inputs; it does not estimate a causal language effect.

Table 6

Pooled controlled retrieval

Normalizer	R@1	R@3	R@5	R@10	MRR
Gemini Flash 3.5	0,727	0,872	0,901	0,932	0,805
MamayLM-Gemma-3-4B	0,724	0,857	0,898	0,929	0,798
Qwen3-4B	0,761	0,888	0,921	0,949	0,829

On the pooled benchmark, the reranked Qwen3-4B pipeline leads Recall@5 at 0.921, ahead of Gemini Flash 3.5 at 0.901 and MamayLM-Gemma-3-4B at 0.898. Qwen3-4B also leads Recall@1 and MRR. Qwen3-4B leads the English and pooled results; Gemini Flash 3.5 slightly leads Ukrainian Recall@5 and MRR.

Common-subset normalizer comparison (English, $n = 1,071$)

Table 7 uses the same 1,071 English doc_id values for every model row. BM25, dense, and reranked values are the ranks of the known source record under each query variant, which pairs normalizer differences and reranking gains directly on this subset.

Table 7

English controlled retrieval

Model	Stage	R@1	R@3	R@5	R@10	MRR
Gemini Flash 3.5	BM25	0,394	0,528	0,579	0,641	0,473
Gemini Flash 3.5	dense	0,549	0,687	0,745	0,819	0,633
Gemini Flash 3.5	rerank	0,679	0,826	0,861	0,898	0,760
MamayLM-Gemma-3-4B	dense	0,637	0,769	0,817	0,878	0,715
MamayLM-Gemma-3-4B	rerank	0,714	0,843	0,888	0,923	0,789
Qwen3-4B	BM25	0,510	0,653	0,704	0,766	0,594
Qwen3-4B	dense	0,615	0,764	0,811	0,863	0,700
Qwen3-4B	rerank	0,739	0,866	0,900	0,933	0,807

At Recall@5, the reranked Qwen3-4B pipeline leads with 0.900, ahead of MamayLM-Gemma-3-4B at 0.888 and Gemini Flash 3.5 at 0.861. The same ordering holds at Recall@1 and MRR. After Holm correction, Qwen3-4B's advantage over Gemini Flash 3.5 is significant for Recall@1 (McNemar exact, adjusted $p = 0.0009$) and MRR (paired bootstrap difference 0.048, 95 % CI [0.023, 0.071], adjusted $p < 0.001$). MamayLM-Gemma-3-4B's numerical advantage does not hold up after correction. Reranking improves every dense pipeline. BM25 remains a useful baseline for generated narratives, while raw and title queries are excluded from the

primary comparison because source-record overlap creates an identity shortcut.

Ukrainian controlled-input evaluation

Table 8 derives from 950 Ukrainian controlled inputs that share a known source doc_id. Each value uses the final reranked list and the same exact-document definitions as the English experiment.

Table 8

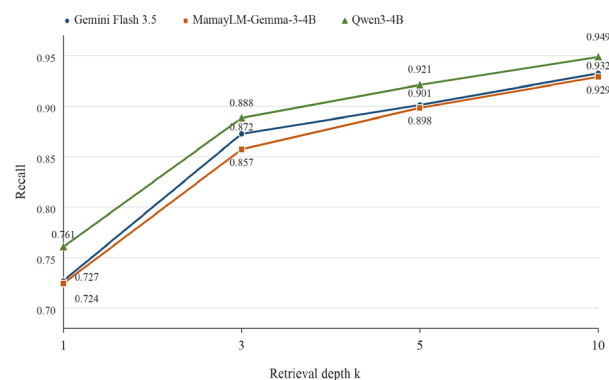
Ukrainian controlled retrieval

Normalizer	R@1	R@3	R@5	R@10	MRR
Gemini Flash 3.5	0,781	0,924	0,946	0,971	0,856
MamayLM-Gemma-3-4B	0,736	0,873	0,909	0,935	0,809
Qwen3-4B	0,785	0,913	0,944	0,966	0,854

The denominator is 950 shared Ukrainian inputs with known source doc_id. Gemini Flash 3.5 has the highest Recall@5 (0.946) and MRR (0.856), while Qwen3-4B has the highest Recall@1 (0.785 versus 0.781). MamayLM-Gemma-3-4B reaches Recall@5 = 0.909. This controlled result demonstrates robustness to Ukrainian input within the tested archive.

Retrieval depth and ranked-list utility

Operational triage rarely depends on a single top-ranked result. An analyst reviews a short list of candidates. The ranked-depth columns in the combined benchmark show how often the relevant debunk appears on that list, which is why Recall@5 is the study's central metric. For the Qwen3-4B reranked system on the combined benchmark, Recall rises from 0.761 at top-1 to 0.888 at top-3, 0.921 at top-5, and 0.949 at top-10. When rank 1 misses the exact match, the next few positions often recover it: reviewing three to five candidates surfaces the relevant record in more than nine of ten cases, as measured by Recall@5. This sustains the exact-document relevance criterion while supporting a ranked-list interface that presents a small set of candidate debunks rather than a single forced match.

Fig. 4. Controlled Recall by retrieval depth ($n = 2,021$)

Source-linked article match quality

The source-linked articles are full, noisy inputs rather than debunk-derived queries. Table 9 aggregates the final human-reviewed scores over 100 articles for each normalizer. Each cell reports the dimension mean followed by the counts assigned scores 0, 1, and 2; no composite score is calculated.

Table 9

Human-reviewed source-linked article rubric

Normalizer	Narr.	Pos.	Abstr.	Spec.
Gemini Flash 3.5	1,90 (1/8/91)	1,97 (0/3/97)	1,85 (0/15/85)	1,74 (1/24/75)
MamayLM-Gemma-3-4B	1,79 (2/17/81)	1,87 (2/9/89)	1,65 (4/27/69)	1,58 (3/36/61)
Qwen3-4B	1,70 (5/20/75)	1,80 (5/10/85)	1,68 (1/30/69)	1,60 (5/30/65)

Gemini Flash 3.5 has the highest mean on Narrative (1.90), Position (1.97), Abstraction (1.85), and Retrieval specificity (1.74). Relative to Gemini, MamayLM-Gemma-3-4B is lower by 5,8 %, 5,1 %, 10,8 %, and 9,2 %; Qwen3-4B is lower by 1,5 %, 8,6 %, 9,2 %, and 8,0 %. The local 4B models' mean scores lie within 11 % of Gemini across the four dimensions. Medians and the 0/1/2 distributions retain information hidden by means. Paired bootstrap intervals quantify per-article mean differences, while paired signed tests with Holm correction test directional differences without treating the scale as continuous. Gemini exceeds both local models on Narrative and Position after correction.

Human rubric means (ordinal 0-2)

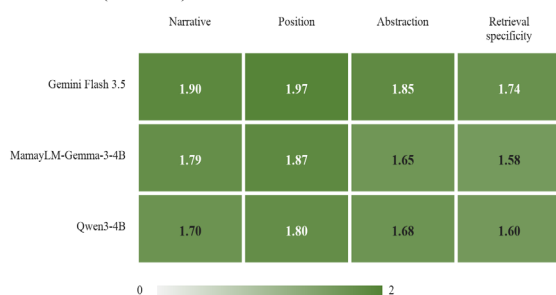
Fig. 5. Human rubric means for source-linked articles ($n = 100$)

Table 10 counts the final human top-three retrieval labels for all 100 articles under each of the four query variants. The denominator is fixed at 100 per variant, usable sums correct and plausible labels, and relevance is narrative-family usefulness rather than exact slug identity.

Gemini Flash 3.5 records 85 % correct and 95 % analyst-usable results. The correct rate for each local 4B model is 79 %, 7,1 % below Gemini, Qwen3-4B reaches 89 % usable, 6,3 % below Gemini, and MamayLM-Gemma-3-4B reaches 88 %, 7,4 % below Gemini.

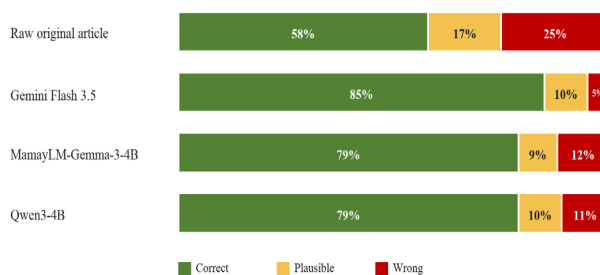
Table 10

Source-linked article retrieval labels

Query variant	Correct	Plaus.	Wrong	Usable
Raw original article	58,0 %	17,0 %	25,0 %	75,0 %
Gemini Flash 3.5	85,0 %	10,0 %	5,0 %	95,0 %
MamayLM-Gemma-3-4B	79,0 %	9,0 %	12,0 %	88,0 %
Qwen3-4B	79,0 %	10,0 %	11,0 %	89,0 %

Raw original-article retrieval reaches 58 % correct and 75 % usable. The normalized variants improve both outcomes over raw after Holm-corrected paired McNemar tests, while the differences among normalizers are smaller. In this evaluation, the local models' correct and usable rates differ from Gemini by no more than 7,4 % in relative terms. Gemini retains the strongest point estimate. These tests concern human top-three labels rather than exact-document accuracy.

Human top-3 retrieval labels

Fig. 6. Human top-three retrieval labels for source-linked articles ($n = 100$)

Secondary exact-document coverage audit

Thirty-nine of the 100 articles have one human-verified exact local debunk. Table 11 derives Recall@k and MRR from the rank of that target in the raw, normalized, and reciprocal-rank-fused lists. The remaining 61 articles remain in every primary human analysis above; they are excluded only from this exact-document audit.

Raw original-article retrieval leads the exact subset at Recall@5 = 0,744 and MRR = 0,609. Among normalized queries, Qwen3-4B matches Gemini Flash 3.5 at Recall@5 and differs by +3,8 % in MRR; its metric differences range from -7,4 % to +8,1 %. MamayLM-Gemma-3-4B differs by -1,0 % in MRR but by -21,7 % at Recall@5. Reciprocal-rank fusion combines complementary raw and normalized lists: raw + Gemini reaches Recall@5 = 0,795 and MRR = 0,578, while raw + Qwen3-4B reaches the highest Recall@1 at 0,487. Qwen3-4B has numerically close point estimates to Gemini on several measures in this subset. The wider MamayLM-Gemma-3-4B Recall@5 gap and the small, coverage-selected denominator require caution.

Table 11

Exact-document retrieval on bridge-resolved articles

Query variant	R1	R3	R5	R10	MRR
Raw original article	0,513	0,718	0,744	0,744	0,609
Gemini Flash 3.5	0,308	0,513	0,590	0,692	0,421
MamayLM-Gemma-3-4B	0,333	0,462	0,462	0,590	0,417
Qwen3-4B	0,333	0,538	0,590	0,641	0,437
Raw + Gemini RRF	0,410	0,718	0,795	0,821	0,578
Raw + Mamay RRF	0,385	0,615	0,667	0,769	0,509
Raw + Qwen RRF	0,487	0,692	0,692	0,718	0,574

Coverage, language, date, and event slices are reported descriptively. The source-linked set contains 24 articles from 2015-2018, 65 from 2019-2021, and 11 from 2022-2023. Keyword-based event grouping identifies 33 Ukraine-related, 23 NATO-Russia, 13 Syria, 7 COVID-19, 1 Afghanistan, and 1 MH17 articles, with overlaps possible. Sparse slices do not support significance claims.

Source-linked article error analysis

The human review identified four component-level failure mechanisms: proposition drift, stance inversion, insufficient abstraction, and retrieval near-miss. Table 12 counts final-score-0 judgments across 100 narratives per model and counts each row once in Any 0. It excludes score-1 degradations, so the total does not reflect the full model error rate.

Table 12

Severe source-linked article failure signals

Normalizer	Narr. 0	Pos. 0	Abstr. 0	Spec. 0	Any 0
Gemini Flash 3.5	1	0	0	1	2
MamayLM-Gemma-3-4B	2	2	4	3	9
Qwen3-4B	5	5	1	5	10

Gemini Flash 3.5 has two narratives with at least one zero, compared with nine for MamayLM-Gemma-3-4B and ten for Qwen3-4B. Expressed relative to Gemini, these counts are 350 % and 400 % higher, but the small Gemini denominator makes the percentages unstable; the absolute differences of 7 and 8 cases are more informative. Although several aggregate retrieval and rubric point estimates are numerically close, Table 12 shows a higher tail risk of severe component failure for the local models. The counts are paired over 100 articles per model and do not measure operational harm outside this source-linked set.

Proposition drift can coexist with coherent retrieval. The article “Tension between China and the

United States over Syria” (title translated from Arabic) concerns a dispute between China and the United States at the United Nations over cross-border aid to Syria. Gemini Flash 3.5 replaces it with a claim that Western sanctions obstruct pandemic assistance. The reviewer assigns Narrative = 0 but Position, Abstraction, and Retrieval specificity = 2. The top three results match the substituted sanctions query, so strong retrieval cannot compensate for a misrepresented source.

Stance inversion creates a more consequential generation failure. In “Russia held ‘responsible’ for downing MH17 by Australia & Netherlands, Moscow questions probe,” Qwen3-4B converts a source that questions Russia's responsibility into the assertion that Russia downed the aircraft. The final scores are Narrative = 0, Position = 0, Abstraction = 2, and Retrieval specificity = 2. Relevant MH17 results do not make the query faithful to the article.

The other two mechanisms isolate abstraction and retrieval. For “A Novichok developer: Navalny photographs refute claims that he was poisoned with the substance” (title translated from Arabic), MamayLM-Gemma-3-4B retains the named expert, symptoms, and incident detail; the reviewer assigns Abstraction = 0 while the top three remain correct. Conversely, its query for “Zelensky simplified Ukrainian citizenship for punitive forces and Russian liberals” (title translated from Russian) receives Narrative = 2, Position = 2, and Abstraction = 2, but Retrieval specificity = 0 because the top three concern Donbas and territorial control rather than citizenship or the alleged beneficiaries. The first case calls for a stronger normalization constraint; the second calls for retriever or corpus improvement.

The categorical retrieval labels capture a wider set of problems than score 0. Gemini Flash 3.5, MamayLM-Gemma-3-4B, and Qwen3-4B have 5, 12, and 11 incorrect labels, respectively, while only 1, 3, and 5 rows have Retrieval specificity = 0. Most remaining wrong results score 1 because they reach the right topic or narrative family but miss a defining actor, mechanism, or proposition. Error review should therefore retain both the anchored rubric and the categorical label rather than collapsing them into one error count.

Experimental environment

GPU normalization, dense retrieval, and reranking runs were executed on RunPod instances equipped with NVIDIA RTX 4090 GPUs. Aggregation, statistical analysis, document preparation, and quality-control scripts ran on macOS. Gemini Flash 3.5 was accessed through the Gemini API. The study evaluates retrieval quality rather than hardware throughput.

Discussion of results

Qwen3-4B leads pooled controlled exact retrieval, while Gemini Flash 3.5 leads the four human dimensions and the correct top-three rate on source-linked articles. The local 4B models are numerically close to Gemini on several aggregate metrics within the tested architecture: pooled retrieval differs by no more than 4,7 %, Qwen3-4B stays within 1,2 % on every Ukrainian metric, and the all-article correct rates are 7,1 % lower. Human rubric gaps reach 11 %, and severe zero-score cases are more frequent for the local models. Numerical proximity on selected point estimates does not establish equivalence, non-inferiority, or interchangeability. Controlled queries reward recovery of one *doc_id*; natural articles reward faithful compression and analyst-useful family matching under incomplete coverage.

Limitations

The corpus is a closed EUvsDisinfo archive with finite editorial, linguistic, and temporal coverage. Exact retrieval cannot detect an unknown narrative, and a semantically similar archived case can lead to false context anchoring. Safe deployment requires corpus-date display, source links, provenance for score and rank, duplicate controls, a no-match/abstain path, and human confirmation before a candidate is treated as evidence.

The controlled queries originate from records in the search corpus, so raw or title-like wording can leak identity and inflate lexical baselines. The Ukrainian set tests controlled cross-language input rather than live social media content. Original-100 also comes from the wider EUvsDisinfo ecosystem [18], and its 2015–2023 dates do not constitute a temporal holdout relative to the 2015–2026 corpus. The 39-case exact bridge is coverage-selected and small, whereas all 100 human labels reflect a single reviewer's interpretation of narrative-family relevance.

The blinded repeat assessment covering 25 % of Original-100 indicated strong intra-rater consistency; agreement between independent reviewers falls outside the present design.

Taken together, these constraints define the scope of the findings as retrospective retrieval within the tested EUvsDisinfo setting. Within this scope, the reported confidence intervals characterize record- or article-level variation under the evaluated protocol. Causal language comparisons and performance of standalone normalizers or live monitoring streams fall outside the present design. Transferability to other languages, publishers, platforms, topics, time periods, and open-web retrieval settings remains a target for external validation using independent

corpora, time-separated data, matched multilingual inputs, and multiple reviewers.

Reproducibility

All experimental code is available at <https://github.com/olehmell/rag-narrative-normalizer-paper>. It contains manifests for the common subset (1,071 *doc_ids*), the Ukrainian controlled subset (950 *doc_ids*), and Original-100 (article identifiers, source URLs, source references, and full texts). It also includes CSV aggregates, retrieval traces, the final result tables, and the provenance of the Original-100 labeling. Python utilities are executed with *uv*, and the package lock is retained. Exact prompt text, decoding parameters, and model revisions are available in the repository.

CONCLUSIONS AND PROSPECTS FOR FURTHER RESEARCH

This study evaluated narrative normalization on 2,021 controlled cross-language inputs and 100 source-linked articles. Qwen3-4B produces the best pooled controlled exact-retrieval result at *Recall@5* = 0,921. For the source-linked articles, Gemini Flash 3.5 obtains the highest human Narrative mean (1,90/2) and correct top-3 rate (85 %); MamayLM-Gemma-3-4B and Qwen3-4B each reach 79 % correct. The raw original article reaches 58%, so the all-100 comparison shows the practical gain from normalization. Thirty-nine articles also support a secondary exact-document coverage audit. In the blinded repeat assessment covering 25 % of Original-100, exact agreement was 93,0 % for rubric decisions and 86,0 % for retrieval decisions.

Local four-billion-parameter language models produced retrieval point estimates close to Gemini Flash 3.5 on several tested benchmarks within this architecture. Human rubric gaps and severe failures still favor Gemini, so deployment should retain model-specific monitoring and human verification. The evidence does not support autonomous disinformation verdicts or generalization beyond English and Ukrainian.

The narrative-retrieval pipeline is intended as one component of a larger multi-agent disinformation-monitoring system built from small language models [22]. That system decomposes analysis into three specialized loops: a manipulation-screening loop that estimates rhetorical risk, a narrative-intentional loop that reconstructs a message as a canonical narrative and a structured bundle of checkable claims, and an evidence-verification loop that performs retrieval-grounded checking with explicit provenance, all coordinated by an orchestration, evidence-memory, and audit layer. The normalization studied here operationalizes the narrative-canonicalization function of the second loop: rewriting incoming text into a broad narrative turns noisy article text into a retrieval-ready query, and matching that narrative against documented debunks supplies a contextualization

signal for verification. An isolated evaluation is necessary because normalization or retrieval errors would otherwise propagate into claim formation and final verdicts.

Future work should introduce negative and no-match examples, calibrate rejection thresholds, and test live monitoring streams. The next system-level test should feed retrieved debunks into claim-bundle formation and the evidence-verification loop, share an evidence board and audit log across loops, and defer absent or low-confidence matches to analyst review. A validated ranked-list interface is the most immediate extension.

REFERENCES

- [1] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *Advances in Neural Information Processing Systems*, vol. 33, 2020, <https://doi.org/10.48550/arXiv.2005.11401>.
- [2] P. Santra, M. Ghosh, D. Ganguly, P. Basuchowdhuri, and S. K. Naskar, “The curious case of contexts in retrieval-augmented generation with a combination of labeled and unlabeled data,” *WIREs Data Mining and Knowledge Discovery*, vol. 15, no. 2, 2025, <https://doi.org/10.1002/widm.70021>.
- [3] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER: A large-scale dataset for fact extraction and verification,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, <https://doi.org/10.18653/v1/N18-1074>.
- [4] S. Shaar, N. Babulkov, G. Da San Martino, and P. Nakov, “That is a known lie: Detecting previously fact-checked claims,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, <https://doi.org/10.18653/v1/2020.acl-main.332>.
- [5] M. Pikuliak, I. Srba, R. Moro, T. Hromadka, T. Smolen, M. Melisek, I. Vykopal, J. Simko, and M. Bielikova, “Multilingual previously fact-checked claim retrieval,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, <https://doi.org/10.18653/v1/2023.emnlp-main.1027>.
- [6] H. Peng et al., “SemEval-2025 Task 7: Multilingual and Crosslingual Fact-Checked Claim Retrieval,” in *Proceedings of the 19th International Workshop on Semantic Evaluation*, 2025, <https://aclanthology.org/2025.semeval-1.323/>.
- [7] G. Da San Martino, A. Barron-Cedeno, H. Wachsmuth, R. Petrov, and P. Nakov, “SemEval-2020 Task 11: Detection of propaganda techniques in news articles,” in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020, <https://doi.org/10.18653/v1/2020.semeval-1.186>.
- [8] J. Piskorski, N. Stefanovitch, G. Da San Martino, and P. Nakov, “SemEval-2023 Task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup,” in *Proceedings of the 17th International Workshop on Semantic Evaluation*, 2023, <https://doi.org/10.18653/v1/2023.semeval-1.317>.
- [9] K. L. Anglin, A. Bertrand, J. Gottlieb, and J. Elefante, “Scaling Up With Integrity: Valid and Efficient Narrative Policy Framework Analyses Using Large Language Models,” *Policy Studies Journal*, vol. 54, no. 2, 2026, <https://doi.org/10.1111/psj.70045>.
- [10] D. Korencic, B. Chulvi, X. B. Casals, A. Toselli, M. Taule, and P. Rosso, “What distinguishes conspiracy from critical narratives? A computational analysis of oppositional discourse,” *Expert Systems*, vol. 41, no. 11, 2024, <https://doi.org/10.1111/exsy.13671>.
- [11] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” *Foundations and Trends in Information Retrieval*, vol. 4, nos. 1–2, pp. 1–174, 2009, <https://doi.org/10.1561/15000000019>.
- [12] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation,” *arXiv*, 2024, <https://doi.org/10.48550/arXiv.2402.03216>.
- [13] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, “Dense passage retrieval for open-domain question answering,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, <https://doi.org/10.18653/v1/2020.emnlp-main.550>.
- [14] O. Khattab and M. Zaharia, “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT,” in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, <https://doi.org/10.1145/3397271.3401075>.
- [15] G. Lin, “Using cross-encoders to measure the similarity of short texts in political science,” *American Journal of Political Science*, vol. 69, no. 4, pp. 1600–1616, 2025, <https://doi.org/10.1111/ajps.12956>.
- [16] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, <https://doi.org/10.18653/v1/D19-1410>.
- [17] European External Action Service, “EUvsDisinfo: EEAS East StratCom Task Force disinformation database,” 2024. <https://euvsdisinfo.eu/> (accessed Jun. 22, 2026).
- [18] J. A. Leite, O. Razuvayevskaya, K. Bontcheva, and C. Scarton, “EUvsDisinfo: A Dataset for Multilingual Detection of Pro-Kremlin Disinformation in News Articles,” in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, <https://doi.org/10.1145/3627673.3679167>.
- [19] Google DeepMind, “Gemini Flash 3.5 technical documentation,” 2025. <https://deepmind.google/>

- technologies/gemini/flash/ (accessed Jun. 22, 2026).
- [20] INSAIT Institute, “MamayLM v1.0 release blog,” Hugging Face, 2025. <https://huggingface.co/spaces/INSAIT-Institute/mamaylm-v1-blog> (accessed Jun. 22, 2026).
- [21] Qwen Team, “Qwen3 technical report,” arXiv, 2025, <https://doi.org/10.48550/arXiv.2505.09388>.
- [22] O. V. Melnychuk, “Methodology of a multi-agent system for detection of manipulations and disinformation” [in Ukrainian], Problems of Informatization and Management, 2025, <https://doi.org/10.18372/2073-4751.84.20898>

Мельничук О. В.

МАЛІ МОВНІ МОДЕЛІ ДЛЯ БАГАТОМОВНОГО ПОШУКУ ДЕЗІНФОРМАЦІЇ

Дезінформаційні кампанії повторюють відомі наративи, змінюючи мову, учасників подій і контекст публікації. Через такі зміни аналітику складно пов'язати новий матеріал із випадками, які вже описали фактчекери. Для розв'язання цієї задачі було розроблено програму багатомовної нормалізації наративів і пошуку документованих спростувань у корпусі EUvsDisinfo. Пошуковий корпус містить 7 538 фактчекінгових записів за 2015–2026 роки. Кожен запис поєднує дезінформаційне твердження зі спростуванням або контекстуальним поясненням. Експеримент охопив матеріали, що стосуються війни Росії проти України, відносин Росії з НАТО, Сирії та COVID-19. У цих даних ті самі смислові схеми з'являються в різних мовах і подіях. Мовна модель перетворює вхідний текст на стислий наративний запит. Вона зберігає основне твердження та позицію джерела, водночас вилучаючи імена, дати й інші ситуативні деталі. Система використовує сформований запит для семантичного пошуку та повторного ранжування записів. Цей модуль може працювати в складі системи моніторингу дезінформації: він пов'яже новий матеріал із відомими наративами та передає знайдені спростування в контур перевірки доказів. У центрі дослідження – можливість комерційної великої мовної моделі Gemini Flash 3.5 із локальними моделями Qwen3-4B і MamayLM-Gemma-3-4B. Отримані результати підтверджують, що малі мовні моделі можуть виконувати нормалізацію наративів у багатомовній пошуковій системі з помірною втратою якості порівняно з комерційною моделлю. Локальне розгортання дає змогу контролювати дані та здійснювати донавчання моделей для покращення якості та стабільності.

Ключові слова: штучний інтелект; RAG; дезінформація; малі мовні моделі; великі мовні моделі; багатомовний пошук; наратив; семантичний пошук; ранжування; інформаційні операції;

Melnichuk O.

SMALL LANGUAGE MODELS FOR MULTILINGUAL DISINFORMATION RETRIEVAL

Disinformation campaigns reuse known narratives while changing the language, actors, and publication context. Such variation makes it difficult for analysts to connect new material with cases already documented by fact-checkers. To address this task, a software application was developed for multilingual narrative normalization and the retrieval of documented debunks from the EUvsDisinfo corpus. The retrieval corpus contains 7,538 fact-checking records published between 2015 and 2026. Each record links a disinformation claim to its debunk or a contextual explanation. The experiment covered materials concerning Russia's war against Ukraine, Russia–NATO relations, Syria, and COVID-19. Across these data, the same semantic patterns recur in different languages and event contexts. The language model converts an input text into a concise narrative query. It retains the main claim and the source's position while removing names, dates, and other contextual details. The system uses the resulting query for semantic retrieval and corpus record reranking. This module can operate as part of a disinformation-monitoring system: it connects new material with known narratives and passes the retrieved debunks to an evidence-verification component. The study compares the capabilities of the commercial large language model Gemini Flash 3.5 with the locally deployed Qwen3-4B and MamayLM-Gemma-3-4B models. The results confirm that small language models can perform narrative normalization in a multilingual retrieval system with a moderate reduction in quality compared with the commercial model. Local deployment provides control over data and enables fine-tuning to improve model quality and stability.

Keywords: artificial intelligence; RAG; disinformation; small language models; large language models; multilingual retrieval; narrative; semantic retrieval; reranking; information operations.

Received: 14.07.2026 p.

Accepted: 17.08.2026 p.

Published: 10.09.2026 p.