

Article

Harnessing LLM Ensembles for KG-Grounded Narrative Extraction: Disinformation vs. Trustworthy News

Justina Mandravickaitė ^{*,†}  and Tomas Krilavičius [†] 

Faculty of Informatics, Vytautas Magnus University, 53361 Akademija, Kaunas District, Lithuania; tomas.krilavicius@vdu.lt

* Correspondence: justina.mandravickaite@vdu.lt

† These authors contributed equally to this work.

Abstract

Due to the rapid spread of disinformation, it is becoming increasingly difficult for the public to understand current events and how discussions and decisions are made in democratic societies. We propose a KG-grounded narrative extraction pipeline to compare disinformation and trustworthy news. English articles (2015–2023), included in EUvsDisinfo cases and matched mainstream coverage, were converted to AMR-based RDF graphs, and LLM ensembles were used to extract characters, events, causal links and framing edges grounded in these graphs. We studied two ensemble policies: a recall-oriented union that retained all model outputs and a precision-oriented consensus that kept only agreed elements, plus an LLM critic that flagged missing links, contradictions and framing inconsistencies. On an expert-annotated subset of 60 articles, the extractor ensemble attained very high precision for characters (0.99) and events (0.97) and solid performance for causal links (0.77) and framing edges (0.84), with similar scores for both classes. Our critic ensemble reached 0.74 precision. Structurally, union and consensus operated over the same grounded nodes but differed significantly in relational density, thus achieving rich vs. skeletal narrative graphs. Linking our narratives to GDELT showed that $\approx 97\%$ of extracted actors and events appeared in global news for both classes, while directional actor pairs from causal links were less often supported for disinformation (0.45) than trustworthy news (0.60). Overall, disinformation and trustworthy articles shared event backbones but diverged in the density and (to a lesser extent) directionality of causal attributions and framing relations.

Keywords: knowledge graphs; AMR; narrative extraction; causal reasoning; framing; LLM ensemble; disinformation



Academic Editors: Alaa Mohasseb and Andreas Kanavos

Received: 27 December 2025

Revised: 11 February 2026

Accepted: 12 February 2026

Published: 16 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The accelerating spread of disinformation makes it increasingly difficult for the public to interpret current events and how discussions and decisions are made in democratic societies [1,2]. As online platforms work in short news cycles, the amount of unverified claims increases. Therefore, readers have to navigate between different narratives, which range from carefully written journalistic articles to deliberately misleading stories [3,4]. In addition to surface inaccuracies, disinformation often changes the structure of the narrative in order to alter perceptions of who the actors are, what events are highlighted, what causal explanations are offered and how selected semantic frames assign responsibility or intent [5–8]. Therefore, automatic detection and analysis of disinformation could benefit from reliable methods for extracting and reproducing narrative elements that are also accessible for practical use [9–11].

Recent studies on large language models (LLMs) have improved information extraction (IE) in terms of events, entities, and relations, but systems that use them still have problems with factual grounding, schema consistency, and cross-document coherence [12–14]. Knowledge graphs (KG) offer an additional structure, as they encode objects, roles, and relationships with clear semantics, which allows consistency checks and linking between different articles [15–17]. However, robust alignment of free-form model outputs to a KG schema remains a challenging task, especially when narratives are heterogeneous and span multiple sources over time [18–20].

Empirically, ensembles and collaborative multi-model systems tend to improve accuracy and robustness by combining models with different inductive biases and failure modes, echoing classic ensemble learning results and recent work on LLM ensembles and compound inference systems [21–26]. For structured extraction in particular, multi-agent and mixture-of-agents frameworks based on several small or medium models have been shown to outperform single-model baselines while remaining cost-efficient and reliable [27].

This study examined whether combining KG grounding with LLM mini-ensembles, and separating extraction and formatting (robust schema-compliant JSON output) roles could improve narrative extraction from news, focusing on disinformation and trustworthy reporting. This separation of roles is consistent with the broader trend towards compound LLM systems, where specialized components handle different stages of the pipeline [26]. For that, we focused on four core narrative structures: characters/actors, events, causal links, and framing. Our KG schema integrated abstract meaning representation (AMR) [28] parses and FRED (a tool for transforming natural language text into machine-interpretable KGs) [29] outputs to derive RDF/OWL semantics, which support denser, more coherent graphs. We further introduce a “critic” component to identify missing links, contradictions and inconsistencies in the extractions, serving as an exploratory diagnostic layer for estimating possible headroom and finding gaps.

To guide the design and evaluation, we formulated the following research questions:

- RQ1: How does grounding narrative extraction in article-level KGs influence the quality, coherence and stability of extracted core narrative elements, i.e., characters and events?
- RQ2: How do different LLM ensemble policies (union vs. consensus), together with a critic component, impact precision–recall trade-offs when extracting more complex relations like causality and framing?
- RQ3: To what extent do the narrative representations of disinformation and trustworthy news differ in structure, particularly in the density and directionality of causal and framing relations?

Overall, our study supported a practical recipe for extracting and analyzing narrative structures that achieved competitive accuracy, reasonable schema alignment and richer graph structure at manageable computational cost. This, in turn, enabled analysis and comparison of how narratives are constructed in disinformation and trustworthy news.

The rest of the paper is structured as follows: Section 2 briefly introduces related work, Section 3 presents materials and methods used in the study, Section 4 reports the results and Sections 5 and 6 ends the paper with discussion and conclusions.

2. Related Work

In this section, we briefly introduce prior research relevant to the methodological design of our study. It is organized into five thematic categories: LLM-based narrative and information extraction, KG-grounded extraction and knowledge-enhanced reasoning, LLM ensembles and aggregation strategies, causality and framing extraction from news, and LLM-as-judge and post-hoc critic frameworks. These categories cover the core components

integrated in our proposed pipeline. We conclude the section with a structured synthesis and gap analysis that positions our contribution in relation to existing approaches.

2.1. LLM-Based Narrative Extraction

Large language models (LLMs) have shifted information extraction (IE) toward a generative paradigm, which includes entities, relations and events. Surveys have reviewed prompting and in-context learning strategies as well as evaluation practices and challenges related to domain transfer and robustness [30,31]. Comparisons between fine-tuned models and prompted LLMs in semantic event extraction indicate complementary strengths as LLMs perform better when labeled data are scarce, while fine-tuned models provide greater stability and tighter control [32].

2.2. KG-Grounded Narrative and Information Extraction

Recent research has explored the links between narrative extraction, disinformation detection, and the use of knowledge graphs (KGs) in combination with large language models (LLMs), motivating the methodological choices made in this study. In particular, the notion of “KG structure as prompt” incorporates KG topology directly into prompts for small language models and has been shown to improve causal reasoning under few-shot settings [33]. Comparative studies of KG-augmented LLMs and classical NLP pipelines on complex narrative structures, such as podcast transcripts, demonstrate that grounding model outputs in structured external knowledge improves robustness and factual accuracy [34]. Related work using KG embeddings further highlights their potential for enriching causal links between news events [35].

More broadly, recent studies show that KGs can reduce hallucinations and improve traceability through pre-, during-, and post-generation integration strategies, search augmentation, and constrained decoding [36,37]. Graph-based approaches likewise indicate that both generic graphs and semantic KGs strengthen fake news detection, with GNN-based models consistently confirming the value of graph-structured evidence [38]. KG-fusion studies further report accuracy gains by combining content features with evidence-based knowledge [39]. Additional LLMs and misinformation analyses highlight the dual role of LLMs and motivate explicit grounding and provenance tracing as key design considerations [40].

2.3. LLM Ensembles and Aggregation Strategies

Beyond grounding, another line of work improves extraction robustness via multi-model ensembles and aggregation strategies. Narrative-focused systems demonstrated that small-model ensembles can surpass baselines on framing-oriented subtasks in multilingual news [41]. Recent systems explicitly used ensembles of multiple small LLMs for narrative-focused tasks such as entity framing in SemEval-2025, where Llama 3.1–8B, Mistral 7B, Phi 4, and Gemma 2 (9B) were combined via a meta-classifier to produce consensus role labels, which outperformed the shared-task baseline across languages [41]. Another SemEval-2025 team evaluated several instruction-tuned LLMs and adopted a simple hard-voting scheme over the top-3 models for final predictions, achieving top ranks on multiple language tracks [42]. Beyond news, a combination of an ensemble of smaller LLMs, RAG pipeline and a moderator model improved inductive narrative coding for Holocaust diaries, illustrating ensemble benefits for sensitive narrative content [43].

2.4. Causality and Framing Extraction from News

Several works have studied how narrative structure, particularly causal relations, can be extracted from unstructured text. For example, causal graph have been extracted from news using event embeddings and ensemble-based causal inference across large temporal

datasets [44], as well as approaches that combine BERT-based causal relation extraction with clustering to construct interpretable causal KGs from financial news [45]. Researchers have also used BERT models and graph neural networks (GNNs) to construct narrative graphs by compressing temporal event-centric KGs into coherent storylines [46]. More recently, generating causal graphs with RoBERTa in combination with linguistic features has been shown to outperform large general-purpose LLMs such as GPT-4 and Claude 3.5 [47].

In parallel, framing has been studied as a key narrative mechanism in news and disinformation. An ensemble-based approach that combined LLMs, static embeddings and commonsense KGs has shown that this combination can improve multilingual frame detection [48]. Also, analyses of multilingual disinformation campaigns further revealed that framing strategies are deliberately adapted to target audiences and that existing framing models often struggle to capture such variations [1]. Recent LLM-based framing studies suggest that while these models can identify frame-related signals, they may overestimate affective or stylistic cues depending on explanations or external constraints [49].

Causality and framing signals are frequently integrated into broader fake news detection pipelines. Several studies have shown that ensemble models that integrated diverse feature extraction and selection methods achieve high accuracy across multiple datasets [50]. These approaches have also been extended to multi-layered systems for bilingual fake news detection in textual, numerical and multimedia forms [51]. Other works explored how KGs can make fake news detection models more scalable and easier to interpret [52]. For example, combining KGs with graph convolutional networks (GCNs) for fake news detection in Vietnamese performed well, even with limited annotated data [53].

2.5. LLM-as-Judge and Post-Hoc Critic Frameworks

An increasing number of research works separate the responsibilities of LLMs, where one model is used for structure extraction and another for formatting or canonization of extracted data according to a target schema or formal language. For example, the EDC framework [54] operationalized this separation in three stages and reported gains in KG construction when extraction was decoupled from normalization. Similarly, multi-phase pipelines that separate raw extraction from schema-constrained normalization or canonicalization have shown consistent benefits for reliability and scalability [54].

Additional evidence shows that a second LLM acting as a post-hoc corrector or formatter can systematically revise and normalize first-pass outputs without retraining [55]. Domain studies further treat schema-conforming rendering as a distinct step, for example, for value and unit normalization, reinforcing the extractor–formatter division as a robust design pattern [56,57].

2.6. Related Work Synthesis and Gap Analysis

As shown in Table 1, existing works have explored individual aspects of narrative extraction, such as KG grounding, causal modeling, framing detection or LLM ensembles. However, these components were typically studied in isolation, as reviewed in the preceding subsections. Few approaches integrated structured KG grounding with role-specialized LLM ensembles and explicit aggregation policies, and even fewer studies included a dedicated component to assess missing links and inconsistencies.

These trends suggest that our approach may be timely and well grounded. Incorporating KG grounding for characters, events, causal links and frames across disinformation versus trustworthy news offers a promising way to move beyond isolated extraction tasks. Our small, role-specialized LLM ensembles with a two-stage split (extractors and formatter), grounded in KGs for verifiability, align well with a small but growing body of related works.

Table 1. Comparison between prior work and this study. Source: authors' contribution.

Aspect	Prior Work	This Work
KG grounding	Partial/task-specific	AMR + FRED-based RDF grounding
Narrative scope	Single elements	Characters, events, causality, framing
LLM strategy	Single model	Role-specialized mini-ensembles
Aggregation	Implicit/ad hoc	Explicit union vs. consensus policies
Critic/validation	Rare or implicit	Dedicated critic component
Evaluation	Intrinsic or task-specific	Expert annotation + GDELT alignment
Disinformation vs. trustworthy news comparison	Limited	Structural narrative comparison

Moreover, comparative analysis of narrative structures between disinformation and trustworthy news remains underexplored. This study addresses these gaps by combining KG-grounded narrative extraction, ensemble-based inference and multi-level evaluation within a single pipeline.

3. Materials and Methods

This section presents the empirical foundation and methodological design of the study. It is divided into two main parts. Section 3.1 presents the materials, including dataset construction, computational environment, preprocessing decisions and knowledge graph construction. Section 3.2 details the KG-grounded narrative extraction pipeline, including choice of models and their configurations, prompting strategy, ensemble aggregation, critic component and evaluation procedures.

3.1. Materials

We describe the data and computational environment underlying the study in this section. Firstly, we introduce the base dataset and the reconstruction of the final news article corpus (Sections 3.1.1 and 3.1.2). Then, this section introduces the hardware setup and software environment to ensure reproducibility (Section 3.1.3). As we did not use the full text of news articles but their extractive summaries, Section 3.1.4 explains the summarization procedure. Finally, Section 3.1.5 presents the construction of article-level knowledge graphs.

3.1.1. Dataset

For this study, we used a part of the dataset for multilingual detection of pro-Kremlin disinformation in news articles [58]. The full dataset consists of 18,249 articles in 42 languages published January 2015–July 2023. As the full texts of these news articles were not publicly available, we used Diffbot API (free for academic purposes) [59] to reconstruct the dataset from URL links, which were publicly available.

We selected only the part of the dataset that contained English-language articles, i.e., a total of 6546 articles (425 disinformation and 6121 trustworthy). Some articles were no longer available or have been modified. So, after filtering, cleaning, and selecting only articles with similar topics in disinformation and trustworthy news classes (NATO-

Russian relations, Russia–Ukraine conflict, Syrian civil war, geopolitical tensions, disinformation/propaganda narratives, conspiracy theories, etc.), the final dataset used in this study was made of 308 disinformation news articles and 302—trustworthy news articles. Labeling news articles as “disinformation” and “trustworthy” was based on the original dataset, i.e., we reused article labels as they were assigned in the original dataset [58].

Table 2 presents article distribution per year in the final dataset. For trustworthy news articles, 2020 had the largest number of articles (109), while for disinformation the articles were distributed more equally, with 2019 and 2020 having the largest numbers (46 each). As for publishers, for disinformation the main publisher was RT (166 articles) and for trustworthy news the articles were less dominated by a single outlet, with BBC having the largest number of articles (28) (Table 3).

Table 2. Article distribution 2015–2023. Reproduced from [60].

Year	Disinformation	Trustworthy News
2015	19	3
2016	26	9
2017	29	14
2018	43	24
2019	46	52
2020	46	109
2021	28	22
2022	43	52
2023	28	17

Table 3. Publishers with the largest number of articles (number in brackets) in the final dataset. Reproduced from [60].

Publishers	
Disinformation	Trustworthy News
RT (166)	BBC (28)
TASS (51)	Polygraph.info (25)
Sputnik (15)	The Guardian (24)
Geopolitica.ru (12)	Radio Free Europe (11)
Global research (11)	DW.com (10)

For each news article in the dataset, publicly available metadata fields included basic bibliographic information (title, author, article_publisher, article_domain, article_url), linguistic and temporal attributes (article_language, date of publication), and variables specific to our disinformation vs. trustworthy news setting (debunk_date, class). The keywords field contained topical descriptors (e.g., “Ukraine”, “NATO”, “COVID-19”). These metadata were used in our narrative element extraction as additional LLM grounding and supporting material.

3.1.2. Dataset Reconstruction and Reproducibility Protocol

For this study, we reconstructed the English part of the *EUvsDisinfo: A Dataset for Multilingual Detection of Pro-Kremlin Disinformation in News Articles* dataset, using its publicly available metadata released on Zenodo [61]. The metadata includes article identifiers,

publishers, URLs, publication dates, language labels, topic keywords and class labels (disinformation vs. trustworthy). The metadata are distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits reuse with attribution.

Full article texts were reconstructed from the original URLs using the Diffbot Article API (/v3/article) [59]. Article retrieval was performed between January and February 2025 using a personal Diffbot API key. For each article URL, the API was queried to extract the main article text and title. If text extraction failed or returned empty content, a second attempt was made using an archived version of the URL (when available) via the Internet Archive [62]. Articles that could not be retrieved successfully after these steps were excluded from further analysis.

All successfully collected article texts were cached locally in JSONL format together with the corresponding article metadata. These cached texts were used consistently for all downstream processing steps, including summarization, knowledge graph construction, narrative extraction and evaluation. No live re-fetching of article content was performed after this initial reconstruction phase.

Following reconstruction, articles were further filtered to ensure that disinformation and trustworthy subsets were similar in terms of the topics discussed in the articles. We assessed topic similarity via sentence embeddings generated with the *all-MiniLM-L6-v2* model. Article titles and keyword fields were embedded, and cosine similarity was used to match articles across classes with a similarity threshold of 0.35 and top- $k = 5$ nearest neighbors. Matching was performed automatically and symmetrically between classes. Articles without sufficiently similar counterparts across classes were excluded to reduce confusion caused by topics.

The Supplementary Materials provide a CSV file listing all article URLs included in the final dataset, together with publication metadata (publisher, domain, title, author, class label and publication date) to enable reconstruction of the dataset using the protocol described above. Complete reproducibility of the reconstructed texts cannot be guaranteed, because some original URLs have changed, been removed or modified since the time of collection. The cached reconstructed texts used in this study are retained by the authors and may be made available upon reasonable request, subject to the access policies of the original publishers.

3.1.3. Hardware Setup and Software Environment

All experiments were conducted in Google Colab Pro [63], which provides a cloud-based environment. The computational workflow relied on external LLM inference via API calls using the OpenRouter [64] platform to access hosted instruction-tuned models. As model inference was performed on third-party cloud infrastructure, no fixed local hardware configuration, such as CPU or GPU specifications, was required or controlled during the experiments.

We used the local environment for data preprocessing, prompt construction, result aggregation and evaluation. The implementation was written in Python 3.12.12 and executed within the Colab runtime. Specific model identifiers, SDK versions, API providers and inference parameters are reported in detail in Appendix E to ensure reproducibility.

Because inference was conducted via API-based access to externally hosted models, computational performance and latency may vary depending on provider-side infrastructure. However, the experimental design does not depend on specialized hardware and can be reproduced in comparable cloud-based notebook environments.

3.1.4. Extractive Summaries

The first step in data processing was text summarization, applied to the news articles in our final dataset. This step was implemented to reduce noise in texts and to use computational resources more efficiently in the later stages. We applied extractive summarization for this task to extract the most important sentences from the given text, while preserving factuality and original wording [65]. Thus, we used article summaries for the extraction task as one part of the input for LLM ensembles, the other two parts being article-level KGs and article metadata.

The cleaned news articles were summarized using two extractive summarization approaches—LexRank Summarizer [66] and BERT Extractive Summarizer [67]. We computed mean ROUGE-1/2/L F1 scores [68] against the original article text to choose from two summary candidates. Summaries generated by the approach that had higher ROUGE values, showing higher source-overlap, were used in the experiments.

The length of the extractive summary was tied to text length, striving for 20–30% of the original text. To ensure adequate quality, we also manually inspected the summaries to correct such issues as occasional incomplete final sentences. The illustrative examples of summaries used in this study are presented in Appendix A.

3.1.5. Knowledge Graphs

To ground LLMs during the extraction task, article-level knowledge graphs (KGs) were constructed from extractive summaries via an approach based on abstract meaning representation (AMR) [28] and generation of linked RDF/OWL ontologies from natural language. AMRs were well-suited for extracting WHO–DID–WHAT information from text, and linked RDF/OWL ontologies, aligned to shared formal semantics and connected to background knowledge, provided the necessary context for narrative understanding. For the latter part, we used the FRED machine-reading tool [29], which turned natural language text into machine-interpretable knowledge graphs.

Abstract meaning representation (AMR) [28] encodes meaning at the sentence level as a single-rooted directed acyclic graph where nodes are concepts, such as lexical items, PropBank [69] frames or special keywords, and edges—semantic relations [28]. AMR integrates previously separate annotation layers (NER, co-references, semantic roles) into one knowledge-based representation, relies on frame-semantic resources such as PropBank [69] and FrameNet [70], and supports links to external knowledge bases such as DBpedia [28,71]. Therefore, AMRs provide rich graphs centered on a predicative core, typically an event and its participants, which corresponds to arguments in the frame structure [72]. This event-rooted organization produces graphs that are well-suited for parsing, comparison and matching, which makes AMR a useful bridge between networks designed for information retrieval and deeper semantic analysis. The example of AMR with description is presented in Appendix B.

We used Text2AMR2FRED [73]—a text-to-KG pipeline for transforming our article summaries into rich article-level KGs. Text2AMR2FRED combines end-to-end AMR parsing with FRED semantic web (SW) machine reader [29] and produces OWL-compliant, interoperable KGs which can be queried and aligned with external knowledge bases [28,29]. By using AMR’s abstraction-friendly representation, Text2AMR2FRED [73] aimed to reduce errors common in component-based pipelines and avoid data augmentation such as lexical substitution [74,75]. This system is event-centric. It is organized around PropBank frames [69] and operates in two steps (text-to-AMR parsing and AMR-to-FRED translation) to produce RDF KGs that follow FRED’s theoretical model [29], which, at its core, uses discourse representation structures (DRSs) from Hans Kamp’s discourse representation theory [76].

These KGs can be aligned with domain ontologies and various other KGs to reveal implicit knowledge hidden in text. They also support structured queries across corpora. The produced KGs are then enriched via Framester [77] and are further linked to resources such as WordNet [78], DBPedia [79] or DOLCE-Zero [80]. Also, a word-sense disambiguation (WSD) step assigns senses to AMR [28] elements that are missing lexical links, which increases the informativeness of a KG [77]. The example result of the Text2AMR2FRED [73] pipeline with a short description is supplied in Appendix C. For parsing our news article summaries to AMRs and then translating them to FRED [29], we used the Python 3.12.12 library `py_amr2fred` [81].

3.2. Methods

This subsection presents the complete KG-grounded narrative extraction pipeline. We begin with ablation evidence and design rationale (Section 3.2.1), followed by a detailed description of the models and configurations used (Section 3.2.2). The prompting strategy for extraction and related tasks is then introduced (Section 3.2.3). The subsection is concluded with a description of the evaluation framework, including narrative pattern analysis, alignment with GDELT, and assessment against an expert-annotated sample.

3.2.1. Relation to Prior Ablation Evidence and Design Rationale

Several design choices in the proposed pipeline, including KG grounding and extractive summarization, were informed by comparative analyses reported in [60]. That prior work evaluated LLM-only, KG-only and KG-augmented narrative extraction pipelines on the same dataset and summaries but employed a different KG construction method based on relational triple extraction. Also, narrative element extraction relied on a single LLM.

In particular, it was shown that a KG-augmented pipeline achieved broader narrative-element coverage and higher embedding-based coherence than both LLM-only and KG-only pipelines [60]. The augmented approach retained most LLM-derived characters and events and added KG-informed elements, producing denser and more internally consistent narrative representations. However, performance on causal links remained similar to that of the LLM-only pipeline, while framing extraction stayed a known limitation. At the same time, the augmented pipeline was also better aligned with external knowledge bases, such as Wikidata, indicating improved factual grounding. These findings motivated the continued use of KG grounding as a core component of the present pipeline.

Other design choices were not evaluated via explicit ablation in the present study and are therefore not claimed to be causally isolated contributors to performance. Extractive summarization was adopted to reduce noise and computational cost while preserving original wording, following established practices in narrative and information extraction pipelines [65].

Ensemble-based extraction and role specialization were motivated by prior evidence that combining models with complementary inductive biases improves robustness and precision–recall trade-offs in structured extraction tasks [22,27]. Schema-constrained JSON normalization and post-hoc formatting were also introduced to ensure consistent, machine-readable outputs and to reduce structural errors in downstream processing, following existing work on multi-stage and normalization-oriented LLM pipelines [54,55].

The present study does not re-run baseline ablations along these axes. Instead, it evaluates a finalized pipeline that integrates and extends these components through new ensemble policies, a dedicated critic module, expert-based evaluation and external validation via GDELT. Accordingly, claims in this article focus on empirical performance and structural patterns observed in the integrated system, and are not intended to isolate the causal contribution of individual components.

3.2.2. Models

We employed a small ensemble of instruction-tuned large language models, each assigned a specific functional role within the pipeline. Mistral-Small-3.1-24B-Instruct was used as a formatter for all tasks. We paired Gemma-3-27B-IT with Mistral-Small-3.1-24B-Instruct for narrative element extraction (characters, events, causal links, framing edges). For identifying missing links, contradictions and framing inconsistencies, we employed Mistral-Small-3.1-24B-Instruct + Gemma-3-27B-IT ensemble as a two-judge critic. Also, ensemble Qwen-2.5-72B-Instruct with Llama-3.1-8B-Instruct was used for entity linking, which was an auxiliary task in this study.

Figure 1 summarizes the main framework. As shown in the figure, extractive summaries, article-level KGs and metadata are used as primary inputs to the narrative extraction component, which produces raw narrative elements. These elements are then passed through a dedicated formatter to obtain a consistent, structured representation. The formatted narrative elements are then provided as the main input to the critic ensemble, while the original summaries, KGs and metadata serve as grounding context to support verification, consistency checking and the identification of missing links or contradictions. Outputs produced by the critic are again processed by the formatter to produce normalized verification results.

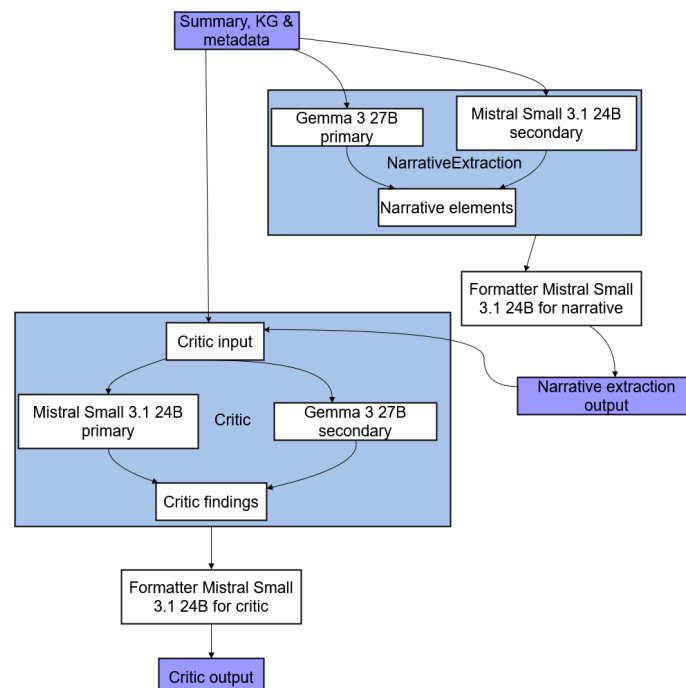


Figure 1. LLM ensemble setup illustrating the narrative extraction, formatting, and critic-based verification components used in the proposed framework. Auxiliary components, such as entity linking, are omitted for clarity. Source: authors' contribution.

We chose Mistral-Small-3.1-24B-Instruct (hereafter, Mistral) as a formatter because it is optimized for low-latency, efficient inference and performs competitively against larger open models on general knowledge, coding and instruction-following benchmarks [82]. This makes it suitable for generating schema-constrained outputs, handling long, nested JSON structures and serving as a stable base model that standardizes outputs from different extractor or critic ensembles. Also, we chose to use a single formatter to reduce formatting errors and make evaluation easier.

Furthermore, for narrative element extraction, we paired Gemma-3-27B-IT (hereafter, Gemma) with Mistral. The instruction-tuned version of Gemma consistently follows in-

structions, supports multiple languages, handles long contexts effectively and successfully competes against larger proprietary models on various reasoning and comprehension benchmarks [83]. Thus, Gemma is suitable as a precise extractor for complex narrative structures like characters, events, causal links and framing edges. In turn, Mistral provides fast, stable generation and solid general language modeling. We used it as a complementary extractor and as a “second opinion” on Gemma’s outputs. Recent studies on collaborative and multi-agent LLM systems for structured scientific information extraction have shown that using multiple models with different strengths for cross-checking or aggregating results leads to better quality outputs than using a single model [22,27]. Our ensemble followed this approach, as Gemma focused on recall and nuanced reasoning, while Mistral improved precision, schema reliability and introduced a different bias. Their agreement formed a high-confidence “core narrative”.

For the critic component, which detects missing links, factual or logical contradictions and framing inconsistencies, we used the Mistral and Gemma pair again, but, in this case, we swapped the roles of the primary and secondary models from the narrative element extraction ensemble. The recent literature on LLM-as-judge suggested that LLM evaluators could match or sometimes outperform humans across various tasks, while also highlighting systematic biases, sensitivity to prompt design and model-specific issues [84,85]. By employing two judges that differ in architecture and in training, we combined insights from both the LLM-as-judge and ensemble-learning literature, which suggested that relying on multiple, diverse evaluators could reduce individual biases and produce more stable, reliable results [21,22,84].

Finally, we performed entity linking as an auxiliary step to support mapping extracted narrative elements to GDELT. For this, we paired Qwen-2.5-72B-Instruct (hereafter, Qwen) with Llama-3.1-8B-Instruct (hereafter, Llama). Qwen is an open-weight model which performs well in reasoning, multilingual understanding and analysis of structured data. It was designed for following instructions and understanding long contexts [86]. Meanwhile, Llama 3 models, including the smaller 8B version, achieve robust results in general-purpose language understanding and perform competitively on standard benchmarks, achieving results that are similar to larger models [87]. Pairing a larger model with a lighter but still capable general model supports research findings that diverse LLM ensembles can provide better accuracy and cost benefits. Such ensembles also improve robustness compared to homogeneous ones or single models [22,23,88]. Therefore, Qwen adds recall and multilingual knowledge for potential entities, while Llama offers complementary insights and helps to correct errors. Thus, our choice is consistent with ensemble strategies like LLM-Blender [24] and routing-based LLM groups [25].

3.2.3. Prompts

In our pipeline, we used two types of prompts: extraction prompts that pointed to article summaries, article-level KGs and article metadata for the extraction and formatting prompts (for details on prompts, see Appendix D). The extraction prompts defined both the schema of narrative elements and a set of principles for the model to follow. The schema distinguished:

1. characters/actors (individuals, organizations, groups, countries) with role labels and optional links to grounding triples;
2. events, including who did what to whom, plus optional time and place if available in the article data;
3. causal links between events with a typed relation such as *cause*, *enable* or *lead to* and a short natural-language justification;

4. framing edges, where a framing actor (e.g., the article author, a quoted politician, a media outlet) portrays a target (actor or event) in a particular way (e.g., *threatening*, *heroic*, *illegitimate*), optionally marked with stance (*support/criticize/neutral*) and linked to supporting text snippet.

The extraction prompts explained this schema, described how to use the article summary text and the knowledge-graph triples, as well as metadata as context, and enforced several principles:

- Stay faithful to what the article actually says (no hallucinated actors or events);
- Treat triples as background evidence rather than extra facts to invent new claims;
- Preserve uncertainty and hedging (mark speculative or contested links instead of turning them into hard facts);
- Separate the narrator's framing from the quoted speakers' framing.

On top of that, the prompt explicitly encouraged the model to look for strategic and narrative-level causality (e.g., "claiming X enables Y's justification", "fear narrative prepares the audience for Z"), not just simple physical causation.

The formatting prompts then took the model's free-form extraction (or critic's comments) and converted it into a strict JSON representation. These prompts restated the schema, listed the exact field names and allowed relation/stance labels and provided in-prompt micro-examples. The formatting instructions required the model to:

1. output each narrative element as a separate object with stable ids;
2. fill only the schema-defined fields;
3. attach text spans and triple identifiers where available;
4. avoid any extra commentary or explanations.

This second stage ensured that outputs from different models and configurations were machine-readable, comparable and easy to aggregate across the dataset. This formatting stage effectively performs schema-constrained JSON normalization across models.

The critic prompts turned the chosen pair of LLMs into a structured quality-control layer. Each critic call received article summary, article metadata, RDF triples for that article and the already-extracted narrative elements (characters/actors, events, causal_links, framing_edges). The prompt then instructed the model not to rewrite the narrative, but to audit it against text, metadata and the triples, i.e.,

- a. Propose missing links (especially missing causal/event links) that are clearly supported by either the summary or the graph, also metadata;
- b. Flag contradictions where extracted narrative elements conflict with the triples, text or metadata;
- c. Flag framing inconsistencies, such as systematically one-sided descriptions of actors or events.

All of this had to be outputted in a strict JSON schema (missing_links, contradictions, framing_inconsistencies) where each issue was grounded with short quotes, natural-language explanations and references to TRIPLE_IDS from the triples block. In our setup, a "primary" critic formulated issues and a "verifier" critic re-ran the same audit from scratch, so we could measure agreement and focus on issues that were stable across the models.

3.2.4. Evaluation

Narrative Pattern Analysis

For each article, we computed narrative-level statistics from the extracted structures and assessed their alignment to article-level KGs constructed via *text2amr2fred* [73] pipeline. For both LLMs in the extraction ensemble (Gemma and Mistral) and for their derived

union and consensus ensembles, we counted the number of characters, events, causal links, and framing edges per article, as well as the number of referenced graph triples for each relation type (event, causal, framing). Every character and event in the narrative carried a *node_iri* which anchored it to a concrete node in the AMR-FRED graph, while causal and framing links stored lists of triple identifiers (*triple references*) that connected narrative elements to concrete RDF statements in the particular graph. Thus, instead of introducing unanchored entities and relations, LLMs were required to reuse existing graph nodes and triples whenever possible. As we used a small number of models with fixed roles, we did not introduce weighted aggregation or composite evaluation scores.

For analysis, we summarized per-article counts and proportions, including the number of narrative nodes, the number of triple references per relation type, and grounding coverage (i.e., whether an element had at least one associated triple reference). In addition to raw counts, we computed simple density measures such as the number of causal links and framing edges per event, and event-slot coverage statistics capturing how often *who*, *what*, *when* and *where* fields were filled for each event. To quantify ensemble effects, we compared union vs. consensus outputs within each article using paired Wilcoxon signed-rank tests separately for disinformation and trustworthy news, reporting medians and IQRs. Given the exploratory nature of this analysis, we reported *p*-values without multiple-comparison correction.

For qualitative pattern summaries, we reported descriptive statistics for Gemma, Mistral, and consensus (we used the union ensemble only as an intermediate candidate pool for ensemble-effect analysis). We further grouped narrative elements into coarse semantic categories using a simple lexicon- and rule-based coding schemes, i.e., for characters, we collapsed labels into *persons*, *organizations*, *places*, *countries* and *other*; for causal links, we analyzed relation types such as *enables*, *leads_to* and *causes*; for framing edges we examined the distribution of frame types (e.g., *support*, *opposition*, *negative*, *accusation*, *blame*, *denial*, *reporting*).

We applied analogous pattern analysis to the critic ensemble. For each article, a primary critic (Mistral) and a verifier critic (Gemma) independently reviewed the extracted narrative elements and proposed three types of issues: *missing links*, *contradictions* and *framing inconsistencies*, providing KG-grounded justifications via triple references. We summarized issue counts and grounding validity by class and critic model. For missing links, we used a lexicon- and rule-based coding scheme over the critic's *candidate_link* text to group suggestions into coarse semantic categories: *causal*, *policy/decision*, *speech/stance*, *background* (historical or contextual facts), *actor role* (descriptions of roles, affiliations or statuses of actors), *temporal*, *location* and *other* category. We grouped contradictions into *background fact gaps*, *missing events*, *role gaps*, *causal gaps* and *other* using heuristic pattern rules over the free-text *issue* field. Framing inconsistencies were already annotated by critic models with discrete types (*causal_flip*, *role_flip*, *sentiment_flip*, *temporal_conflict*, *modality_shift*), which we analyzed directly. Finally, to characterize ensemble effects for our critic models, we quantified critic-critic agreement using multiple matching schemes (exact string equality, character *n*-gram soft matching, a relaxed TF-IDF/Jaccard matcher and a relaxed matcher with an additional embedding-based gate), and summarized agreement statistics by class (disinformation vs. trustworthy) and critic model.

Alignment with GDELT

GDELT (Global Database of Events, Languages and Tones) is a massive open-source news database that monitors broadcast, print and online media in nearly every country in over 100 languages, encoding events, actors, locations, topics and tone into a structured, continuously updated feed of global news [89–91]. Recent studies have used GDELT to

study global climate-change adaptation recognition [92], news-flow networks [93], refugee portrayals [94] and ESG-macroeconomic dynamics [95], showing that it provides broad, multi-country coverage suitable for large-scale media analysis. Using GDELT as an external evaluation resource for our extracted narrative elements, therefore, provided an independent, globally sampled point of comparison. It allowed us to assess how far model-derived narratives align with, complement or diverge from established patterns in large-scale news coverage. Recent methodological work also emphasizes that, despite noise and the need for cleaning, GDELT offers high coverage and integrity across countries and media outlets [89,96,97], which fits well with using it as a large, imperfect but informative reference for our KG-grounded narrative extraction.

In order to match our extractions to GDELT, we first combined outputs of the two models of extraction ensemble using fixed, type-specific rule-based ensemble schemes (union, de-duplication and gated inclusion), as detailed in Appendix F). Then, for matching itself, different queries have been constructed using the diverse parameters. For actors and events (“Do they show up in GDELT nearby in time & space?”), we used GDELT purely as a reflection check. For each article, we took the fused actors/characters (persons, organizations, countries) and locations (characters/actors, marking place, city or country, plus event “where” labels that were part of event extractions) and searched GDELT within a ± 7 -day window around the article’s publication date. We then looked for events which *Actor1Name/Actor2Name* overlapped with our actor set and that also either mentioned one of our locations (via *ActionGeo_** fields) or matched an acronym form of an actor or organization name. Any such hits were counted as external “support” that the actors and event contexts extracted by our pipeline corresponded to entities and situations that also appeared in global news coverage.

For causal links (“Do the same directional actor pairs appear in GDELT events?”), we used GDELT as a directional interaction proxy rather than trying to match our full cause-related semantics. For each extracted causal link, we took its source and target actors and formed ordered pairs ($A \rightarrow B$) for the article. We then looked in GDELT within ± 7 days of the article’s publication date for events where *Actor1* matched *A* and *Actor2* matched *B*. Such hits were counted as external support that this actor pair was seen interacting in that direction in broader news. This did not prove that GDELT encoded the same causal claim, but it provided a practical signal of how often our extracted causal relations aligned with real directional interactions between the same actors in global news.

Finally, for framing (“Does the surrounding GDELT event mix & polarity look compatible with the narrative’s framing?”) we again used GDELT only as a coarse external signal. For each article, we first identified all GDELT events that matched our actors/locations within a ± 7 -day window, then examined their event-level attributes rather than trying to match our fine-grained narrative frames directly. We used several metrics present in GDELT—textual tone and the *Goldstein Scale*—to get a rough sense of negativity or positivity around those events. We also mapped each event’s *EventRootCode* into broad buckets (*cooperation*, *verbal conflict*, *material conflict*, *neutral*) and computed the share of matched events per article that fell into each bucket. We treated these distributions as a very coarse reflection of the context, surrounding the event, as a proxy for assessing framing extractions.

Evaluation Against Expert-Judged Sample

Sampling procedure. For evaluation, we used a stratified random sampling procedure to address both article-level and element-level structure. First, each predicted element was assigned to strata defined by $\text{class} \times \text{element_type} \times \text{confidence tier}$, where $\text{class} \in \{\text{disinformation}, \text{trustworthy}\}$, $\text{element_type} \in \{\text{character}, \text{event}, \text{causal link}, \text{framing edge}\}$, and $\text{tier} \in A, B, C$ based on ensemble support and grounding. Within each *element_type*

and class, we then randomly selected 30 disinformation and 30 trustworthy news articles per type. From each selected article, we sampled up to 3 Tier-A elements, up to 2 Tier-B elements and up to 2 Tier-C elements at random from all candidates of particular element type in that article. This two-stage process (article sampling, then element sampling within articles and tiers) ensured that both classes were represented, that all element types and confidence tiers were covered and that multiple elements per article were annotated to make both micro (element-level) and macro (article-level) precision metrics meaningful. The diagram, summarizing the sampling procedure, is supplied in Figure A3 in Appendix G.

Annotation guidelines. For expert evaluation, we asked a domain expert to judge the correctness of models' outputs using a part-credit scheme 0, 0.5, 1. The annotation task included self-contained context: basic article metadata (identifier, class label, title, etc.), an extractive summary of the article and one or more quoted passages that the extractor system claimed as evidence. For narrative elements (characters, events, causal links and framing edges), the expert also saw a structured description of the predicted element (e.g., a character name with type and mention span, an event with who/what/where/when and its evidence sentence, or a causal link connecting two events with a relation label and an evidence passage). For critic items (missing links, contradictions and framing inconsistencies), the expert was provided with the critic's short label and textual description of the issue, together with the quoted passage and a compact list of relevant narrative elements to which the critic referred. The expert was explicitly instructed to base judgments only on the provided text excerpts and element descriptions and not to use external knowledge or their own interpretation beyond what was explicitly stated.

A score of 1 was assigned when the system output was judged to be fully correct. For narrative elements, this meant that the element corresponded to a real entity, event, causal relation or framing relation clearly expressed in the quoted passage and described with sufficient specificity (e.g., correct direction of a causal link or correct target of a framing edge). For critic items, a score of 1 was given when the critic correctly interpreted the passage and identified a genuine problem in the narrative representation (e.g., an important missing relation or a framing mismatch). A score of 0.5 denoted a partially correct item, where the underlying entity, event or relation was recognizable, or the critic pointed to a real issue, but the description was incomplete, overly generic or distorted (e.g., wrong strength of the claim, missing a key participant). A score of 0 was assigned when the output was not supported by the quoted passage, the text was misread, the content was fabricated or, in the case of critic items, the described inconsistency did not hold given the text and the listed narrative elements.

Evaluation metrics. We evaluated the quality of the extracted narrative elements against the expert-judged sample using part-credit precision at two levels. Each system output (character, event, causal link, framing edge or critic item) received a score in $\{0, 0.5, 1\}$ from the expert, where 1 denoted a fully correct prediction, 0.5 a partially correct prediction and 0—an incorrect prediction according to the guidelines described above. Micro part-credit precision was then computed over all items of a given type as the sum of these scores divided by the total number of items, capturing the overall proportion of correct items among all predictions. To account for variability across articles, we also reported macro part-credit precision at the article level, where for each article we first computed the mean score over its items of a given type and then averaged these article-level means across the judged subset.

4. Results

This section reports the empirical findings of our study in three stages. First, Section 4.1 examines extracted narrative structures and patterns, including the effects of ensemble

aggregation on narrative elements and critic outputs, as well as descriptive distributions across element types. Next, Section 4.2 evaluates the extracted elements through alignment with GDELT as an external reference. Finally, Section 4.3 presents the results of expert-based evaluation, including precision assessment and qualitative error analysis.

4.1. Analysis of Narrative Structures and Patterns

4.1.1. Analysis of Ensemble Effects

Ensemble effects on narrative element extraction. We compared, within each news article, the narrative structures produced by two ensemble strategies. The first was a union ensemble, which collected all narrative elements accepted by either extractor, and the second was a more conservative consensus ensemble, which retained only elements agreed upon by both models. Differences between these two variants we evaluated separately for disinformation and trustworthy news using paired Wilcoxon signed-rank tests at the article level. Tables 4 and 5 report the corresponding results.

Table 4. Union vs. consensus ensemble for disinformation (Wilcoxon signed-rank tests, paired per article). Source: authors' contribution.

Metric	Union Median [IQR]	Consensus Median [IQR]	<i>p</i>
$n_{\text{characters}}$	4.00 [3.50–5.00]	4.00 [3.50–5.00]	0.1573
n_{events}	5.00 [4.00–6.00]	5.00 [4.00–6.00]	0.0833
$n_{\text{causal links}}$	6.00 [4.00–7.00]	0.00 [0.00–1.00]	3.25×10^{-21}
$n_{\text{framing edges}}$	2.00 [1.00–2.00]	1.00 [0.00–1.00]	6.78×10^{-18}
$n_{\text{event triple refs}}$	8.00 [5.00–10.00]	8.00 [5.00–10.00]	0.1025
$n_{\text{causal triple refs}}$	8.00 [5.00–11.50]	0.00 [0.00–2.50]	1.24×10^{-20}
$n_{\text{framing triple refs}}$	2.00 [1.00–3.50]	1.00 [0.00–1.00]	2.16×10^{-16}

Table 5. Union vs. consensus ensemble for trustworthy news (Wilcoxon signed-rank tests, paired per article). Values are median [IQR]. Source: authors' contribution.

Metric	Union Median [IQR]	Consensus Median [IQR]	<i>p</i>
$n_{\text{characters}}$	5.00 [4.00–6.00]	5.00 [4.00–6.00]	0.0588
n_{events}	5.00 [5.00–7.00]	5.00 [5.00–7.00]	0.0455
$n_{\text{causal links}}$	8.00 [5.00–11.00]	0.00 [0.00–1.00]	7.68×10^{-17}
$n_{\text{framing edges}}$	2.00 [2.00–2.00]	1.00 [0.00–1.00]	1.01×10^{-14}
$n_{\text{event triple refs}}$	6.00 [4.00–10.00]	6.00 [4.00–10.00]	0.0633
$n_{\text{causal triple refs}}$	7.00 [3.00–14.00]	0.00 [0.00–2.00]	1.12×10^{-15}
$n_{\text{framing triple refs}}$	2.00 [1.00–3.00]	0.00 [0.00–1.00]	1.56×10^{-12}

Our analysis revealed that characters were stable across union and consensus variants in both classes. Events were largely stable but showed a small though statistically detectable shift in trustworthy news, even though both classes had identical medians in both ensemble variants (Table 5). Because the medians were identical, the Wilcoxon significance reflected differences in the distribution tails rather than a change in central tendency. Also, anchored nodes (*node_iri*) were rarely removed by the consensus filter. We observed a similar pattern for event triple references, which showed no statistically significant differences between union and consensus, indicating preserved event grounding in the AMR-FRED graphs.

In contrast, relational structure differed strongly between the two ensemble variants. For disinformation, the median number of causal links per article dropped from 6.0 in the union ensemble to 0.0 in the consensus ensemble, indicating that many articles contained no consensus causal links, and framing edges decreased from 2.0 to 1.0 (Table 4). A similar pruning effect was observed in trustworthy news, where causal links dropped from a

median of 8.0 to 0.0, and framing edges decreased from 2.0 to 1.0 (Table 5). The same pattern held for triple-reference counts, i.e., while event triple references remained largely unchanged, causal and framing triple references fell from moderately large medians in the union variant to mostly zero or one triple per article in the consensus variant.

Overall, the union ensemble produced substantially denser relational structure (6–8 median causal links per article) than the consensus ensemble (median: 0), while preserving core actors and events (Tables 4 and 5). This contrast between an exploratory union view and a conservative consensus view is useful for downstream analysis, as it enables comparisons of how densely disinformation and trustworthy narratives layer causal and framing structure over largely shared, graph-anchored actors and events within each article.

Ensemble effects on the LLM critic. We applied a critic ensemble to analyze the extracted narrative structures, with Mistral acting as the *primary* critic and Gemma as the *verifier*. For each article, both models independently reviewed the extracted narrative elements and flagged three types of issues: *missing links*, *contradictions*, and *framing inconsistencies*. Table 6 summarizes per-article issue counts by class and critic model, together with the proportion of cited triple references that successfully resolve in the article-level KG.

Table 6. Per-article issues by critic model and class (means across articles). The proportions of valid triples indicate the mean share of triple references that successfully resolve in the KG for each issue type. Source: authors’ contribution.

Critic	Class	n_{miss}	n_{contra}	n_{frame}	n_{all}	Prop. Valid _{miss}	Prop. Valid _{contra}	Prop. Valid _{frame}
Primary	Disinformation	4.92	1.58	1.71	8.20	1.00	1.00	1.00
Primary	Trustworthy news	4.84	1.55	1.68	8.06	1.00	0.99	0.99
Verifier	Disinformation	4.97	1.08	1.58	7.63	0.93	0.92	0.89
Verifier	Trustworthy news	5.00	1.40	1.71	8.11	0.97	0.98	0.98

Both critics flagged a similar total number of issues per article (around eight on average), with missing links constituting the largest category, followed by contradictions and framing inconsistencies. The primary (Mistral) critic tended to flag more contradictions, particularly in disinformation articles, while the verifier (Gemma) critic was slightly more conservative on contradictions but comparable on framing. Triple references were highly verifiable in the KG as the primary critic achieved 0.99–1.00 validity across issue types, while the verifier achieved 0.89–0.98 depending on issue type and class (Table 6).

Article-level correlations between critics were near zero (e.g., $r = 0.04$ for total issues), indicating that the two models often focused on different subsets of issues (Table 7). In this setting, such diversity is intentional. We used complementary critics to broaden issue coverage rather than to maximize inter-critic agreement.

Table 7. Pearson correlations between primary (Mistral) and verifier (Gemma) critics at the article level (one row per article). Source: authors’ contribution.

Measure	r
Total issues (n_{all})	0.04
Missing links (n_{missing})	−0.01
Contradictions ($n_{\text{contradictions}}$)	0.13
Framing inconsistencies (n_{framing})	0.10

To assess agreement at the level of individual issues, we compared critic outputs under four matching schemes: exact string equality, character n -gram based soft matching, a relaxed TF-IDF similarity combined with token-level Jaccard overlap, and a further

relaxed variant incorporating sentence embeddings as a gating mechanism. Table 8 reports mean agreement scores across all articles. Exact matching produced near-zero agreement, reflecting substantial surface-level variation in how the two critic models phrased their findings. Agreement increased under soft matching and improved further under the relaxed schemes, with the embedding-gated matcher achieving the highest overall agreement (19.3%). Gains were most pronounced for contradictions (from 10.7% to 15.4%) and framing inconsistencies (from 5.5% to 13.1%).

The two critics also differed in how strictly they grounded their outputs. The primary critic almost always cited triples that resolved in the KG across issue types, whereas the verifier critic occasionally referenced triples that could not be resolved, particularly for disinformation articles. This asymmetry suggests a trade-off in critic behavior: the primary critic maintained stricter grounding fidelity, while the verifier sometimes extrapolated beyond the available triples.

Table 8. Mean agreement between primary and verifier critics under different matching schemes (percentages of matched issues; averages across articles). Source: authors' contribution.

Matching Scheme	Missing	Contradictions	Framing	Overall
Exact (serialized)	0.2	0.9	0.5	0.5
Soft	22.9	6.2	3.3	10.8
Soft relaxed	27.9	10.7	5.5	14.7
Soft relaxed + embed.	29.4	15.4	13.1	19.3

Even under embedding-gated matching, the mean overall agreement was 19.3% (Table 8). Agreement was highest for missing links (around 29%), intermediate for contradictions (about 15%), and lowest for framing inconsistencies despite substantial gains from relaxed matching. Overall, these patterns indicate that the critic ensemble provides complementary perspectives: agreement under relaxed matching can serve as a high-precision signal, while disagreements preserve diversity for more exploratory error analysis.

Ensemble effects on entity linking. For each node in article-level KGs (*node_iri*, label, snippet), we first retrieved a small set of Wikidata [98] candidates (QID, label, gloss) via fuzzy matching. Nodes with no candidates (4391/6006 mentions) were excluded, and nodes with a single candidate (142/6006 mentions) were treated as already linked. For the remaining nodes with two or more candidates, we instructed Qwen to select the best candidate with an associated confidence score (details on entity linking provided in Appendix D.5).

On this candidate-available subset, Qwen alone linked 1348/1615 mentions (83.5%). We then applied Llama only to low-confidence Qwen cases (confidence < 0.7) as a verifier. This targeted fallback added 6 new links, increasing coverage on the candidate-available set to 1354/1615 (83.8 %, +0.37 percentage points) (Table 9). While the improvement was numerically modest, it demonstrated the intended efficiency trade-off, i.e., applying a secondary model only to uncertain cases produced incremental coverage gains without the cost of full dual-model linking.

Table 9. Coverage by assignment method on mentions with at least one Wikidata candidate ($N = 1615$). Source: authors' contribution.

Method	Coverage (Rows)	Coverage (%)	Δ Rows vs. Qwen	Δ pp vs. Qwen
Qwen-only	1348	83.50	0	0.00
Llama-on-lowconf	1354	83.84	6	0.37
Two-way consensus	1354	83.84	6	0.37

A simple two-way consensus (union of Qwen and Llama decisions with Llama priority when both provide a QID) produced the same coverage here, since Llama was only invoked for a small subset of low-confidence mentions. Entity linking results were utilised as an additional resource to facilitate matching extracted narrative elements to GDELT (more details are provided in Appendix F). As entity candidates were retrieved via live Wikidata [98] access using the MediaWiki API [99] at the time of the study, this step may introduce ambiguity due to fuzzy matching and occasional inconsistencies in owl:sameAs statements. No fixed Wikidata dump version was assumed, and entity linking errors were therefore treated as an inherent source of noise rather than exhaustively corrected.

4.1.2. Narrative Pattern Analysis

In the previous subsection, we analyzed ensemble effects, showing that the high-recall union ensemble (Gemma $\cup$ Mistral) provided a large candidate pool of narrative elements, while the consensus ensemble retained only a compact, high-confidence subset. Having established how much structure was pruned from union to consensus, we now focus on the patterns produced by the two single-model extractors (Gemma and Mistral) and by their consensus output. Therefore, in this section, we report descriptive statistics for the two single-model ensembles (Gemma and Mistral) and for their consensus ensemble. The union ensemble was used as an intermediate candidate pool in the ensemble analysis, but was not evaluated as a standalone variant, since it was not intended for further use.

Characters/actors. The coverage of character extraction was very high across ensembles (Table 10), as at least one character was identified in over 96% of articles in all model-class combinations. The distribution of character types (Table 11) revealed differences between models and between classes. For Gemma, in disinformation narratives, the largest share of characters consisted of persons (48.9%) and organizations (36.4%), though places (8.6%) and countries (4.7%) had smaller shares. In trustworthy news, persons dominated (64.0%), and organizations accounted for a smaller share (30.4%), while places and countries had few occurrences. Consensus ensemble followed Gemma closely, with disinformation split mainly between persons (49.7%) and organizations (35.6%), while trustworthy articles were focused on persons (65.6% persons, 30.2% organizations).

Table 10. Per-article counts of characters by model and class. Source: authors' contribution.

Model	Class	Mean	Median	Min	Max	% ≥ 1
Gemma	Disinformation	4.50	4	0	11	98.3
Gemma	Trustworthy news	5.59	5	1	13	100.0
Mistral	Disinformation	7.18	7	2	16	100.0
Mistral	Trustworthy news	7.99	8	2	17	100.0
Consensus	Disinformation	4.41	4	0	11	98.3
Consensus	Trustworthy news	5.31	5	1	13	95.7

Table 11. Distribution of character types by model and class (row-wise percentages). Source: authors' contribution.

Model	Class	Person	Org.	Place	Country	Other
Gemma	Disinformation	48.9	36.4	8.6	4.7	1.5
Gemma	Trustworthy news	64.0	30.4	2.9	2.3	0.4
Mistral	Disinformation	31.9	50.0	13.2	2.7	2.2
Mistral	Trustworthy news	45.4	41.3	8.1	3.6	1.5
Consensus	Disinformation	49.7	35.6	8.4	4.8	1.5
Consensus	Trustworthy news	65.6	30.2	2.2	1.8	0.2

Mistral displayed a different behavior, especially for disinformation, where organizations were the most frequent character type (50.0%), followed by persons (31.9%) and places (13.2%). This suggested a stronger focus on institutional and geographic actors. In trustworthy articles, Mistral was more balanced between persons (45.4%) and organizations (41.3%) but still extracted places more often (8.1%) than Gemma or Consensus. Overall, trustworthy news tended to contain slightly more characters per article than disinformation, while disinformation narratives, especially under Mistral, had a relatively larger share of organizational and place-based actors.

Events. Events were also found in almost all articles. The proportion of articles containing at least one event ranged from 95.7% to 100% across evaluated settings and classes (Table 12). Trustworthy news contained more events per article than disinformation, particularly under Mistral (11.24 vs. 7.99 events per article for trustworthy vs. disinformation).

Table 12. Per-article counts of events by model and class. Source: authors' contribution.

Model	Class	Mean	Median	Min	Max	% ≥ 1
Gemma	Disinformation	5.13	5	0	9	98.3
Gemma	Trustworthy news	6.01	6	3	12	100.0
Mistral	Disinformation	7.99	8	0	23	99.2
Mistral	Trustworthy news	11.24	11	0	32	95.7
Consensus	Disinformation	5.01	5	0	9	97.5
Consensus	Trustworthy news	5.73	5	0	11	95.7

The coverage of event slots (Table 13) revealed that Gemma, Mistral and Consensus almost always fill the *what* slot (100% of events in every condition) and provide *who* information in most events (Gemma: 88.0% vs. 86.4% for disinformation vs. trustworthy news; Mistral: 97.5% vs. 94.1%, respectively; Consensus: 99.5% vs. 99.1%, respectively). Temporal and spatial anchoring varied more. For Gemma, *when* was present in 36.4% of events for disinformation and for 43.1% of events—for trustworthy news. Meanwhile, *where* was filled in 42.8% of events for disinformation and in 48.5% of events—for trustworthy news. Mistral tended to under-specify time in disinformation events (*when* slot was filled in only 20.9% of events) but improved for trustworthy news (*when* slot was filled in 36.6% of events). As for spatial anchoring, coverage of slot *where* was comparable to or higher (34.6% vs. 49.5%) than that of slot *when*. Interestingly, the consensus ensemble produced fewer events, but they were better anchored, i.e., *when* was filled in 44.0% vs. 58.3% and *where*—in 55.4% vs. 65.5% of events for disinformation and trustworthy news, respectively. This suggested that consensus selectively retained events that were more complete in their temporal and spatial specification.

Table 13. Event slot coverage by model and class (percentage of events in which a slot is present). Source: authors' contribution.

Model	Class	Who	What	When	Where
Gemma	Disinformation	88.0	100.0	36.4	42.8
Gemma	Trustworthy news	86.4	100.0	43.1	48.5
Mistral	Disinformation	97.5	100.0	20.9	34.6
Mistral	Trustworthy news	94.1	100.0	36.6	49.5
Consensus	Disinformation	99.5	100.0	44.0	55.4
Consensus	Trustworthy news	99.1	100.0	58.3	65.5

Causal links. We found that causal links were common in outputs of individual models, but much sparser in the consensus ensemble (Table 14). Gemma produced about

2.35 and 2.66 causal links per article in disinformation and trustworthy news, respectively, and causal links were present in almost all articles (over 98%). Mistral achieved higher counts of causal links, particularly for trustworthy articles (3.82 vs. 6.38 causal links per article in disinformation vs. trustworthy news, respectively). When normalized by the number of events, Gemma produced roughly 0.46 causal links per event in both classes, while Mistral provided 0.48 vs. 0.57 causal links per event in disinformation vs. trustworthy news. In contrast, the consensus ensemble retained only about 0.42 vs. 0.45 causal links per article, corresponding to roughly 0.08 causal links per event and causal structure in only about one third of the articles.

Table 14. Per-article counts of causal links by model and class. Source: authors' contribution.

Model	Class	Mean	Median	Min	Max	% ≥ 1
Gemma	Disinformation	2.35	2	0	6	98.3
Gemma	Trustworthy news	2.66	3	1	4	100.0
Mistral	Disinformation	3.82	4	0	8	95.0
Mistral	Trustworthy news	6.38	5	0	17	95.7
Consensus	Disinformation	0.42	0	0	3	34.5
Consensus	Trustworthy news	0.45	0	0	3	31.2

The distribution of causal link types varied greatly across models (Table 15). Gemma provided a relatively balanced mix of *leads_to*, *enables*, and *causes*. In disinformation 38.9% of links had *leads_to*, 34.6%–*enables* and 23.2%–*causes*. A similar pattern was observed in trustworthy news (*leads_to*–45.3%, *enables*–30.8%, *causes*–21.9%). Mistral, in contrast, strongly favored *enables*, i.e., in disinformation, 69.0% of causal links were *enables*, 23.5% were *causes* and only 7.3% were *leads_to*. Meanwhile, in trustworthy news, the share of *enables* increased to 72.3%, *causes* accounted for 20.2% and *leads_to*–for 6.7%.

Table 15. Distribution of causal relation types by model and class (row-wise percentages). Source: authors' contribution.

Model	Class	Enables	Leads to	Causes	Other
Gemma	Disinformation	34.6	38.9	23.2	3.2
Gemma	Trustworthy news	30.8	45.3	21.9	2.0
Mistral	Disinformation	69.0	7.3	23.5	0.2
Mistral	Trustworthy news	72.3	6.7	20.2	0.7
Consensus	Disinformation	64.0	2.0	34.0	0.0
Consensus	Trustworthy news	52.4	2.4	45.2	0.0

The number of causal links in the consensus ensemble was significantly lower, however, *enables* and *causes* were the most pronounced among the remaining links. In disinformation, 64.0% of consensus links were *enables* and 34.0% are *causes*, while in trustworthy news, *causes* increased to 45.2% and *enables* remained high at 52.4%. Overall, consensus appeared to act as a strong precision filter, retaining only about 18% of Gemma's causal links and an even smaller fraction of Mistral's causal links, while favoring more canonical causal types.

Framing edges. We found that framing edges were the sparsest narrative element in Gemma. It was denser in Mistral, but it was strongly filtered in the consensus ensemble (Table 16). Gemma produced about 0.91 and 1.08 framing edges per article in disinformation and trustworthy news, respectively, with framing being present in approximately 78–83% of articles. Mistral achieved much richer framing structure, i.e., 3.88 and 4.65 framing edges per article for disinformation and trustworthy news, respectively, and this element was observed in approximately 96% of articles. Normalized by events, Gemma had

approximately 0.18 framing edges per event in both classes, while Mistral had 0.49 vs. 0.41 framing edges per event in disinformation vs. trustworthy news, respectively. Consensus ensemble retained only 0.60 and 0.67 framing edges per article for disinformation and trustworthy news (about 0.12 per event), with framing edges being present in over half of the articles.

The distribution of framing types (Table 17) revealed clear model- and class-specific tendencies. For Gemma, disinformation narratives were dominated by explicitly evaluative frames such as *negative* (24.1%), *accusation* (9.3%) and *blame* (4.6%), with some positive frames (4.6%) as well. Trustworthy narratives retained a small portion of *negative* (10.0%) and *accusation* (8.0%) framing but also showed slightly more *denial* (5.0%) and a broader range of additional framing types (e.g., *justification*, *accountability*, *legal responsibility*) which were less frequent and, therefore, were not individually listed in Table 17.

Table 16. Per-article counts of framing edges by model and class. Source: authors' contribution.

Model	Class	Mean	Median	Min	Max	% ≥ 1
Gemma	Disinformation	0.91	1	0	4	78.2
Gemma	Trustworthy news	1.08	1	0	3	82.8
Mistral	Disinformation	3.88	3	0	11	96.6
Mistral	Trustworthy news	4.65	3	0	15	95.7
Consensus	Disinformation	0.60	1	0	3	53.8
Consensus	Trustworthy news	0.67	1	0	3	54.8

Table 17. Distribution of selected framing types by model and class (row-wise percentages). Source: authors' contribution.

Model	Class	Support	Opposition	Negative	Positive	Accusation	Blame	Denial	Reporting
Gemma	Disinformation	0.0	0.0	24.1	4.6	9.3	4.6	2.8	0.9
Gemma	Trustworthy news	0.0	1.0	10.0	4.0	8.0	3.0	5.0	1.0
Mistral	Disinformation	18.2	14.1	6.9	10.4	1.9	5.2	0.2	4.1
Mistral	Trustworthy news	25.2	10.0	0.0	0.0	1.4	1.4	3.7	0.7
Consensus	Disinformation	0.0	0.0	23.9	7.0	8.5	5.6	1.4	1.4
Consensus	Trustworthy news	0.0	1.6	6.5	3.2	9.7	3.2	4.8	1.6

Mistral, on the other hand, highlighted stance-oriented frames such as *support* and *opposition*. In disinformation articles, *support* accounted for 18.2% and *opposition* for 14.1% of frames, with additional *positive* (10.4%), *negative* (6.9%), *blame* (5.2%) and *reporting* (4.1%) framing. In trustworthy news, *support* became even more prominent (25.2%) and *opposition* remained high (10.0%), while more specialized frames such as *denial*, *statement*, *claim*, *criticism*, *explanation*, and *fact-checking* also appeared but less frequently, thus, for brevity, these were not listed separately in Table 17. These patterns suggested that Mistral captures a rich spectrum of argumentative and meta-journalistic stances, especially in trustworthy news.

The consensus ensemble again behaved similarly to Gemma in framing composition, particularly for disinformation, where *negative* (23.9%), *accusation* (8.5%) and *blame* (5.6%) frames dominated. Meanwhile, in trustworthy news, consensus ensemble displayed slightly lower *negative* framing (6.5%) but a comparable share of *accusation* (9.7%) and a slightly higher proportion of *denial* (4.8%). Quantitatively, the consensus ensemble retained about 62–66% of Gemma's framing edges but only about 14–15% of Mistral's framing edges. This again indicated that it largely followed Gemma's more compact framing decisions while discarding much of Mistral's additional structure.

Critic. For missing links, each critic model proposed a short candidate statement that should be added to the outputs of the narrative extraction ensemble (e.g., a causal connection, a policy decision or a stance-bearing speech act). We grouped these statements into eight broad semantic classes based on the candidate text using a simple lexicon- and rule-based scheme: *causal* (explicit cause-effect or “leads to” language), *policy/decision* (sanctions, bans, referendums, formal decisions), *speech/stance* (statements, claims, criticism, support or opposition), *background* (historical or contextual facts), *actor role* (descriptions of roles, affiliations or statuses of actors), *temporal* (timing or stage information), *location* (spatial setting) and *other* category for remaining cases. Table 18 presents the distribution of these categories for both critic models, separately for disinformation and trustworthy news.

Table 18. Semantic categories of missing-link suggestions by critic and class (row-wise percentages). Source: authors’ contribution.

Critic	Class	Causal	Policy/ Decision	Speech/ Stance	Background	Actor Role	Temporal	Location	Other
Primary (Mistral)	Disinformation	8.7	10.4	21.5	2.7	7.5	3.2	21.5	24.3
Primary (Mistral)	Trustworthy news	10.7	10.9	19.3	2.0	9.6	1.8	23.8	22.0
Verifier (Gemma)	Disinformation	11.8	8.8	17.2	2.2	4.4	1.9	14.9	38.9
Verifier (Gemma)	Trustworthy news	11.2	10.1	17.6	0.4	8.4	1.9	19.1	31.2

For the primary (Mistral) critic, missing-link suggestions covered a broad range of semantic types. In disinformation, about 8.7% of missing links were explicitly *causal*, 10.4% focused on *policy/decision* content and 21.5% captured *speech/stance*. Also, many missing links consisted of missing *actor role* descriptions (7.5%) and *location* information (21.5%). For trustworthy news, the primary critic showed a similar profile, but with slightly higher shares of missing *causal* (10.7%) and *actor role* links (9.6%) and share of *speech/stance* (19.3%) and *location* (23.8%).

In the verifier (Gemma) critic, we observed broadly similar patterns but with a larger “Other” category, especially for disinformation. In both classes, it produced slightly more *causal* suggestions and fewer *speech/stance* links than the primary critic. However, the verifier critic still identified many missing links for *policy/decision*, *actor role* and *location* information. Overall, class-wise differences between disinformation and trustworthy news were modest compared to the differences between critic models, as both of them mainly proposed additional descriptive, stance-related, and contextual links (roles, locations), with a smaller but still significant subset of *causal* and *policy/decision* suggestions that extended narrative element extractions.

In addition, contradictions, marked by the critic models, were related to mismatches where extracted narrative elements conflicted with the triples, text or metadata. A heuristic categorization of the *issue* field suggested several recurring patterns:

- *Background fact gaps*, where historical or contextual facts (e.g., the launch date of the G7) are mentioned in the text but not encoded in the graph;
- *Missing events*, where high-level meetings, media appearances, or specific actions are absent from the narrative graph;
- *Role gaps*, where an actor’s role (e.g., as independent researcher, journalist, or official) is underspecified;
- *Causal gaps*, where the critic explicitly notes that a causal relationship described in the text is missing from the graph.

Table 19 presents a summary of the distribution of these categories across all contradictions flagged by our two critics. While the majority of contradictions belonged to the “Other” category, which reflected that our heuristic coding scheme was coarse, these

four types recurred systematically, especially for the primary critic, and gave us a useful high-level view of the kinds of discrepancies critic models highlighted.

Table 19. Types of contradictions flagged by the critics (row-wise percentages, aggregated over all articles). Source: authors' contribution.

Critic	Background Fact Gaps	Missing Events	Role Gaps	Causal Gaps	Other
Primary (Mistral)	1.8	8.1	3.9	3.0	83.1
Verifier (Gemma)	0.4	4.7	1.9	3.1	89.9

Also, a manual inspection of “Other” category revealed several repeating motifs, including stance and evaluation gaps (e.g., missing descriptions of an event as a “disaster” or an accusation as false), attribution and affiliation gaps (e.g., misreported group affiliations or recognition of new authorities) and intent or strategic framing gaps (e.g., missing statements about strategic goals, pressures or narrative motives). For simplicity, we kept these phenomena within the “Other” category in our quantitative analysis but noted them here to illustrate the breadth of contradictions that our critic models revealed.

Overall, class-wise differences in terms of the distribution of contradiction types were modest. Disinformation showed slightly more background fact and missing-event gaps, while trustworthy news contained slightly more role and causal gaps. However, in both cases *Other* category remained dominant. Given the heterogeneity and relatively small absolute counts of non-residual categories, we reported these class differences qualitatively and left a more fine-grained typology of contradictions to future work.

Framing inconsistencies were the most structured critic findings, as each issue was annotated with a discrete *type*. We observed five main types: *causal_flip* (narrative element extractions reversed or omitted the causal direction implied by the combination of article summary, article-level KG and metadata), *role_flip* (the extracted source/target or actor/patient roles differed from those supported by the input evidence), *sentiment_flip* (the polarity or evaluation of an actor or event in the extraction conflicted with the sentiment expressed in the inputs), *temporal_conflict* (inconsistent temporal ordering relative to the timeline suggested by the inputs) and *modality_shift* (changes in certainty, speculation or obligation compared to how the situation was framed in the inputs). Table 20 shows the distribution of these types by critic.

Table 20. Types of framing inconsistencies flagged by the critics (row-wise percentages, aggregated over classes). Source: authors' contribution.

Critic	Causal_Flip	Role_Flip	Sentiment_Flip	Temporal_Conflict	Modality_Shift
Primary (Mistral)	49.3	36.5	5.0	9.2	0.0
Verifier (Gemma)	40.9	4.6	50.7	0.9	2.9

The primary (Mistral) critic focused mainly on *causal_flip* and *role_flip* inconsistencies, as about 49.3% of its framing issues concerned causal direction and 36.5%—who was portrayed as acting upon whom. Temporal conflicts accounted for 9.2%, while sentiment flips were relatively rare (5.0%). In contrast, the verifier (Gemma) critic was dominated by *sentiment_flip* issues, which accounted for 50.7% of its framing inconsistencies. In second place in terms of frequency was *causal_flip* (40.9%), while *role_flip* (4.6%), *temporal_conflict* (0.9%) and *modality_shift* (2.9%) were relatively rare. These complementary profiles suggested that the primary critic was more sensitive to structural narrative roles and causal framing, while the verifier critic was more attuned to evaluative and affective

framing. Again, disinformation and trustworthy news showed qualitatively similar distributions within each critic, with only minor shifts in the relative frequencies of causal vs. sentiment-oriented inconsistencies.

To summarize, these patterns showed that in our critic ensemble, both models specialized in different aspects of narrative quality. On average, each critic flagged approximately five missing-link suggestions per article (4.84–5.00 across classes), making missing links the largest issue category compared to contradictions (1.08–1.58) and framing inconsistencies (1.58–1.71) (Table 6). The primary critic highlighted structural contradictions and framing inconsistencies around causality and roles, while the verifier critic provided a complementary focus on evaluative and affective framing. Class effects (disinformation vs. trustworthy) were visible but subtle and much smaller than the differences between critic profiles themselves.

4.2. Mapping Extracted Narrative Elements to GDELT

Roughly 97% of our extracted actors and events had a corresponding mention in GDELT (most often via GKG actor fields rather than the events table). We observed this for both disinformation and trustworthy news (Table 21). Such a high rate suggested that our pipeline rarely introduced idiosyncratic entities, but focused on people, organizations and places that also appeared in global news. Also, GDELT support rates were almost identical for our two classes (0.965 vs. 0.977; Table 21), suggesting largely overlapping actor and event contexts. This meant that the main differences were not in their specificity, but in how these shared actors and events were causally linked and rhetorically presented.

Table 21. GDELT support for fused actors and events: number of articles (*N*), count with at least one supporting GDELT hit (supported), count with no hit (not found), and support rate per class. Source: authors' contribution.

Class	<i>N</i>	Supported	Not Found	Support Rate
Disinformation	308	297	11	0.965
Trustworthy news	302	295	7	0.977
Overall	610	592	18	0.971

Furthermore, the directional interaction proxy revealed a clear difference between disinformation and trustworthy news that we did not observe at the level of actors and events alone. For each article, we derived ordered actor pairs ($A \rightarrow B$) from the extracted causal links by linking actors in the source event to actors in the target event and then queried GDELT for events within ± 7 days of the article's publication date whose *Actor1Name* refers to *A* and *Actor2Name* refers to *B* (after simple normalization). As shown in Table 22, such directional pairs were supported for about 60% of trustworthy news but only about 45% of disinformation. In other words, causal relations in trustworthy news more often connected actors in ways that also appeared as directed interactions in global news. We treated this only as a coarse alignment signal, not as validation. Meanwhile, a larger share of disinformation causal links pointed to actor–actor relations that GDELT did not see in that direction. We did not interpret this as GDELT providing “true” causal labels, but as an indication that actor pairs implied by extracted causal links were less frequently observed as directed interactions in GDELT within the same time window.

Table 22. GDELT support for causal links, using directional actor-pair matches (Actor1 → Actor2 within ± 7 days) as a proxy. N_{eval} is the number of articles, *Supported* counts articles with at least one matched directional pair, and *Support rate* is the proportion of such articles per class. Source: authors' contribution.

Class	N_{eval}	Supported	Not Supported	Support Rate
Disinformation	308	137	171	0.445
Trustworthy news	302	182	120	0.602
Overall	610	319	291	0.522

Across all three framing signals (event-tone framing, Goldstein polarity and event-type framing), GDELT was broadly consistent with disinformation and trustworthy news appearing in a similar conflict-centric news context. Therefore, their frames were difficult to distinguish. First, event-tone framing showed that, among articles with at least one matched GDELT event (approximately 72% of disinformation and 87% of trustworthy news), almost all matched events were labeled as having negative tone (about 96% in both classes), with virtually no positive-tone events. This reflected conflict-related topics in our data rather than class-specific framing differences (Table 23).

Table 23. GDELT event-tone framing: number of articles (N), number with at least one matched event (*Covered*), coverage rate, and counts/rates of negatively and positively toned events among covered articles (within ± 7 days). Source: authors' contribution.

Class	N	Covered	Coverage Rate	Neg. n	Neg. Rate	Pos. n	Pos. Rate
Disinformation	308	223	0.724	215	0.964	0	0.000
Trustworthy news	302	263	0.871	253	0.962	0	0.000
Overall	610	486	0.796	468	0.963	0	0.000

In contrast, Goldstein polarity produced almost opposite results as it was extremely positive for the same matched events (around 98% for both disinformation and trustworthy news). This result highlighted that Goldstein polarity captured the expected impact (e.g., cooperation vs. conflict) rather than surface sentiment and again provided nearly identical results for both classes (Table 24).

Table 24. GDELT Goldstein polarity framing: number of articles (N), number with at least one matched event (*Covered*), coverage rate, and counts/rates of strongly positive and negative Goldstein events among covered articles (within ± 7 days). Source: authors' contribution.

Class	N	Covered	Coverage Rate	Neg. n	Neg. Rate	Pos. n	Pos. Rate
Disinformation	308	223	0.724	0	0.000	218	0.976
Trustworthy news	302	266	0.881	0	0.000	260	0.977
Overall	610	489	0.802	0	0.000	478	0.976

Finally, grouping *EventRootCode* items into broad buckets (cooperation, verbal conflict, material conflict, neutral/other) showed similar average per-article shares across classes with roughly 18% of cooperation, 15% of verbal conflict, 14% of material conflict and over 50% of neutral/other for both disinformation and trustworthy news (Table 25).

Table 25. GDELT event-type framing via *EventRootCode*. For each class, *Covered* is the number of articles with at least one matched GDELT event (within ± 7 days), and the shares are the average per-article proportions of events mapped to cooperation, verbal conflict, material conflict, and neutral/other buckets. Source: authors' contribution.

Class	Covered	Coop Share	Verbal Share	Material Share	Neutral Share
Disinformation	223	0.182	0.146	0.143	0.529
Trustworthy news	266	0.179	0.160	0.141	0.520
Overall	489	0.181	0.152	0.142	0.525

To summarize, these results indicated that GDELT's framing-related attributes were too coarse and homogeneous across classes for our fine-grained narrative frames. They confirmed that both types of articles had comparably negative, conflict-centric news contexts, but did not capture subtler framing differences that our narrative extraction was designed to model.

4.3. Evaluation Against Expert-Judged Subset

4.3.1. Precision-Based Analysis

We assessed the quality of the extracted narrative structures on a stratified expert-annotated subset, which consisted of 60 articles (30 disinformation and 30 trustworthy news). For each article, the expert inspected individual system outputs (characters, events, causal links, framing edges and critic findings) together with their supporting evidence passages. Then the expert assigned a part-credit score in $\{0, 0.5, 1\}$ according to the guidelines we described in Section 3.2.4. This score reflected how well the prediction matched what the quoted passage actually stated (for narrative elements) or how accurately the critic ensemble described a problem in the extracted narrative representation. From these individual scores, we computed micro part-credit precision (aggregating over all annotated items) and macro part-credit precision (averaging per-article means and giving each article equal weight). The summary of results is presented in Table 26.

Table 26. Part-credit precision of the narrative element extractions by class and element type on the expert-annotated subset (60 articles). Source: authors' contribution.

Class	Element Type	Micro Precision	Macro Precision	#Elements
Disinformation	Characters	0.989	0.989	87
Trustworthy news	Characters	1.000	1.000	86
Overall	Characters	0.994	0.994	173
Disinformation	Events	0.966	0.967	88
Trustworthy news	Events	0.972	0.972	88
Overall	Events	0.969	0.969	176
Disinformation	Causal links	0.764	0.768	70
Trustworthy news	Causal links	0.780	0.764	75
Overall	Causal links	0.772	0.766	145
Disinformation	Framing edges	0.846	0.814	52
Trustworthy news	Framing edges	0.845	0.822	58
Overall	Framing edges	0.845	0.818	110

The precision scores showed that narrative extractor was highly precise for core elements (characters/actors and events). For characters, overall micro and macro precision were both 0.994, while disinformation reached 0.989 (micro 0.9885 and macro 0.9889) and trustworthy news—1.000 (both micro and macro). As for events, overall micro and macro

precision were both 0.969. Disinformation achieved micro and macro precision of 0.966 and 0.967, respectively, while trustworthy news achieved 0.972 for both. Causal links were more challenging but still achieved overall micro/macro precision of 0.772/0.766. Disinformation achieved 0.764/0.768 micro/macro precision and trustworthy news—0.780/0.764. For framing edges, overall micro/macro precision was 0.845/0.818, with very similar class-specific values (0.846/0.814 micro/macro precision for disinformation and 0.845/0.822—for trustworthy news).

Our critic component, which identified missing links, contradictions and framing inconsistencies in extractions of narrative elements, attained an overall micro part-credit precision of 0.741 and macro precision of 0.735 (Table 27). Broken down by class and issue type, critic precision was the highest for missing links and contradictions (typically in the 0.73–0.82 range) and lower for framing inconsistencies (approximately 0.65–0.72), with consistently higher scores for trustworthy articles than for disinformation (Table 28), indicating that our critic provided non-trivial diagnostic coverage (micro precision 0.741) but lower reliability than the narrative extractor, particularly for framing inconsistencies.

Finally, Table 29 shows that critic precision depended both on issue type and on which model acted as critic, i.e., for contradictions, the primary critic is slightly more reliable in both disinformation and trustworthy articles (especially strong on trustworthy news), whereas for framing inconsistencies and missing links, the verifier critic achieved higher micro precision for framing inconsistencies and missing links, particularly on disinformation (e.g., 0.861 vs. 0.550 for framing inconsistencies). Overall, both sides achieved micro part-credit precision above 0.80 on missing links in trustworthy articles, but framing inconsistencies remained the noisiest category across classes.

Table 27. Overall part-credit precision of the critic component on the expert-annotated subset. Source: authors' contribution.

	Micro Precision	Macro Precision	#Elements
Overall critic	0.741	0.735	353

Table 28. Micro part-credit precision of the critic component by article class and issue type. Source: authors' contribution.

Class	Issue Type	Micro Precision
Disinformation	Missing links	0.727
	Contradictions	0.713
	Framing inconsistencies	0.647
Trustworthy news	Missing links	0.817
	Contradictions	0.816
	Framing inconsistencies	0.724

Table 29. Micro part-credit precision of the critic component by class, issue type and model side (primary vs. verifier). Source: authors' contribution.

Class	Issue Type	Side	#Elements	Micro Precision
Disinformation	Contradictions	Primary	37	0.7297
	Contradictions	Verifier	17	0.6765
Trustworthy news	Contradictions	Primary	34	0.8529
	Contradictions	Verifier	23	0.7609
Disinformation	Framing inconsistencies	Primary	40	0.5500
	Framing inconsistencies	Verifier	18	0.8611
Trustworthy news	Framing inconsistencies	Primary	35	0.6714
	Framing inconsistencies	Verifier	23	0.8043
Disinformation	Missing links	Primary	30	0.6500
	Missing links	Verifier	36	0.7917
Trustworthy news	Missing links	Primary	29	0.8276
	Missing links	Verifier	31	0.8065

4.3.2. Error Analysis

Characters/actors. Expert judgment over the character sample produced 171 fully correct extractions (score 1.0) and two partially correct extractions (score 0.5). No character item scored as a completely incorrect item (score 0). Both partially correct cases concerned character specification rather than hallucination (detailed examples are provided in Table A5 in Appendix G.2). In the first partially correct case, our extractor marked *Germany* as a unitary actor in a context where the article referred more narrowly to specific German institutions, which the expert judged as an over-generalized but text-grounded actor. In the second partially correct case, the extractor extracted *the state*, which was clearly discussed in the article but remained too vague to count as a well-formed, narrative-level actor.

Events. The results indicated that the majority of extracted events were judged to be correct (94.89%) and only 5.11% of events had errors (partial or full). The overall distribution of scores is shown in Table A6 in Appendix G.2. Since our extractor ensemble was instructed to support its event decisions, where it was possible, with KG triples, we also examined the relationship between correctness and grounding. Grounded events constituted approximately two-thirds of the sample, used for expert evaluation (116/176), and attained a similarly high level of correctness as non-grounded events (details are presented in Table A7 in Appendix G.2). There were both partially correct and incorrect events among grounded instances, which suggested that the interaction between the LLM and the KG could occasionally lead to over-specific or misaligned event formulations.

To better understand the nature of these errors, we qualitatively analyzed the expert explanations for the events that were partially or fully incorrect (nine instances in total). We derived four recurring error types (detailed summary is presented in Table A8 in Appendix G.2). We summarized the recurring error types without detailed case studies. Most errors involved subtle semantic issues rather than completely hallucinated events. Broadly, these error types captured cases where event labels overstated the strength or scope of the source text, under-specified key distinctions or drifted from the article's framing or contextual constraints despite being topically related.

Causal links. We observed that individual models and consensus ensemble had different error tendencies (details and examples are provided in Tables A9 and A10 in Appendix G.2). For Gemma, the overall error rate (score 0 or 0.5) was around 42% (28/66). The dominant issues were that identified events were usually correct, but causality was overstated or too weakly supported, especially for *leads_to* and *enables* causal links. Many

of these links receive a score of 0.5, where the expert accepted the events but downgraded the causal strength. However, fully incorrect items were relatively rare (12%) in comparison to Mistral. For the latter model, the overall error rate was around 82% (42/51). The main issues were missing or undefined events, which constituted the largest part of errors, followed by explicitly incorrect causal links and noisy “Other” errors. Mistral behaves as a very liberal generator, willing to invent or loosely ground both events and causal links. For the consensus ensemble, the overall error rate was approximately 21% (6/28). The issues were subtle or borderline causality, when events were correct, but the causal link was weak, implicit or interpretative.

For trustworthy news, *causes* edges were particularly reliable (no links scored as completely incorrect), while *enables* edges were noticeably noisier (almost half judged incorrect), which suggested that models tended to over-claim enabling effects in factual reporting. In disinformation articles, the three relation types (*causes*, *enables*, and *leads_to*) were more balanced, although *enables* remained the most error-prone relation type. However, this tendency was weaker than in trustworthy articles.

Framing edges. We analyzed expert judgments for 110 sampled framing edges (details are presented in Tables A11 and A12 in Appendix G.2). Overall, 90 predictions (81.8%) were judged correct, 6 (5.5%) partially correct, and 14 (12.7%) incorrect. This produced an aggregate error rate of 18.2%. Most errors involved selecting an incorrect framing-type, where actor-event relation was supported by the article, but the specific stance label (e.g., *blame* vs. *criticism*) was misassigned. Fewer errors corresponded to unsupported or absent frames, indicating that grounding in article-level KGs substantially reduced outright hallucinations but did not fully resolve fine-grained semantic distinctions.

In addition, error rates were evenly distributed across source models (Gemma vs. Mistral) and classes (disinformation vs. trustworthy). Consistent with our confidence-tier design, errors were concentrated in lower-confidence predictions. Our sample contained no Tier A (consensus) framing edges, while Tier C items (single-model ungrounded or disagreement cases) accounted for the majority of errors, compared to fewer errors in Tier B (single-model grounded) cases.

Critic. For critic, 256/353 items (72.5%) were judged as fully correct, 11/353 (3.1%) as partially correct and 86/353 (24.4%) as incorrect (details of error analysis are displayed in Tables A13 and A14 in Appendix G.2). Among the critic issues, *framing inconsistencies* were the most challenging as only 66.4% of framing-related critic items were judged fully correct, with 29.3% marked as incorrect. In contrast, *missing links* and *contradictions* achieved slightly higher correctness (76.2% and 74.8%, respectively) and lower error rates (22.2% and 21.6%). This pattern was consistent with the intuition that fine-grained framing judgments were harder to identify than detecting omitted links or explicit contradictions.

We also examined errors by class (disinformation vs. trustworthy news). Across all critic categories, critic items attached to trustworthy news were more often judged correct and less often judged incorrect. Therefore, disinformation achieved 121/178 fully correct items (68.0%) and 51/178 incorrect ones (28.7%), while trustworthy news reached 135/175 fully correct (77.1%) items and 35/175 incorrect ones (20.0%). Within each issue type, we observed the same trend, which suggested that critic models were somewhat more reliable on trustworthy news, while they tended to produce more false positives on disinformation content.

Qualitative inspection of expert explanations revealed distinct error profiles for each issue type. For *missing links*, most incorrect cases (score 0) occurred when the critic claimed a causal or framing relation was absent, even though it was already captured in the extracted narrative elements, or when it inferred a relation that was more speculative than the passage of evidence text supported. For *contradictions*, incorrect items typically occurred when our critic ensemble alleged a conflict that was not evidenced in the quoted

segment (e.g., only one side was present, or the supposed opposing statement was missing). As for *framing inconsistencies*, we found that errors often resulted from over-interpreting neutral language as strongly evaluative, thus treating minor shifts in emphasis as genuine inconsistencies, or mislabeling cases better characterized as missing links or omissions. Partially correct items (score 0.5) usually reflected correct detection of a problem but an imprecise characterization of its type or severity, e.g., identifying a missing nuance but labeling it as a full inconsistency, or treating a reasonably supported additional relation as a hard error rather than an optional enrichment.

5. Discussion

This section provides an interpretation of the empirical results related to our study's methodological objectives while addressing the broader literature on narrative extraction, knowledge graph grounding and ensemble-based LLM systems. Thus, Section 5.1 describes the main findings and reflects on how the proposed pipeline balances robust extraction of characters and events with more tentative modeling of causal and framing relations. Section 5.2 discusses practical considerations related to transferability and implementation of using our approach in applied analytical settings. In Section 5.3, we outline the study's limitations and directions for future research, including aspects related to methodology, computational costs, and evaluation.

5.1. Main Findings

Our comparison of union and consensus ensembles reflects earlier research on knowledge base population and ensemble information extraction, where unions of systems usually maximize recall and intersections or more conservative aggregation schemes produce high-precision subsets at the expense of coverage [100,101]. This traditional recall-precision trade-off was reproduced in our setting by the dense union graphs and sparse consensus graphs in a structured narrative context. While causal and framing edges nearly collapsed under consensus, characters and events essentially stayed the same between the two variants. Additionally, both our ensembles used the same triple identifiers and AMR-FRED node IRIs. This approach followed a recent tendency in Text-to-KG research, which has argued that AMR-based RDF graphs provide logically coherent, interpretable backbones for downstream reasoning [73,102,103]. This is consistent with previous research showing that relation extraction, particularly causal and discourse-level structure, is more model-dependent and prone to errors, whereas entity and event detection tends to be more stable across models [104].

Meanwhile, our pipeline's GDELT matches functioned as an external validity check. Approximately 97% of fused actors and events had at least one GDELT hit for both disinformation and trustworthy news. This indicated that the extracted narrative elements corresponded to people, organizations and places that appeared in global news streams rather than being idiosyncratic entities [96,97,105]. This observation was consistent with research findings showing that GDELT offers a fairly comprehensive, if noisy, map of major news agendas [105,106]. From the perspective of mis- and disinformation studies, these results, again, supported the view that misleading content often built on the same factual event backbone as mainstream coverage but differed in how those shared events were sequenced and interpreted [107–109].

The directional actor-pair analysis produced a clearer signal between disinformation and trustworthy news as a proxy for causal relationships. A larger proportion of disinformation links involved actor–actor relations that were not recorded in the same direction in GDELT, even though the underlying actors and events were similarly grounded. On the other hand, causal links in trustworthy news were much more likely to connect actors in

ways that also appeared as directed interactions in GDELT. This pattern was consistent with research that argued that disinformation frequently deviated from mainstream accounts by reassigning causal responsibility, exaggerating or downplaying specific ties and reconfiguring evaluative frames around similar actors and events [110–112], even though we did not treat GDELT as a causal ground truth.

The overall picture of a shared, conflict-heavy event environment for both classes was largely confirmed by the framing-related GDELT attributes. This was reflected in the nearly universal negative tone and consistently high Goldstein scores observed for matched events. Similarly, the distribution of `EventRootCode` buckets (cooperation vs. verbal/material conflict vs. neutral) was nearly the same across classes. It was useful as a background signal to confirm that our extracted narratives inhabited the same macro-level event space as global news and provide context to assess how closely their causal links align with mainstream interaction patterns.

Finally, the expert-annotated subset demonstrated that our KG-grounded narrative extractor consistently performed well for both disinformation and trustworthy news, achieving very high precision for core elements (characters and events) with overall part-credit scores around 0.97–0.99. Entity and event identification are relatively mature information extraction (IE) tasks, and modern neural and LLM-based systems typically achieve high precision once type inventories are well defined and contextual cues are clear, while greater challenges arise when these units are linked [113–115]. From this angle, our findings implied that, regardless of article class, combining LLMs with article-level KGs produced an extractor that is nearly as reliable as an expert for determining “who” and “what happened” in the news.

Conversely, greater difficulty of modeling narrative relations was reflected in the lower but still strong precision for framing edges (overall ≈ 0.84) and causal links (overall ≈ 0.77). Because they frequently rely on subtle discourse cues and world knowledge, causal and other event–event relations continue to be among the most difficult IE targets [113,114,116]. Even in fully supervised settings, relation classifiers and narrative-structure components were less accurate than event detection, according to recent research on event-centric news narratives and narrative chains [117,118]. Consistent with these findings, our results indicate that although narrative relationships can be modeled fairly reliably, they remain noisier than object and event nodes, particularly when extracted from full news articles rather than short sentences.

Additionally, the critic component achieved an overall precision of roughly 0.74 when identifying missing links and contradictions. Studies evaluating LLM-based IE and data-extraction assistants in applied domains (e.g., systematic reviews or key information extraction from documents) typically report accuracy levels in the 70–80% range [30,119]. Similar behavior was demonstrated by our critic, which successfully identified a large number of missing links and contradictions. These trends align with recent studies that support the use of nuanced metrics and human-in-the-loop validation, especially for higher-level semantic properties [120].

Taken together, the expert evaluation supports a functional division of labor within the pipeline, where core narrative units can be extracted with near-expert reliability, while causal and framing relations benefit from aggregation, cross-checking and selective expert inspection. This is consistent with research on public discourse and narrative extraction, which also found that causal relationships and framing choices were the most vulnerable and informative elements of automatically generated narrative representations [115,121]. To summarize, our results suggested a division of labor, where automatically extracted characters and events could be used with high accuracy, while causal links, framing edges and critic findings were best handled as structured hypotheses to be aggregated, cross-checked, and selectively inspected rather than treated as unfiltered ground truth.

While individual components have been studied previously, we combined AMR-FRED grounding with role-specialized LLM mini-ensembles for narrative extraction in a unified pipeline and systematically evaluated the resulting precision–coverage trade-offs. We quantified the effects of ensemble choice on narrative elements and distinguished a recall-oriented union view from a high-precision consensus skeleton over the same set of grounded elements. Our findings demonstrated that while disinformation and trustworthy news articles were comparably grounded in article-level KGs, they differed mainly in the density and configuration of causal and framing edges that were layered on top of this common event backbone. This structurally supported claims made in the misinformation literature that misleading content often repurposes widely shared factual cores while reconfiguring causal attributions, responsibilities and evaluative framing instead of creating entirely new event worlds [112,122–124]. In this way, our KG-anchored union and consensus graphs provided a concrete, inspectable example of how narrative causality and framing differed between disinformation and trustworthy news.

5.2. Transferability and Implementation Considerations

Although the proposed pipeline is research-oriented, several aspects of the design support its use in applied settings. The narrative extraction components for characters and events achieved precision comparable to expert annotation in our evaluation, indicating that they can serve as reliable inputs for downstream analysis with limited manual verification. These components could support tasks such as actor tracking, event indexing, timeline construction, or exploratory corpus analysis within semi-automated workflows.

On the other hand, causal links, framing edges and critic outputs are intentionally treated as structured hypotheses rather than definitive facts. This design aligns well with semi-automated workflows in journalism, policy analysis and disinformation research, where analysts benefit from candidate relations that can be inspected, aggregated or filtered rather than automatically trusted. In this sense, the system is best understood as a decision-support tool rather than a fully autonomous narrative interpreter. Depending on computational constraints and task requirements, components such as AMR-FRED grounding, ensemble aggregation, the critic module, and GDELT alignment may be included or excluded. For example, lightweight configurations may rely only on single-model extraction and event grounding, while more demanding analytical settings can incorporate ensembles and critic-assisted inspection.

Finally, despite relying on multiple LLM calls, the overall computational cost remained modest, indicating that the approach is feasible for medium-scale analytical use without specialized hardware. These properties support the claim that the contribution is not only methodological, but also applicable in analytical settings that require interpretable, KG-grounded narrative analysis with human oversight.

5.3. Limitations and Future Work

Despite the promising results, several limitations of the proposed approach should be acknowledged. They also point toward concrete directions for future work.

L1. Extractive summaries vs. full text

In this study, we deliberately relied on extractive summaries rather than full article texts as input to the narrative extraction pipeline. While summaries inevitably exclude some discourse-level nuance, this choice reflects a trade-off between textual coverage and the feasibility of constructing interpretable article-level representations.

Extractive summarization provided us with a practical compromise by preserving original wording and core factual content while substantially reducing text length, thereby enabling AMR parsing [28] and KG construction. This design choice allowed the pipeline to

prioritize semantic depth and interpretability of the extracted narratives over exhaustive textual coverage. Future work could explore selective full-text parsing, hybrid summary–full-text strategies, or the integration of discourse-aware representations alongside KG grounding to better capture long-range discourse structure.

L2. Language dependence and semantic preprocessing cost of AMR–FRED grounding

Article-level KGs in this study were constructed using the Text2AMR2FRED [73] pipeline, which relies on deep semantic parsing via abstract meaning representation (AMR) [28] followed by ontology-driven RDF generation via the FRED [29] framework. While this approach produces rich, event-centric and interpretable semantic structures, it introduces a substantial preprocessing burden that is largely independent of the downstream LLM components.

In particular, AMR parsing and ontology-based grounding are computationally and linguistically demanding, and currently mature primarily for English. Extending this pipeline to other languages would require reliable multilingual AMR parsers, cross-lingual sense inventories and language-independent ontology alignment. Unlike syntactic frameworks such as universal dependencies, efforts toward universal meaning representation are currently at an early stage.

As a result, while the proposed pipeline remains relatively inexpensive in terms of LLM inference, the main computational constraints arise from semantic preprocessing. This shifts the primary bottleneck from model inference to representation construction. Future work could therefore explore lighter-weight or multilingual alternatives to AMR–FRED grounding, hybrid symbolic–neural representations or selective KG construction strategies that trade some semantic richness for improved scalability.

L3. Ensemble aggregation trade-offs

The use of union and consensus aggregation strategies reflects a deliberate precision–recall trade-off. While union graphs maximize coverage and hypothesis generation, consensus graphs provide high-precision narrative skeletons but at the cost of increased sparsity, especially for causal and framing edges. We plan to explore more adaptive aggregation strategies, including confidence-weighted voting or task-specific ensemble policies (e.g., recall-oriented causality vs. precision-oriented framing), in the future.

L4. Critic reliability and LLM-as-judge limitations

The critic component relies on LLMs acting as automated judges. Although effective in identifying missing links and contradictions, the critic was also prone to incorrect critiques, inconsistent reasoning, and sensitivity to prompt formulation. For this reason, critic outputs are best interpreted as structured hypotheses rather than definitive errors. In the future, we plan to integrate symbolic validators, constraint-based checks or human-in-the-loop verification to improve reliability. Prior evaluations of LLM-based information extraction assistants report accuracy levels in the 70–80% range and emphasize their role as decision-support tools rather than replacements for human judgment [30,119].

L5. Domain and language scope

The present study focuses exclusively on English-language news articles within a limited geopolitical domain (e.g., Russia–Ukraine conflict, NATO relations). This restricts generalizability to other languages, cultural contexts and narrative traditions. Extending the pipeline to multilingual settings will require language-specific models, prompts and potentially different grounding resources, as well as systematic evaluation across domains.

Without such adaptations, precision for character and event extraction is likely to decrease due to morphological variation, causal link reliability may be affected by language-

specific discourse markers, and framing edges may become more sensitive to culturally grounded rhetorical patterns.

L6. Scope of external validation

We used GDELT as an external reference to assess the presence and directional alignment of actors and events in global news coverage. However, GDELT does not encode fine-grained narrative structures or causal semantics and therefore cannot serve as a ground truth for narrative correctness. Future work could explore complementary external resources, such as domain-specific knowledge bases, media-bias or framing lexicons, expert-curated datasets or cross-corpus narrative alignment. Existing research has shown that while GDELT provides broad coverage of protest, conflict and disaster events, it remains relatively coarse with respect to subtle framing and narrative distinctions [96,105].

L7. Computational cost

The proposed pipeline relies on multiple LLMs, aggregated multi-model outputs and post-hoc validation, and would be expected to be more computationally demanding than single-model approaches. In practice, however, the computational footprint remained modest. Across all experiments, including exploratory runs, the pipeline issued approximately 16,000 LLM inference requests and processed roughly 60 million tokens in total, resulting in an overall cost of approximately \$20 using API-based access to open-weight instruction-tuned models.

These figures indicate that, despite the use of ensembles and a critic component, the approach is feasible for medium-scale research settings and does not require specialized hardware. Nevertheless, the design prioritizes robustness and interpretability over efficiency. Our future work includes exploring routing strategies and smaller specialized models to further reduce computational cost while preserving narrative quality.

6. Conclusions

In this work we introduced a practical pipeline for KG-grounded narrative extraction that combines AMR-FRED knowledge graphs with role-specialized mini-ensembles of large language models and a dedicated critic component. The goal was to produce structured and inspectable narrative representations that connect characters, events, causal relations and framing in shared semantic structures.

Across expert evaluation and external validation, the pipeline was reliable for core narrative elements and revealed meaningful differences between disinformation and trustworthy news, particularly in terms of causal and framing structure. By providing both high-recall and high-precision views of the same grounded narratives, the pipeline supports exploratory analysis as well as more conservative structural comparison.

Overall, the main contribution of this work is methodological. It shows that combining knowledge graph grounding, small LLM ensembles and post-hoc validation into a single workflow can support robust narrative extraction without sacrificing transparency or interpretability. Anchoring narrative differences in explicit semantic structures provides a reusable methodological foundation for future work on disinformation, narrative dynamics, and information manipulation across domains. Finally, the entire experimental pipeline was run with a modest computational budget, demonstrating that KG-grounded, ensemble-based narrative extraction is feasible without large-scale infrastructure.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16041962/s1>, File S1: URLs and metadata of news articles used in this study.

Author Contributions: Conceptualization, J.M. and T.K.; methodology, J.M. and T.K.; software, J.M. and T.K.; validation, J.M. and T.K.; formal analysis, J.M. and T.K.; investigation, J.M. and T.K.; resources, J.M. and T.K.; data curation, J.M. and T.K.; writing—original draft preparation, J.M. and T.K.; writing—review and editing, J.M. and T.K.; visualization, J.M. and T.K.; supervision, J.M. and T.K.; project administration, J.M. and T.K.; funding acquisition, J.M. and T.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research Council of Lithuania (LMTLT), grant agreement No. S-PD-24-88.

Data Availability Statement: This study used a publicly available base dataset consisting of metadata from the EUvsDisinfo dataset, released under the CC BY 4.0 license. The base dataset includes article identifiers, URLs, publishers, publication dates, language labels, topic keywords, and class labels, but does not include full news article texts. Due to copyright restrictions, the full article texts cannot be redistributed. For this study, article texts were reconstructed at the time of analysis from the URLs provided in the base EUvsDisinfo metadata using the Diffbot Article API and cached locally for downstream processing. Because original URLs may change, be removed, or be updated over time, exact reconstruction of the article texts cannot be guaranteed. However, the described reconstruction protocol enables partial replication of the dataset. To support reproducibility, the Supplementary Materials provide a CSV file listing all article URLs included in the final dataset, together with associated publication metadata (publisher, domain, title, author, class label, and publication date). Reconstructed article texts, extractive summaries, article-level knowledge graphs, and derived narrative annotations produced in this study are text-derived research artifacts that include lexicalized or text-aligned representations originating from copyrighted source materials and therefore cannot be made publicly available. Prompt templates and additional methodological details are provided in Appendixes D and E.

Acknowledgments: The authors thank the reviewers for their constructive feedback, which helped to improve this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMR	Abstract Meaning Representation
BERT	Bidirectional Encoder Representations from Transformers
DRS	Discourse Representation Structure
GCN	Graph Convolutional Network
GDELT	Global Database of Events, Language, and Tone
GKG	GDELT Global Knowledge Graph
GNN	Graph Neural Network
IE	Information Retrieval
IQR	Interquartile Range
KG	Knowledge Graph
LLM	Large Language Model
NER	Named Entity Recognition
OWL	Web Ontology Language
RAG	Retrieval-Augmented Generation
RDF	Resource Description Framework
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
TF-IDF	Term Frequency-Inverse Document Frequency
WSD	Word-Sense Disambiguation

Appendix A. Examples of News Article Summaries

In this section, a few short fragments from the extractive news article summaries are presented for illustrative purposes together with their ROUGE scores. Full extractive summaries are not reproduced due to their substantial overlap with the original news article text. The summary extracted with the method that achieved higher ROUGE score, which reflects higher overlap with original text, was used in the experiments (Table A1).

Table A1. Short illustrative excerpts from extractive summaries. Ellipses indicate omitted text. Source: authors' contribution.

Title	LexRank Summary	BERT Extractive Summary	ROUGE (LexRank)	ROUGE (BERT)
EU supports Zelensky's peace plan–Michel https://en.interfax.com.ua/news/general/885670.html , accessed on 10 January 2025	"...support for Ukraine is the first issue..." "...damages caused by Russia to Ukraine..." "...Europeans have understood that this is more than an attack..."	"...supports the Peace Plan proposed by President Volodymyr Zelensky..." "...damages caused by Russia to Ukraine must be compensated..." "...EUR 300 billion of Russian assets should be used for reconstruction..." "...more than an attack on Ukraine..."	0.291	0.292
International Space Station preparing for 'space taxis' https://www.cnn.com/2015/02/20/international-space-station-preparing-for-space-taxis.html , accessed on 10 January 2025	"...prepare the International Space Station for two new docks..." "...Boeing's CST and SpaceX's Dragon spacecraft..." "...setting up cables on the first two walks..." "...first time vessels have docked on the American side..." "...\$4.2 billion contract to develop a transportation system..."	"...pacewalk outside the ISS was postponed..." "...analysis of spacesuits..." "...two new docks for the spacecraft..." "...cargo mission..." "...first time vessels have docked on the American side..."	0.622	0.59
The Maduro Diet: Food v. Freedom in Venezuela https://www.csis.org/analysis/maduro-diet-food-v-freedom-venezuela , 10 January 2025	"...unable to control inflation..." "...disintegrated Venezuela's economy..." "...dependent on CLAP deliveries..." "...politically motivated food system..." "...90% of CLAP boxes come from Mexico..."	"...Venezuela's humanitarian crisis deepens..." "...tighten political control..." "...poverty has risen to nearly 90%..." "...90% of CLAP boxes come from Mexico..." "...weaponization of the CLAP program..."	0.294	0.212

Appendix B. AMR Example

Figure A1 presents an exemplary sentence "The dog wants to go" parsed to AMR [28] representation. This AMR representation encodes two events, a *wanting* and a *going*, linked by a shared participant. The main predicate is the instance *z0/want-01*, where *want-01* is the PropBank [69] sense which denotes a wanting or desiring event. Its argument :ARG0 is *z1/dog*, indicating that the dog fills the ARG0 role of *want-01*, i.e., the *wanter*. The argument :ARG1 of *z0* is another event (*z2/go-02*), which captures the fact that the thing wanted is an eventuality, namely the *going* event. For *go-02*, the ARG0 role corresponds to the *goer*, and this role is again filled by *z1*, the same dog. Thus, the AMR graph expressed

clausal embedding (the wanting event takes the going event as its ARG1 “thing wanted”) and subject control (the dog is both the *wanter* of want-01 and the *goer* of go-02) by reusing the variable z1 in both argument positions.

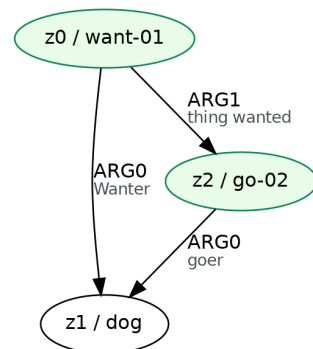


Figure A1. AMR [28] example—parsed sentence “The dog wants to go”. This AMR was generated using Text2AMR2FRED WebApp [125] for illustration only. Source: authors’ contribution.

Appendix C. Text2AMR2FRED

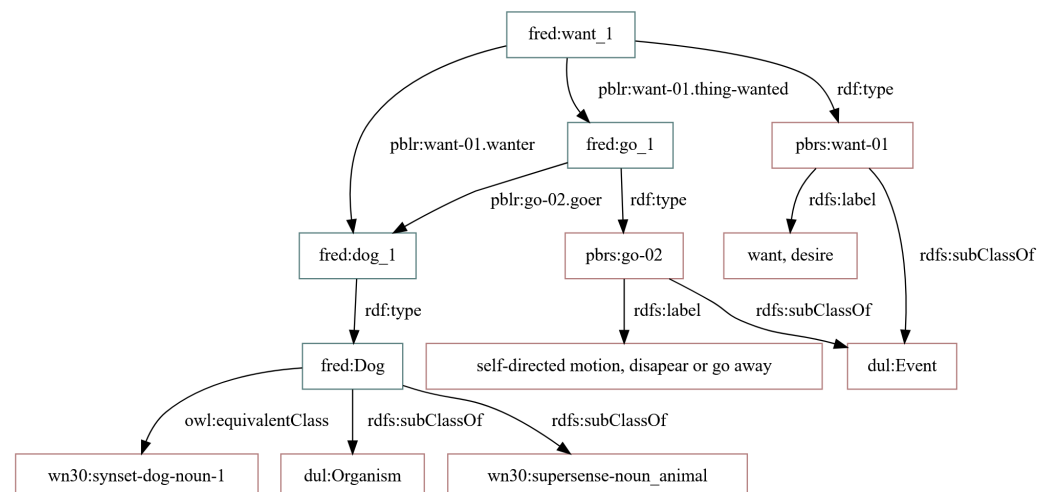


Figure A2. Text2AMR2FRED [73] example for the sentence “The dog wants to go”. This representation was generated using Text2AMR2FRED WebApp [125] for illustration only. Source: authors’ contribution.

Figure A2 presents an example, where an exemplary sentence ‘The dog wants to go’ was processed via Text2AMR2FRED [73] pipeline. The sentence was first converted into an AMR [28] graph that captured its underlying predicate-argument structure—a wanting event whose experiencer (the dog) also controlled an embedded going event. FRED [29] then mapped these AMR predicates to sense-specific PropBank [69] rolesets (want-01 and go-02), creating corresponding FRED [29] individuals for the events (want_1, go_1) and for the participant (dog_1). The semantic roles were made explicit using PropBank [69] local roles as OWL object properties, so that dog_1 was linked to want_1 via want-01.wanter, to go_1—via go-02.goer and want_1 was linked to go_1 via want-01.thing-wanted. At the same time, the participant dog_1 was typed as an instance of the class Dog, which was aligned with a specific WordNet [78] synset and classified as both a dul:Organism and a noun_animal supersense. In this way, the pipeline turned a short sentence into a richly structured OWL/RDF representation in which events were grounded in frame and sense inventories, arguments were connected by explicit role properties and entities were linked to a background lexical-ontological system.

Appendix D. Prompts

Appendix D.1. System Prompt for Narrative Extraction

You extract narrative elements grounded ONLY in the provided text excerpt and RDF triples. Triples are listed as lines:

TRIPLE_ID | subject_iri - predicate_iri - object_iri.

Use BOTH sources:

- Triples: for node_iri selection and source_triples/evidence_triples.
- Text excerpt: for labels, names, evidence_text, and to infer causal and framing relations.

Grounding rules:

- node_iri must be "" or EXACTLY one TRIPLE_ID from the Triples section.
- source_triples / evidence_triples must be lists of real TRIPLE_IDs.
- Every event, causal link, and framing edge must cite at least one TRIPLE_ID or short textual quote.

Causality guidelines:

- "causes": direct responsibility.
- "enables": creates conditions, cover, opportunity, or motivation.
- "leads_to": indirect/long-term consequences.
- Treat strategic/psychological relations as causal when framed as reasons or calculations.
- Reporting/quoting events are NOT causes unless explicitly described as producing an effect.

Return STRICT JSON only, matching the schema:

```
{
  "narrative": {
    "characters": [...],
    "events": [...],
    "causal_links": [...],
    "framing_edges": [...]
  }
}
```

Appendix D.2. Prompt Template (Per-Article Input)

Article ID: {article_id}

Text excerpt:

"""

{summary}

"""

Triples (TRIPLE_ID | subject_iri - predicate_iri - object_iri):

{triples}

{meta}

Rules:

- Extract characters, events, causal links, and framing edges.
- Always copy node_iri and TRIPLE_IDs exactly.
- Use RDF triples for structure; text for naming and evidence_text.
- Include physical/logistical AND strategic/psychological causal links.
- If text presents A as the rationale for B, use enables or leads_to.
- Reporting verbs are not causal unless stated otherwise.
- Do not invent entities or unsupported events.
- Return JSON only, following the schema.

Appendix D.3. JSON Schema Hint Provided to the Models

```
{
  "narrative": {
    "characters": [
      {"id": "c1", "name": "", "label": "", "type": "",
        "node_iri": "", "mentions": [""]}
    ],
    "events": [
      {"id": "e1", "label": "", "what": "", "when": "", "where": "",
        "time": "", "node_iri": "", "participants": [""],
        "who": [""], "evidence_text": "", "source_triples": [""]}
    ],
    "causal_links": [
      {"source_event": "e1", "target_event": "e2",
        "relation": "causes", "evidence_triples": [""],
        "evidence_text": ""}
    ],
    "framing_edges": [
      {"source": "c1", "target": "e1", "relation": "frames",
        "frame_type": "", "evidence_text": "",
        "source_triples": [""]}
    ]
  }
}
```

Appendix D.4. JSON Validation and Repair Pipeline (Summary)

To ensure that model outputs were always valid, we implemented a four-stage post-processing pipeline:

1. Direct JSON extraction: Detect and parse the first balanced {...} block from model output.
2. LLM-based repair: If parsing fails, the raw output is sent to a repair model instructed to return a single valid JSON object conforming to the schema.
3. Schema normalization: We unify key aliases (e.g., who vs. participants, text_evidence vs. evidence_text), ensure required lists exist, and enforce grounding rules for node_iri and TRIPLE_ID fields.
4. Fallback: If all repair attempts fail, we return a schema-valid empty structure containing all required keys.

This pipeline guarantees that every article produced a structurally valid JSON narrative object suitable for quantitative evaluation, error analysis and cross-model comparison.

Appendix D.5. Entity Linking Prompts and Linking Procedure

Appendix D.5.1. System Prompt for Entity Linking

You link RDF entities to Wikidata. Choose exactly one QID or NONE.

Return strict JSON:

```
{"choice": "QID or NONE", "confidence": 0-1, "justification": "one sentence"}.
```

Appendix D.5.2. Prompt Template (Per-Entity Input)

Entity label: "{label}"

Local context: "{snippet}"

Candidates (label - QID - short note):

```
{cands}
```

Rules:

- Prefer precision over recall.
- If no candidate clearly fits, choose NONE.
- Consider geopolitical and temporal cues in context.
- Output strictly the JSON object with fields: choice, confidence, justification.

Appendix D.5.3. Input/Output Interface

For each article, entity linking is performed over RDF nodes, using as input: an entity surface label, a short local textual context snippet and a candidate set retrieved upstream. The input is stored as a semicolon-separated CSV with columns:

```
article_id ; node_iri ; label ; snippet ; candidates_json
```

where candidates_json is a JSON array of candidate objects with keys label, qid, and note. The linker outputs one JSON object per entity (JSONL), containing the chosen Wikidata QID (or NONE), a confidence score in $[0, 1]$, and a one-sentence justification.

Appendix D.5.4. Linking Procedure (Auto-Decisions + LLM Linking)

To reduce model calls and enforce deterministic behavior in trivial cases, we applied two automatic decisions before querying the LLM:

1. AUTO_NONE (zero candidates): if no candidates are available, the system outputs choice=NONE with a default confidence, without calling the LLM.
2. AUTO_SINGLETON (one candidate): if exactly one candidate with a valid QID is available, the system selects that QID with confidence 1.0, without calling the LLM.

If two or more candidates were available, we queried the LLM (Qwen in our implementation) using the prompts in Appendixes D.5.1 and D.5.2. The candidate list was rendered as one item per line:

```
- candidate_label - QID - short note
```

Robust parsing. Model outputs were expected to be a single JSON object. If parsing failed, the system recorded choice=NONE and retained the raw model output for debugging. (Optionally, an additional JSON-repair stage can be inserted here, analogous to Appendix D.4).

Resumable Execution. Entity linking was executed in a resumable manner: previously processed entity keys (article_id, node_iri) were skipped when resume=true. This enabled incremental linking over large collections and supported recovery from interrupted runs without duplicating outputs.

Appendix E. LLM Configuration and Inference Parameters

All LLM calls were made through the OpenAI SDK ($v \geq 1.51.0$) using the OpenRouter-compatible endpoint (`openrouter.ai/api/v1`). For transparency and reproducibility, we report the exact model identifiers used for each pipeline component, decoding parameters (temperature, top- p and maximum generated tokens) as well as robustness settings (time-outs, retry budgets and rate-limiting/backoff) to ensure stable execution under provider-imposed rate limiting (HTTP 429) and short-term server overload (5xx).

Appendix E.1. Narrative Extractor (Grounded Extraction)

The narrative extractor uses Gemma as the primary extraction model, with Mistral as a secondary model used for a second opinion and JSON validation and repair. Extraction ran at low temperature to favor deterministic JSON outputs.

Table A2. Narrative extractor: models and inference parameters. Source: authors' contribution.

Item	Setting
API client	OpenAI SDK via OpenRouter base URL
Primary model	google/gemma-3-27b-it (gemma27)
Secondary model	mistralai/mistral-small-3.2-24b-instruct (mistral24)
Temperature	0.1 (extraction)
Max generated tokens	2048
Top- p	(not set; provider default)
Timeout	120 s per request
Retries	3 attempts
Backoff	exponential: sleep $(2.0)^{\text{attempt}}$ seconds
Streaming	disabled (<code>stream=False</code>)

Appendix E.2. Critic (Verification + Repair)

The critic ensemble used Mistral as the main critic and repair model and Gemma—as a verifier (secondary model). The critic was run deterministically (temperature 0.0). To mitigate throttling on shared/free routes, calls were paced with an RPM limiter and used header-aware reset handling when available.

Table A3. Critic: models and inference parameters. Source: authors' contribution.

Item	Setting
API client	OpenAI SDK via OpenRouter base URL
Models available	mistralai/mistral-small-3.2-24b-instruct (mistral24); google/gemma-3-27b-it (gemma27)
Temperature	0.0 (critic), 0.0 (repair)
Max generated tokens	2000
Top- p	1.0
Retries	up to 6 attempts
Base backoff	6.0 s (exponential; capped; jittered)
Rate limiting	target 12 RPM with jitter (uniform ± 0.25)
429 handling	sleep until X-RateLimit-Reset if present; else exponential backoff
5xx handling	exponential backoff for overload/502/503/500

Appendix E.3. Entity Linking

Entity linking used Qwen on the free route (with deterministic decoding). Requests were short (low max tokens) and paced with the same limiter pattern to reduce 429 errors.

Table A4. Entity linking: models and inference parameters. Source: authors' contribution.

Item	Setting
API client	OpenAI SDK via OpenRouter base URL
Model	qwen/qwen-2.5-72b-instruct:free (qwen72)
Temperature	0.0 (extraction), 0.0 (repair; if used)
Max generated tokens	300
Top- <i>p</i>	1.0
Retries	up to 6 attempts
Base backoff	6.0 s (exponential; capped; jittered)
Rate limiting	target 12 RPM with jitter (uniform ± 0.25)
429 handling	header-aware X-RateLimit-Reset or exponential backoff
5xx handling	exponential backoff for overload/502/503/500

Appendix F. Rules for Combining Outputs of Extraction Ensemble for GDELT

In order to match extracted narrative elements to GDELT, we first combined outputs of extraction ensemble using these type-specific ensemble rules:

- Characters/actors were taken from the union of both models and canonicalized via Wikidata entity linking when available. Otherwise, we de-duplicated them by normalized surface forms. We retained all person entities and restricted non-person actors by an evidence-based score, keeping only the top-K characters/actors per article.
- Events were also taken from the union of both models and de-duplicated using simple string/IRI overlap criteria. We found event-level entity linking to be brittle and therefore did not use it for merging.
- For causal links, we treated the primary model (Gemma) as a high-precision baseline and allowed the secondary model (Mistral) to contribute additional links only if they passed a lightweight gating heuristic based on lexical causal cues (e.g., *causes*, *leads to*, *results in*), evidence span length and optional Wikidata links on the endpoints. We encoded this as a scalar score combining these factors and retained only links above a fixed threshold.
- Framing edges followed the same scheme as causal links, but we relaxed the lexical cue requirement and instead required a non-empty evidence span and a predicted frame_type (e.g., *threat*, *support*, *opposition*).

Appendix G. Details of Evaluation Against Expert-Judged Sample

Appendix G.1. Sampling Procedure

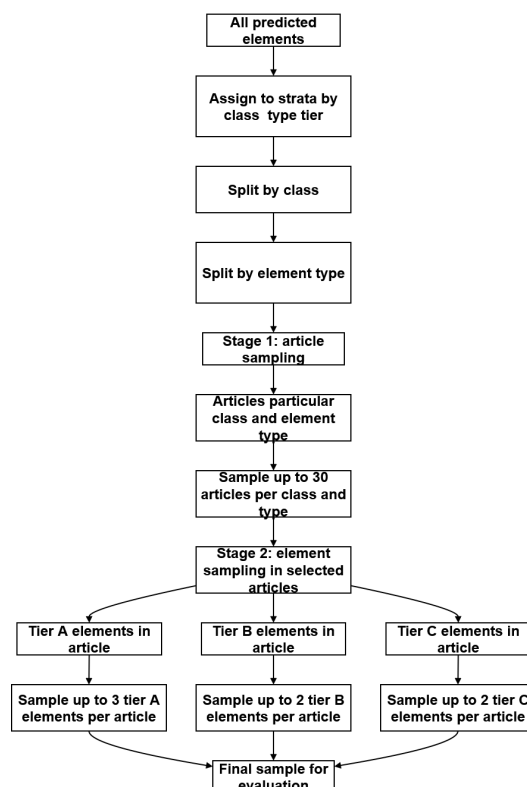


Figure A3. Sampling scheme. Source: authors' contribution.

Appendix G.2. Error Analysis Tables

Appendix G.2.1. Characters

This subsection reports all observed character-level extraction errors, presented in Table A5.

Table A5. Characters: partially correct character extractions (score 0.5) in the expert-judged sample. Source: authors' contribution.

Article ID	Class	Character	Error Type	Expert Comment (Summary)
dis_5-9	disinformation	Germany	Over-generalised actor	The article refers to specific German entities, but the model collapsed them into a single, broad actor <i>Germany</i> , which was text-grounded but too coarse.
dis_2-15	disinformation	the state	Vague/underspecified actor	<i>The state</i> was clearly discussed in the article, yet it was treated as an abstract collective rather than a clearly identifiable narrative actor, and was therefore judged only partially correct.

Appendix G.2.2. Events

This subsection presents all observed event-level extraction errors. Table A6 summarizes expert judgments for event extraction, Table A7 reports distribution of grounded elements in relation to judgment status and Table A8 provides error types with frequencies.

Table A6. Overall distribution of expert judgments for event extraction. Source: authors' contribution.

Score	Interpretation	Count	Proportion (%)
1.0	Correct	167	94.89
0.5	Partially correct	7	3.98
0.0	Incorrect	2	1.14
Total		176	100.00

Table A7. Judgment scores by grounding status (grounded). Source: authors' contribution.

Grounded?	0.0 (Incorrect)	0.5 (Partial)	1.0 (Correct)	Total
False	1	2	57	60
True	1	5	110	116

Table A8. Qualitative error types derived from expert explanations for non-perfect events (score_j1 < 1.0, $n = 9$). Source: authors' contribution.

ID	Name	Description	Freq.
A	Underspecified/too generic	The event label is semantically related to the passage but lacks essential specificity (e.g., missing key participants or the precise nature of the event). The label collapses multiple or more detailed events into a vague formulation.	2
B	Overinterpretation/exaggeration	The label asserts more than is explicitly supported by the article, often upgrading a quoted claim or local statement into a stronger factual event (e.g., inferring "state intervention is reinforced" from a statement that there is "no alternative to state aid").	3
C	Semantic drift/misframing	The label is connected to the content but reframes the event or changes its focus (e.g., interpreting a quoted statement as an event of "accepting" someone into separatist forces, or focusing on a formal "denouncement" when the text only reports statements).	2
D	Missing context/wrong event boundaries	The label refers to the correct topic or discourse element (e.g., a tweet or suggestion) but omits crucial qualifiers or temporal/contextual details, resulting in a partially correct but incomplete representation of the event.	2

Appendix G.2.3. Causal Links

Table A9 summarizes the proportion of correct items as well as proportion of correct types of causal links per class and per model. Table A10 presents main error types with descriptions, frequencies and model tendencies.

Table A9. Causal links: mean score by class, support type, and relation. Source: authors' contribution.

Class	Group	Mean Score
Disinformation	Consensus	0.96
	Gemma-only	0.75
	Mistral-only	0.23
	causes	0.69
	enables	0.62
	leads_to	0.55
Trustworthy	Consensus	0.81
	Gemma-only	0.70
	Mistral-only	0.36
	causes	0.86
	enables	0.45
	leads_to	0.68

Table A10. Error types by model support type for links with score < 1.0 (counts). Source: authors' contribution.

Error Type	Short Description	Gemma	Mistral	Consensus	Model Pattern
Missing/underspecified event	Event not provided/not defined/impossible to assess	1	18	0	Almost entirely Mistral
Events supported, causality not	Events grounded, but causal link not established or unclear	15	10	2	Mostly Gemma and Mistral; consensus rarely
Weak or implicit causality	Causality weak, implied, ambiguous, or goal-like	7	1	3	Mainly Gemma; some subtle consensus cases
Explicitly incorrect causal link	Annotator explicitly judged link as incorrect	2	3	0	Hard errors mostly Mistral
No clear causal relationship (strict)	Text explicitly judged to lack a causal relationship	3	1	0	Small, slightly skewed toward Gemma
Other/mixed	Temporal, multi-step, rhetorical, or combined issues	3	10	1	Mostly Mistral, noisy speculative edges

Appendix G.2.4. Framing Edges

Table A11 provides a summary of expert evaluation scores and frequencies by error type. Table A12 presents error details in terms of grounding, class, model source and sample tier.

Table A11. Framing-edge evaluation scores and error types. Source: authors' contribution.

Category	Count	Share
Correct (1.0)	90	81.8%
Partially correct (0.5)	6	5.5%
Incorrect (0.0)	14	12.7%
Wrong framing type	7	–
Unsupported/absent	2	–
Over-interpretation	1	–
Other/mixed	10	–

Table A12. Framing-edge errors across models, grounding, article class, and confidence tier. Source: authors' contribution.

Dimension	Group	Errors	Total
Model	Gemma	10	–
	Mistral	10	–
Grounding	Grounded	18	88
	Ungrounded	2	22
Class	Disinformation	10	52
	Trustworthy	10	58
Tier	A (consensus)	0	0
	B (single-model, grounded)	7	–
	C (ungrounded or disagreement)	13	–

Appendix G.2.5. Critic

In Table A13 we summarized overall expert judgments, while in Table A14 we displayed critic issues by category and class.

Table A13. Expert-judged correctness across critic categories. Source: authors' contribution.

Critic Category	<i>n</i>	Correct (1)	Partial (0.5)	Incorrect (0)	Correct (%)	Incorrect (%)
Framing inconsistencies	116	77	5	34	66.4	29.3
Missing links	126	96	2	28	76.2	22.2
Contradictions	111	83	4	24	74.8	21.6

Table A14. Expert-judged correctness by critic category and class. Source: authors' contribution.

Critic Category	Class	<i>n</i>	Correct (1)	Partial (0.5)	Incorrect (0)	Correct (%)
Framing inconsistencies	Disinformation	58	36	3	19	62.1
Framing inconsistencies	Trustworthy	58	41	2	15	70.7
Missing links	Disinformation	66	47	2	17	71.2
Missing links	Trustworthy	60	49	0	11	81.7
Contradictions	Disinformation	54	38	1	15	70.4
Contradictions	Trustworthy	57	45	3	9	78.9

References

1. Sinelnik, A.; Hovy, D. Narratives at Conflict: Computational Analysis of News Framing in Multilingual Disinformation Campaigns. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), Bangkok, Thailand, 11–16 August 2024; pp. 225–237. <https://doi.org/10.18653/v1/2024.acl-srw.21>.
2. Terp, S.J.; Breuer, P.C. DISARM: A Framework for Analysis of Disinformation Campaigns. In Proceedings of the IEEE CogSIMA, Salerno, Italy, 6–10 June 2022; pp. 1–8. <https://doi.org/10.1109/cogsima54611.2022.9830669>.
3. Vykopal, I.; Pikuliak, M.; Srba, I.; Moro, R.; Macko, D.; Bielikova, M. Disinformation capabilities of large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; pp. 14830–14847. <https://doi.org/10.18653/v1/2024.acl-long.793>.
4. Santos, F.C.C. Artificial intelligence in automated detection of disinformation: A thematic analysis. *Journal. Media* **2023**, *4*, 679–687. <https://doi.org/10.3390/journalmedia4020043>.
5. Zhao, J.; Tu, J.; Du, H.; Xue, N. Media attitude detection via framing analysis with events and their relations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 17197–17210. <https://doi.org/10.18653/v1/2024.emnlp-main.954>.
6. Knorr, C.; Pentzold, C. Causal Attributions (Framing). *DOCA-Database Var. Content Anal.* **2024**, *1*, 1–5. <https://doi.org/10.34778/2zaa>.
7. Frermann, L.; Li, J.; Khanehzar, S.; Mikolajczak, G. Conflicts, Villains, Resolutions: Towards models of Narrative Media Framing. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 8712–8732. <https://doi.org/10.18653/v1/2023.acl-long.486>.
8. Wang, G.; Frederick, R.; Duan, J.; Wong, W.B.; Rupar, V.; Li, W.; Bai, Q. Detecting misinformation through framing theory: The frame element-based model. *J. Comput. Soc. Sci.* **2025**, *8*, 72. <https://doi.org/10.1007/s42001-025-00403-w>.
9. Artene, C.G.; Opriș, C.; Buțincu, C.N.; Leon, F. Finding patient zero and tracking narrative changes in the context of online disinformation using semantic similarity analysis. *Mathematics* **2023**, *11*, 2053. <https://doi.org/10.3390/math11092053>.
10. Schütz, M. Disinformation detection: Knowledge infusion with transfer learning and visualizations. In Proceedings of the 45th European Conference on Information Retrieval (ECIR), Dublin, Ireland, 2–6 April 2023; pp. 468–475. https://doi.org/10.1007/978-3-031-28241-6_54.
11. González-Silot, S.; Montoro-Montarroso, A.; Cámara, E.M.; Gómez-Romero, J. Enhancing Disinformation Detection with Explainable AI and Named Entity Replacement. *arXiv* **2025**, arXiv:2502.04863. <https://doi.org/10.48550/arXiv.2502.04863>.
12. Qi, Y.; Peng, H.; Wang, X.; Xu, B.; Hou, L.; Li, J. ADELIE: Aligning Large Language Models on Information Extraction. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 7371–7387. <https://doi.org/10.18653/v1/2024.emnlp-main.419>.

13. Saxena, Y.; Chopra, S.; Tripathi, A.M. Evaluating consistency and reasoning capabilities of large language models. In Proceedings of the 2024 Second International Conference on Data Science and Information System (ICDSIS2024), Hassan, India, 17–18 May 2024; IEEE: New York, NY, USA, 2024; pp. 1–5. <https://doi.org/10.1109/ICDSIS61070.2024.10594233>.
14. Rajan, S.S.; Soremekun, E.; Chattopadhyay, S. Knowledge-based consistency testing of large language models. *arXiv* **2024**, arXiv:2407.12830. <https://doi.org/10.48550/arXiv.2407.12830>.
15. Padia, A.; Ferraro, F.; Finin, T. Enhancing Knowledge Graph Consistency through Open Large Language Models: A Case Study. In Proceedings of the AAAI Symposium Series, Stanford, CA, USA, 25–27 March 2024; Volume 3, pp. 203–208. <https://doi.org/10.1609/aaais.v3i1.31201>.
16. Ma, S.; Xu, C.; Jiang, X.; Li, M.; Qu, H.; Guo, J. Think-on-graph 2.0: Deep and interpretable large language model reasoning with knowledge graph-guided retrieval. *arXiv* **2024**, arXiv:2407.10805. <https://doi.org/10.48550/arXiv.2407.10805>.
17. Yu, P.; Deng, C.; Dai, B.; Wang, X.; Wen, Y. KaLM: Knowledge-aligned Autoregressive Language Modeling via Dual-view Knowledge Graph Contrastive Learning. *arXiv* **2024**, arXiv:2412.04948. <https://doi.org/10.48550/arXiv.2412.04948>.
18. Sun, L.; Wang, X.; Li, Y. Pyramid-Driven Alignment: Pyramid Principle Guided Integration of Large Language Models and Knowledge Graphs. *arXiv* **2024**, arXiv:2410.12298. <https://doi.org/10.48550/arXiv.2410.12298>.
19. Zheng, S.; Bai, H.; Zhang, Y.; Su, Y.; Niu, X.; Jaitly, N. Kglens: Towards efficient and effective knowledge probing of large language models with knowledge graphs. *arXiv* **2023**, arXiv:2312.11539. <https://doi.org/10.48550/arXiv.2312.11539>.
20. Chen, S.; Zhang, Q.; Dong, J.; Hua, W.; Li, Q.; Huang, X. Entity alignment with noisy annotations from large language models. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 15097–15120. <https://doi.org/10.52202/079017-0482>.
21. Mohammed, A.; Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 757–774.
22. Lu, J.; Pang, Z.; Xiao, M.; Zhu, Y.; Xia, R.; Zhang, J. Merge, ensemble, and cooperate! A survey on collaborative strategies in the era of large language models. *arXiv* **2024**, arXiv:2407.06089. <https://doi.org/10.48550/arXiv.2407.06089>.
23. Nasser, H.; Trad, F.; Chehab, A. Stacking Large Language Models is All You Need: A Case Study on Phishing URL Detection. *J. Artif. Intell. Soft Comput. Res.* **2025**, *15*, 337–356. <https://doi.org/10.2478/jaiscr-2025-0017>.
24. Jiang, D.; Ren, X.; Lin, B.Y. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 9–14 July 2023, Toronto, ON, Canada, 2023; pp. 14165–14178. <https://doi.org/10.18653/v1/2023.acl-long.792>.
25. Lu, K.; Yuan, H.; Lin, R.; Lin, J.; Yuan, Z.; Zhou, C.; Zhou, J. Routing to the expert: Efficient reward-guided ensemble of large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 16–21 June 2024, Mexico City, Mexico, 2024; pp. 1964–1974. <https://doi.org/10.18653/v1/2024.naacl-long.109>.
26. Chen, L.; Davis, J.; Hanin, B.; Bailis, P.; Stoica, I.; Zaharia, M.; Zou, J. Are More LLM Calls All You Need? Towards the Scaling Properties of Compound AI Systems. In Proceedings of the Advances in Neural Information Processing Systems 37, Vancouver, BC, Canada, 10–15 December 2024; Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C., Eds.; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2024; Volume 37, pp. 45767–45790. <https://doi.org/10.52202/079017-1455>.
27. Li, X.; Huang, Z.; Quan, S.; Peng, C.; Ma, X. SLM-MATRIX: A multi-agent trajectory reasoning and verification framework for enhancing language models in materials data extraction. *npj Comput. Mater.* **2025**, *11*, 241. <https://doi.org/10.1038/s41524-025-01719-x>.
28. Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; Griffitt, K.; Hermjakob, U.; Knight, K.; Koehn, P.; Palmer, M.; Schneider, N. Abstract meaning representation for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Sofia, Bulgaria, 8–9 August 2013; pp. 178–186. <https://aclanthology.org/W13-2322/>.
29. Gangemi, A.; Presutti, V.; Reforgiato Recupero, D.; Nuzzolese, A.G.; Draicchio, F.; Mongiovì, M. Semantic web machine reading with FRED. *Semant. Web* **2017**, *8*, 873–893. <https://doi.org/10.3233/SW-160240>.
30. Xu, D.; Chen, W.; Peng, W.; Zhang, C.; Xu, T.; Zhao, X.; Wu, X.; Zheng, Y.; Wang, Y.; Chen, E. Large language models for generative information extraction: A survey. *Front. Comput. Sci.* **2024**, *18*, 186357. <https://doi.org/10.1007/s11704-024-40555-y>.
31. Zhao, X.; Deng, Y.; Yang, M.; Wang, L.; Zhang, R.; Cheng, H.; Lam, W.; Shen, Y.; Xu, R. A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers. *ACM Comput. Surv.* **2024**, *56*, 1–39. <https://doi.org/10.1145/3674501>.
32. Kuculo, T.; Abdollahi, S.; Gottschalk, S. Transformer-Based Architectures Versus Large Language Models in Semantic Event Extraction: Evaluating Strengths and Limitations. *Semant. Web* **2025**, *16*, 22104968251363759. <https://doi.org/10.1177/22104968251363759>.
33. Susanti, Y.; Färber, M. Knowledge graph structure as prompt: Improving small language models capabilities for knowledge-based causal discovery. In *The Semantic Web–ISWC 2024, Baltimore, MD, USA, 11–15 November 2024*; Demartini, G., Hose, K., Acosta, M., Palmonari, M., Cheng, G., Skaf-Molli, H., Ferranti, N., Hernández, D., Hogan, A., Eds.; Springer: Cham, Switzerland, 2024; pp. 87–106. https://doi.org/10.1007/978-3-031-77844-5_5.

34. Leng, X.; Liang, J.; Mauro, J.; Wang, X.; Bertozzi, A.L.; Chapman, J.; Lin, J.; Chen, B.; Ye, C.; Daniel, T.; et al. Narrative Analysis of True Crime Podcasts With Knowledge Graph-Augmented Large Language Models. *arXiv* **2024**, arXiv:2411.02435. <https://doi.org/10.48550/arXiv.2411.02435>.
35. Khatiwada, A.; Shirai, S.; Srinivas, K.; Hassanzadeh, O. Knowledge Graph Embeddings for Causal Relation Prediction. In Proceedings of the DL4KG@ ISWC, Online, 24 October 2022. Available online: <https://ceur-ws.org/Vol-3342/paper-8.pdf> (accessed on 13 November 2025).
36. Lavrinovics, E.; Biswas, R.; Bjerva, J.; Hose, K. Knowledge graphs, large language models, and hallucinations: An nlp perspective. *J. Web Semant.* **2025**, *85*, 100844. <https://doi.org/10.1016/j.websem.2024.100844>.
37. Ibrahim, M.; Tinni, A.; Cheng, Z.; Fowler, A. A survey on augmenting knowledge graphs with large language models. *Discov. Artif. Intell.* **2024**, *4*, 76. <https://doi.org/10.1007/s44163-024-00175-8>.
38. Lakzaei, B.; Abdi, M.; Rezvani, R. Disinformation detection using graph neural networks: A review. *Artif. Intell. Rev.* **2024**, *57*, 102. <https://doi.org/10.1007/s10462-024-10702-9>.
39. Dar, R.A.; Hashmy, R.; Anwar, M.S.; B"ohm, P.; Frnda, J. Semantic knowledge graph fusion for fake news detection: Unifying content-based features and evidence-based analysis in the COVID-19 infodemic. *PLoS ONE* **2025**, *20*, e0321919. <https://doi.org/10.1371/journal.pone.0321919>.
40. Chen, C.; Shu, K. Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Mag.* **2024**, *45*, 354–368. <https://doi.org/10.1002/aaai.12188>.
41. Aryal, S.K.; Dhungana, P. HowardUniversityAI4PC at SemEval-2025 Task 10: Ensembling LLMs for Multi-lingual Multi-Label and Multi-Class Meta-Classification. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval 2025), Vienna, Austria, 31 July–1 August 2025; pp. 1585–1592. <https://aclanthology.org/2025.semeval-1.209.pdf>.
42. Liu, J.; Liu, Y.; Wang, T.; Xu, X.; Wang, Y.; Jiang, Y. Team QUST at SemEval-2025 Task 10: Evaluating Large Language Models in Multiclass Multi-label Classification of News Entity Framing. *arXiv* **2025**, arXiv:2506.21564. <https://doi.org/10.48550/arXiv.2506.21564>.
43. Parfenova, A. Unveiling Hidden Stories: Automated Narrative Extraction from Holocaust Diaries with Ensemble LLMs. In Proceedings of the Text2Story'25 Workshop, Lucca, Italy, 10 April 2025; CEUR Workshop Proceedings; Volume 3964, pp. 81–92. Available online: <https://ceur-ws.org/Vol-3964/paper7.pdf> (accessed on 9 December 2025).
44. Maisonnave, M.; Delbianco, F.; Tohme, F.; Milios, E.; Maguitman, A.G. Causal graph extraction from news: A comparative study of time-series causality learning techniques. *PeerJ Comput. Sci.* **2022**, *8*, e1066. <https://doi.org/10.7717/peerj-cs.1066>.
45. Tan, F.A.; Paul, D.; Yamaura, S.; Koji, M.; Ng, S.K. Constructing and interpreting causal knowledge graphs from news. *Proc. AAAI Symp. Ser.* **2023**, *1*, 52–59. <https://doi.org/10.1609/aaais.v1i1.27476>.
46. Yan, Z.; Tang, X. Narrative graph: Telling evolving stories based on event-centric temporal knowledge graph. *J. Syst. Sci. Syst. Eng.* **2023**, *32*, 206–221. <https://doi.org/10.1007/s11518-023-5561-0>.
47. Li, Z.; Pan, R.; Pi, X. Beyond LLMs A Linguistic Approach to Causal Graph Generation from Narrative Texts. In Proceedings of the 7th Workshop on Narrative Understanding, Albuquerque, NM, USA, 4 May 2025; Curran Associates, Inc.: Red Hook, NY, USA, 2025, pp. 36–51. <https://doi.org/10.18653/v1/2025.wnu-1.10>.
48. Heinisch, P.; Plenz, M.; Frank, A.; Cimiano, P. ACCEPT at SemEval-2023 Task 3: An ensemble-based approach to multilingual framing detection. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Toronto, ON, Canada, 13–14 July 2023; pp. 1347–1358. <https://doi.org/10.18653/v1/2023.semeval-1.187>.
49. Jiang, Y.; Lai, S.; Guo, L.; Ishwar, P.; Wijaya, D.; Betke, M. Exploring an alternative computational approach for news framing analysis through community detection in framing element networks. *Journal. Mass Commun. Q.* **2025**. <https://doi.org/10.1177/10776990251328597>.
50. Ilyas, M.A.; Rehman, A.; Abbas, A.; Kim, D.; Naseem, M.T.; Allah, N.M. Fake News Detection on Social Media Using Ensemble Methods. *Comput. Mater. Contin.* **2024**, *81*, 4525–4549. <https://doi.org/10.32604/cmc.2024.056291>.
51. Munir, S.; Asif Naeem, M. BiL-FaND: Leveraging ensemble technique for efficient bilingual fake news detection. *Int. J. Mach. Learn. Cybern.* **2024**, *15*, 3927–3949. <https://doi.org/10.1007/s13042-024-02128-0>.
52. Manos, A.; Filippidou, D.E.; Pavlidis, N.; Karanasios, G.; Vachtanidis, G.; D'Ulizia, A.; D'Andrea, A. Knowledge graphs and machine learning in fake news and disinformation detection. In Proceedings of the 2024 International Conference on Engineering and Emerging Technologies (ICEET), Dubai, United Arab Emirates, 27–28 December 2024; IEEE: New York, NY, USA, 2024; pp. 1–6. <https://doi.org/10.1109/ICEET65156.2024.10913780>.
53. Nguyen, V.D.K.; Do, P. Fake news detection using knowledge graph and graph convolutional network. *J. Intell. Fuzzy Syst.* **2023**, *45*, 11107–11119. <https://doi.org/10.3233/JIFS-233260>.
54. Zhang, B.; Soh, H. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), Miami, Florida, USA, 12–16 November 2024; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 9820–9836. <https://doi.org/10.18653/v1/2024.emnlp-main.548>.

55. Zhong, Z.; Zhou, K.; Mottin, D. Harnessing Large Language Models as Post-hoc Correctors. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 14559–14574. <https://doi.org/10.18653/v1/2024.findings-acl.867>.
56. Dagdelen, J.; Dunn, A.; Lee, S.; Walker, N.; Rosen, A.S.; Ceder, G.; Persson, K.A.; Jain, A. Structured information extraction from scientific text with large language models. *Nat. Commun.* **2024**, *15*, 1418. <https://doi.org/10.1038/s41467-024-45563-x>.
57. Brinkmann, A.; Baumann, N.; Bizer, C. Using LLMs for the Extraction and Normalization of Product Attribute Values. In *Advances in Databases and Information Systems (ADBIS 2024)*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2024; pp. 217–230. https://doi.org/10.1007/978-3-031-70626-4_15.
58. Leite, J.A.; Razuvaevskaya, O.; Bontcheva, K.; Scarton, C. 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, Boise, ID, USA, 21–25 October 2024*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 5380–5384. <https://doi.org/10.1145/3627673.3679167>.
59. Diffbot Technologies Corp. Diffbot: A Tool for Automated Web Data Extraction. Available online: <https://www.diffbot.com/> (accessed on 15 October 2025).
60. Mandravickaitė, J. Narrative Structure Extraction in Disinformation and Trustworthy News: A Comparison of LLM, KG, and KG-Augmented Pipelines. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops, Hannover, Germany, 9–11 September 2025*; pp. 86–103. Available online: <https://aclanthology.org/2025.konvens-2.7/> (accessed on 16 October 2025).
61. Leite, J.; Razuvaevskaya, O.; Bontcheva, K.; Scarton, C. EUvsDisinfo: A Dataset for Multilingual Detection of Pro-Kremlin Disinformation in News Articles (Dataset), 2024. Available online: <https://zenodo.org/records/10514307> accessed on 15 January 2026).
62. Internet Archive. Available online: <https://archive.org/> (accessed on 5 February 2026).
63. Google. *Google Colaboratory*; Google: Mountain View, CA, USA. Available online: <https://colab.research.google.com> (accessed on 8 February 2026).
64. OpenRouter. Available online: <https://openrouter.ai/> (accessed on 8 February 2026).
65. Hofmann-Coyle, E.; Kulkarni, M.; Xie, L.; Maddela, M.; Preotiu-Pietro, D. Extractive entity-centric summarization as sentence selection using bi-encoders. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Online, 21–23 November 2022; pp. 326–333. <https://doi.org/10.18653/v1/2022.aacl-short.40>.
66. Erkan, G.; Radev, D.R. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* **2004**, *22*, 457–479. <https://doi.org/10.1613/jair.1523>.
67. Miller, D. Leveraging BERT for extractive text summarization on lectures. *arXiv* **2019**, arXiv:1906.04165. <https://doi.org/10.48550/arXiv.1906.04165>.
68. Lin, C.Y. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches out, Barcelona, Spain, 25–26 July 2004*; pp. 74–81. Available online: <https://aclanthology.org/W04-1013/> (accessed on 15 February 2026).
69. Palmer, M.; Gildea, D.; Kingsbury, P. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.* **2005**, *31*, 71–106. <https://doi.org/10.1162/0891201053630264>.
70. Ruppenhofer, J.; Ellsworth, M.; Petruck, M.R.; Johnson, C.R.; Scheffczyk, J. *FrameNet II: Extended Theory and Practice*; ICSI: Berkeley, CA, USA, 2016.
71. Oral, E.; Acar, A.; Eryigit, G. Abstract meaning representation of Turkish. *Nat. Lang. Eng.* **2024**, *30*, 171–200. <https://doi.org/10.1017/S1351324922000183>.
72. Pustejovsky, J.; Lai, K.; Xue, N. Modeling quantification and scope in Abstract Meaning Representations. In *Proceedings of the First International Workshop on Designing Meaning Representations, Florence, Italy, 1 August 2019*; pp. 28–33. <https://doi.org/10.18653/v1/W19-3303>.
73. Gangemi, A.; Asprino, L.; De Giorgis, M.; Troianiello, G.; Alam, M.; Nuzzolese, A.G. Text2AMR2FRED, a Tool for Transforming Text into RDF/OWL Knowledge Graphs via Abstract Meaning Representation. In *Proceedings of the ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, Athens, Greece, 6–10 November 2023*; CEUR Workshop Proceedings; Volume 3485, p. 5. <https://hdl.handle.net/11584/426131>.
74. Lacerra, C.; Pasini, T.; Tripodi, R.; Navigli, R. Alasca: An automated approach for large-scale lexical substitution. In *Proceedings of the IJCAI 2021, Montreal, QC, Canada, 19–27 August 2021*; pp. 3836–3842. <https://doi.org/10.24963/ijcai.2021/528>.
75. Lacerra, C.; Tripodi, R.; Navigli, R. Genesis: A generative approach to substitutes in context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021*. <https://doi.org/10.18653/v1/2021.emnlp-main.844>.

76. Kamp, H. A theory of truth and semantic representation. In *Meaning and the Dynamics of Interpretation*; Brill: Kowloon, Hong Kong, 2013; pp. 329–369. https://doi.org/10.1163/9789004252882_014.
77. Gangemi, A.; Alam, M.; Asprino, L.; Presutti, V.; Recupero, D.R. Framester: A wide coverage linguistic linked data hub. In *Knowledge Engineering and Knowledge Management*; Springer: Cham, Switzerland, 2016; pp. 239–254. https://doi.org/10.1007/978-3-319-49004-5_16.
78. Fellbaum, C. *WordNet: An Electronic Lexical Database*; MIT Press: Cambridge, MA, USA, 1998.
79. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; Van Kleef, P.; Auer, S.; et al. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semant. Web* **2015**, *6*, 167–195. <https://doi.org/10.3233/SW-140134>.
80. Paulheim, H.; Gangemi, A. Serving DBpedia with DOLCE—more than just adding a cherry on top. In *The Semantic Web-ISWC 2015, Bethlehem, PA, USA, 11–15 October 2015*; Springer: Cham, Switzerland, 2015; pp. 180–196. https://doi.org/10.1007/978-3-319-25007-6_11.
81. py_amr2fred: A Python Library for Translating AMR Graphs to FRED RDF. Available online: https://github.com/infovillasimius/py_amr2fred (accessed on 1 December 2025).
82. Mistral AI Team. Mistral Small 3. Available on: <https://mistral.ai/news/mistral-small-3> (accessed on 18 November 2025).
83. Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; Rouillard, L.; et al. Gemma 3 Technical Report. *arXiv* **2025**, arXiv:2503.19786. <https://doi.org/10.48550/arXiv.2503.19786>.
84. Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; Liu, Y. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. *arXiv* **2024**, arXiv:2412.05579. <https://doi.org/10.48550/arXiv.2412.05579>.
85. Huang, H.; Bu, X.; Zhou, H.; Qu, Y.; Liu, J.; Yang, M.; Xu, B.; Zhao, T. An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 27 July–1 August 2025*; pp. 5880–5895. <https://doi.org/10.18653/v1/2025.findings-acl.306>.
86. Hui, B.; Yang, J.; Cui, Z.; Yang, J.; Liu, D.; Zhang, L.; Liu, T.; Zhang, J.; Yu, B.; Lu, K.; et al. Qwen2.5 Technical Report. *arXiv* **2024**, arXiv:2412.15115. <https://doi.org/10.48550/arXiv.2412.15115>.
87. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783. <https://doi.org/10.48550/arXiv.2407.21783>.
88. Huang, Y.; Feng, X.; Li, B.; Xiang, Y.; Wang, H.; Liu, T.; Qin, B. Ensemble learning for heterogeneous large language models with deep parallel collaboration. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS'24), Vancouver, BC, Canada, 10–15 December 2024*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 119838–119860. <https://doi.org/10.52202/079017-3808>.
89. Buckingham, K.; Brandt, J.; Anderson, W.; do Amaral, L.F.; Singh, R. The untapped potential of mining news media events for understanding environmental change. *Curr. Opin. Environ. Sustain.* **2020**, *45*, 92–99. <https://doi.org/10.1016/j.cosust.2020.08.015>.
90. Lu, X.; Chan, F.K.S.; Li, N.; Chen, C.; Chen, W.Q.; Chan, H.K. Improving urban flood resilience via GDELT GKG analyses in China's Sponge Cities. *Sci. Rep.* **2022**, *12*, 20317. <https://doi.org/10.1038/s41598-022-24370-8>.
91. Piazza, A.; Vasudevan, S. Mapping the cyberthreat landscape in healthcare using GDELT: A multimethod approach. *Health Secur.* **2025**, *23*, 186–197. <https://doi.org/10.1089/hs.2024.0016>.
92. Mudassar, M.; Lin, T.; Wang, X.; Geng, H. Global trends and influential factors of climate change adaptation recognition in GDELT. *Sci. Rep.* **2025**, *15*, 26887. <https://doi.org/10.1038/s41598-025-11794-1>.
93. Alipour, S.; Di Marco, N.; Avallé, M.; Etta, G.; Cinelli, M.; Quattrociochi, W. The drivers of global news spreading patterns. *Sci. Rep.* **2024**, *14*, 1519. <https://doi.org/10.1038/s41598-024-52076-6>.
94. Bassoli, M.; Greppi, T. Different refugees, different portrayals. A GDELT-based comparative analysis of the mediatic portrayal of Syrian refugees in 2015 and Ukrainian refugees in 2022 across seven European countries. *Eur. J. Commun.* **2025**, *41*, 41–60. <https://doi.org/10.1177/02673231251394694>.
95. Xiao, H.; Li, X.; Yang, J.; Ye, X. Can AI-Driven National ESG in Big Data Help Improve Macroeconomic Forecasting? *J. Econ. Forecast.* **2025**, *2*, 84–103. Available online: <https://ideas.repec.org/a/rjr/romjef/vy2025i2p84-103.html> (accessed on 15 February 2026).
96. Hoffmann, M.; Santos, F.G.; Neumayer, C.; Mercea, D. Lifting the veil on the use of big data news repositories: A documentation and critical discussion of a protest event analysis. *Commun. Methods Meas.* **2022**, *16*, 283–302. <https://doi.org/10.1080/19312458.2022.2128099>.
97. Hong, D.; Fu, Z.; Zhang, X.; Pan, Y. Research on the Development and Application of the GDELT Event Database. *Data* **2025**, *10*, 158. <https://doi.org/10.3390/data10100158>.
98. Vrandečić, D.; Pintscher, L.; Krötzsch, M. Wikidata: The making of. In *Proceedings of the Companion Proceedings of the ACM Web Conference 2023, Austin, TX, USA, 30 April–4 May 2023*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 615–624. <https://doi.org/10.1145/3543873.3585579>.

99. MediaWiki Developers. MediaWiki API. Available online: https://www.mediawiki.org/wiki/API:Main_page (accessed on 5 February 2026).
100. Viswanathan, V.; Rajani, N.F.; Bentor, Y.; Mooney, R. Stacked ensembles of information extractors for knowledge-base population. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 27–31 July 2015; pp. 177–187. <https://doi.org/10.3115/v1/P15-1018>.
101. Martínez-Rodríguez, J.L.; Hogan, A.; López-Arévalo, I. Information Extraction Meets the Semantic Web: A Survey. *Semant. Web* **2020**, *11*, 255–335. <https://doi.org/10.3233/SW-180333>.
102. De Giorgis, S.; Gangemi, A.; Russo, A. Neurosymbolic graph enrichment for grounded world models. *Inf. Process. Manag.* **2025**, *62*, 104127. <https://doi.org/10.1016/j.ipm.2025.104127>.
103. Gangemi, A.; Graciotti, A.; Meloni, A.; Nuzzolese, A.G.; Presutti, V.; Reforgiato Recupero, D.; Russo, A. py-amr2fred: A Python Library for Converting Text into OWL-Compliant RDF KGs. In *The Semantic Web. ESWC 2025*; Springer: Cham, Switzerland, 2025; pp. 65–83. https://doi.org/10.1007/978-3-031-94578-6_4.
104. Wolfe, T. Report Linking: Information Extraction for Building Topical Knowledge Bases. Ph.D. Thesis, Johns Hopkins University, Baltimore, MD, USA, 2017.
105. Kwak, H.; An, J. A first look at global news coverage of disasters by using the gdelt dataset. In *Social Informatics*; Springer: Cham, Switzerland, 2014; pp. 300–308. https://doi.org/10.1007/978-3-319-13734-6_22.
106. Stoehr, N.; Torroba Hennigen, L.; Valvoda, J.; West, R.; Cotterell, R.; Schein, A. An Ordinal Latent Variable Model of Conflict Intensity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, ON, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 4817–4830. <https://doi.org/10.18653/v1/2023.acl-long.265>.
107. Scheufele, D.A.; Krause, N.M. Science Audiences, Misinformation, and Fake News. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 7662–7669. <https://doi.org/10.1073/pnas.1805871115>.
108. Broda, E.; Strömbäck, J. Misinformation, Disinformation, and Fake News: Lessons from an Interdisciplinary, Systematic Literature Review. *Ann. Int. Commun. Assoc.* **2024**, *48*, 139–166. <https://doi.org/10.1080/23808985.2024.2323736>.
109. Shu, K.; Wang, S.; Lee, D.; Liu, H. Mining disinformation and fake news: Concepts, methods, and recent advancements. In *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*; Springer International Publishing: Cham, Switzerland, 2020; pp. 1–19. https://doi.org/10.1007/978-3-030-42699-6_1.
110. Hellman, M. Narrative Analysis and Framing Analysis of Disinformation. In *Security, Disinformation and Harmful Narratives: RT and Sputnik News Coverage about Sweden*; Springer Nature: Cham, Switzerland, 2024; pp. 101–121. https://doi.org/10.1007/978-3-031-58747-4_4.
111. Boudana, S.; Segev, E. Fake News Makes the News: Definitions and Framing of Fake News in Mainstream Media. *Journal. Pract.* **2026**, *20*, 679–698. <https://doi.org/10.1080/17512786.2024.2379898>.
112. Tay, L.Q.; Lewandowsky, S.; Hurlstone, M.J.; Kurz, T.; Ecker, U.K. Thinking clearly about misinformation. *Commun. Psychol.* **2024**, *2*, 4. <https://doi.org/10.1038/s44271-023-00054-5>.
113. Liu, K.; Chen, Y.; Liu, S.; Zuo, X.; Zhao, J. Extracting Events and Their Relations from Texts: A Survey on Recent Research Progress and Challenges. *AI Open* **2020**, *1*, 22–39. <https://doi.org/10.1016/j.aiopen.2021.02.004>.
114. Detroja, K.; Bhensdadia, C.K.; Bhatt, B.S. A Survey on Relation Extraction. *Intell. Syst. Appl.* **2023**, *19*, 200244. <https://doi.org/10.1016/j.iswa.2023.200244>.
115. Santana, B.S.; Campos, R.; Amorim, E.; Nunes, S.; Jorge, A.M. A Survey on Narrative Extraction from Textual Data. *Artif. Intell. Rev.* **2023**, *56*, 8393–8435. <https://doi.org/10.1007/s10462-022-10338-7>.
116. Bose, P.; Srinivasan, U.; Thakur, S.; Kelly, D.; Pakhomov, S.V.; Liu, H. A Survey on Recent Named Entity Recognition and Relationship Extraction on Clinical Text. *Appl. Sci.* **2021**, *11*, 8319. <https://doi.org/10.3390/app11188319>.
117. Das, R.; Chandra, A.; Lee, I.T.; Pacheco, M.L. Media Framing through the Lens of Event-Centric Narratives. In Proceedings of the 6th Workshop on Narrative Understanding, Miami, FL, USA, 15 November 2024; pp. 85–98. <https://doi.org/10.18653/v1/2024.wnu-1.15>.
118. Norambuena, B.F.K.; Mitra, T.; North, C. A Survey on Event-Based News Narrative Extraction. *ACM Comput. Surv.* **2023**, *55*, 1–39. <https://doi.org/10.1145/3584741>.
119. Schmidt, L.; Hair, K.; Graziosi, S.; Campbell, F.; Kapp, C.; Khanteymoori, A.; Craig, D.; Engelbert, M.; Thomas, J. Exploring the Use of a Large Language Model for Data Extraction in Systematic Reviews: A Rapid Feasibility Study. *arXiv* **2024**, arXiv:2405.14445. <https://doi.org/10.48550/arXiv.2405.14445>.
120. Zheng, S.; Cao, W.; Xu, W.; Bian, J. Revisiting the Evaluation of End-to-End Event Extraction. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 2–5 August 2021; pp. 4609–4617. <https://doi.org/10.18653/v1/2021.findings-acl.405>.

121. Pournaki, A.; Willaert, T. A Graph-Based Approach to Extracting Narrative Signals from Public Discourse. *arXiv* **2024**, arXiv:2411.00702. <https://doi.org/10.48550/arXiv.2411.00702>.
122. Kirmayer, L.J. Science and Sanity: A Social Epistemology of Misinformation, Disinformation, and the Limits of Knowledge. *Transcult. Psychiatry* **2024**, *61*, 795–808. <https://doi.org/10.1177/13634615241296301>.
123. Musi, E.; Reed, C. From Fallacies to Semi-fake News: Improving the Identification of Misinformation Triggers Across Digital Media. *Discourse Soc.* **2022**, *33*, 349–370. <https://doi.org/10.1177/09579265221076609>.
124. Nascimento, I.K.D. The Psychological Cycle of Misinformation: An Integrative Model of Legitimation Perception in Social Networks. *Front. Soc. Psychol.* **2025**, *3*, 1621794. <https://doi.org/10.3389/frsps.2025.1621794>.
125. Text2AMR2FRED WebApp (Demonstration Interface). Available online: <https://arco.istc.cnr.it/txt-amr-fred/> (accessed on 18 December 2025).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.