BetweenTheLines - Cross Source News Analysis

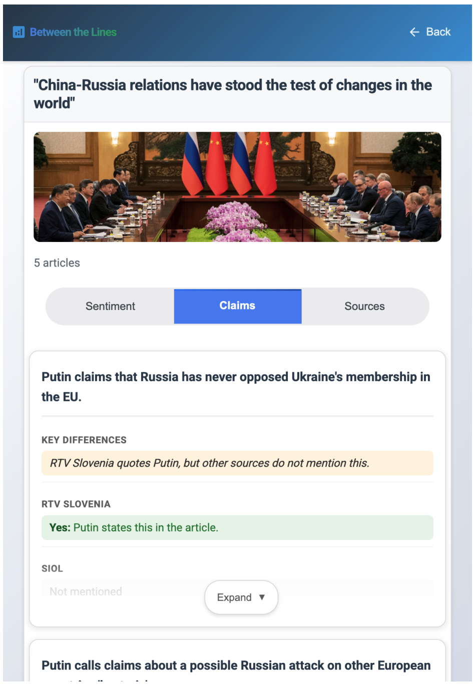
Georgi Trajkov
geotrajkov0@gmail.com
Jožef Stefan Institute
Ljubljana, Slovenia

Marko Grobelnik
marko.grobelnik@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Adrian Mladenic Grobelnik
adrian.m.grobelnik@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Different news outlets covering the same event often emphasize, omit, or frame facts differently, making cross-source comparison essential for understanding media bias and information diversity. Large language models (LLMs) can automate this analysis, but simple single-LLM prompt approaches tend to underperform when processing large amounts of data [1]. Platforms like Ground News [2] and Event Registry [3] provide publisher and article-level bias scores but cannot track how individual claims and entities are portrayed by articles. The fundamental challenge is determining whether LLM prompt architecture affects accuracy when classifying claim presence across multiple news sources. We show that a multi-prompt LLM architecture reduces classification errors 7-fold (from 33.0% to 4.67%) compared to single-prompt approaches. Our pipeline first extracts all claims and entities from articles collectively, then evaluates each article separately for claim presence (confirmed/contradicted/partial/absent) and entity sentiment. This decomposition virtually eliminates false positives, major errors dropped from 28.0% to 0.79% across 797 manually validated claim-publisher pairs from Slovene news. The results demonstrate that task decomposition, not LLM sophistication, drives accuracy in cross-source analysis. This finding enables scalable media monitoring at $0.01 per event, making systematic bias detection accessible to journalists and researchers worldwide.



"China-Russia relations have stood the test of changes in the world"

5 articles

Putin claims that Russia has never opposed Ukraine's membership in the EU.

KEY DIFFERENCES

RTV Slovenia quotes Putin, but other sources do not mention this.

RTV SLOVENIA

Yes: Putin states this in the article.

SIOL

Not mentioned

Putin calls claims about a possible Russian attack on other European

Figure 1: Analyzed event in BetweenTheLines mobile webapp, showing the claims tab

1 Introduction

Different news sources (publishers) covering the same event (groups of articles reporting on the same story) often cover facts differently. While existing platforms like Event Registry [3] and Ground News [2] provide valuable bias indicators and sentiment scores, they do not track how specific entities (People, Organizations, Countries) and claims (Factual Claims) within articles are portrayed across publishers. Getting insight into these differences is usually time-consuming for the user.

Thus we present BetweenTheLines, (Figure 1) a system that automatically identifies claims and entities in an event, and tracks their portrayal in each individual publisher. For example, when analyzing political coverage, we can see how the same entity is portrayed differently by 2 publishers, and how one publisher omitted a claim while the other did not.

Our key technical contribution is demonstrating that multi-prompt LLM architecture outperforms single-stage approaches for this task.

2 Related Work

Cross-source news analysis is an under-discussed area of research which is important for understanding media bias, information diversity, and narrative framing across different outlets. This section reviews existing approaches to cross-source news analysis, event aggregation systems, and LLM-based content analysis pipelines.

2.1 Cross-Source News Analysis Platforms

Ground News represents a prominent platform for cross-source news comparison, classifying publishers along the left-right political spectrum. The platform has gained widespread adoption in educational institutions, with libraries at Harford Community College [4] and West Virginia University [5] integrating it into their media literacy curricula. For each news event, Ground News allows users to compare coverage by publisher on aggregate. While these aggregated summaries can reveal different emphases across the political spectrum, the platform does not provide article-by-article comparisons or track how specific entities and claims are portrayed between articles.

2.2 Event-Centric News Aggregation

Event Registry [6, 3] pioneered event-centric news aggregation by clustering articles from multiple publishers around identified news events. The platform provides article-level sentiment scores using VADER sentiment analysis [7] and allows filtering

SiKDD 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.sikdd.26

by various parameters including language, location, and publisher credibility. Each article has a sentiment score, a level of granularity above Ground News. Still there is no analysis for how specific entities and claims within those articles are portrayed.

Our work builds upon Event Registry's foundation, by combining its event-based aggregation, with more granular entity and claim analyses through LLM processing. Unlike Ground News's publisher-level political bias ratings or Event Registry's article sentiment scores, we provide fine-grained analysis of how specific entities and claims are portrayed differently across publishers.

3 Application and analysis Architecture

3.1 Application architecture

"BetweenTheLines is a news-analysis web app 1, developed with Claude Code [8]." The backend is built using Flask [9] and PostgreSQL [10]. It uses Event Registry [6, 3] analysis service for event and article fetching, and integrates both Google Gemini [11] and OpenAI [12] LLMs.

3.2 Analysis Service overview

The analysis service consists of two modules, claims analysis and sentiment analysis, with more thorough exploration of the former due to it's less subjective nature. Figure 2 illustrates our three-stage LLM pipeline.

Stage 1: Extraction. We begin by sending all articles from an event to a single LLM call. This extracts two lists (Table 1) for entities and claims that appear in the articles.

Stage 2: Classification. With the lists from stage 1, a parallel LLM call is made twice for each publisher, once for claims, once for entities. The calls return categorized data. Claims are categorized by presence, and entities by sentiment. The results of these categorizations are referred to as entity-publisher and claim-publisher pairs.

Stage 3: Key Differences. Summarizes how different publishers covered each claim or entity. This requires one LLM call per claim/entity, running in parallel.

The final results are structured into a tabular or card format, depending on device, where users can compare coverage across publishers at a glance (Figure 1).

3.3 Language

We decided for all prompts to be in Slovene, and to analyze only Slovene articles. This came after empirically observing a decrease in errors when the language of the prompts and articles was the same. It also language consistency for evaluation.

All showcased prompts and results are originally Slovene, and were translated to English for the paper.

3.4 Event Filtering

Events and articles are fetched from the Event Registry API[3].

Articles are then filtered to only retain the newest article for each unique publisher in an event. To retain only the most relevant events, we discard any events with less than 3 articles.

To prevent context overloading maximum article limit is 10. Then final article list is prepared for each event, and the title, body, publisher name, and article link is stored for every article.

3.5 Extraction

Extraction for an event is done after filtering, in a single LLM call to gpt-4o-mini [13], in which the contents of all articles are included in the prompt along with instructions for extracting 2 lists (Table 1) in JSON format: entities for sentiment analysis and claims for claims analysis.

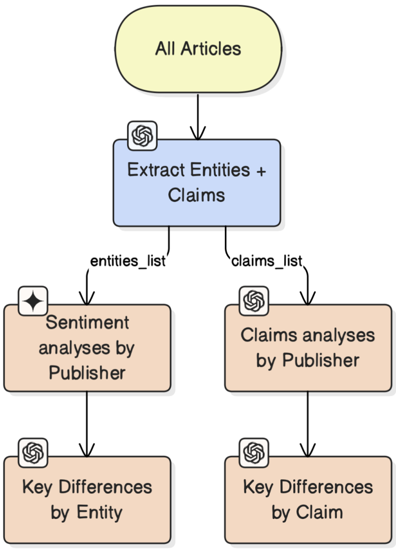

All Articles
Extract Entities + Claims
entities_list
claims_list
Sentiment analyses by Publisher
Claims analyses by Publisher
Key Differences by Entity
Key Differences by Claim

Figure 2: three Stage process flow of analysis. Extracted results lead to multiple parallel LLM calls.

The prompt focuses on extracting 8-15 claims and 8-15 entities that are central to the story, explicitly excluding news publishers unless they are the subject of the news story:

```
Analyze all these news articles and extract two comprehensive
     lists in JSON format:
  1. All significant CLAIMS made across all articles
  2. All important ENTITIES (people, organizations, countries, etc
       .) mentioned across all articles
```

A 2-step extraction process was also tested, where each article is prompted for claims and entities contained in it, and then the results are aggregated. However, this led to very large lists with duplicate names written differently (e.g., USA vs United States Government vs United States), for little performance gain.

Another issue we faced was the publisher names themselves being in the entities list, even in situations where they are not a direct part of the article. This led us to add additional rules in the extraction prompt to not include them:

```
-EXCLUDE news publishers/sources (like BBC, CNN, Reuters, etc.)
      UNLESS they are actually subjects of the news story itself
- Focus on entities that are the SUBJECT of the news, not the
      source reporting it
```

Entities	Claims
Vladimir Putin	Putin claims that Russia has never opposed Ukraine's membership in the EU.
Xi Jinping	Putin calls claims about a possible Russian attack on other European countries "hysteria."
Russia	Putin says that Russia is forced to respond to the West's attempt to take over the post-Soviet space.
China	Putin and Trump discussed the security of Ukraine.
Ukraine	Putin and Xi signed about 20 agreements in the fields of energy, aviation, artificial intelligence, and agriculture.

Table 1: Example of first 5 entities and claims received from extraction prompt for Russia–China summit.

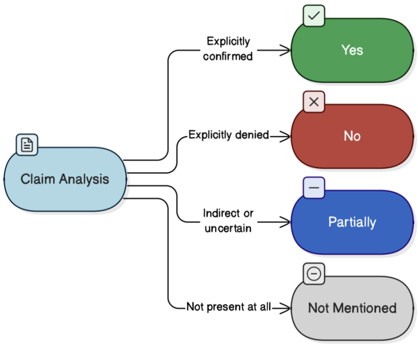


Figure 3: Claims analysis decision tree, 4 options depending on whether and how a claim is mentioned

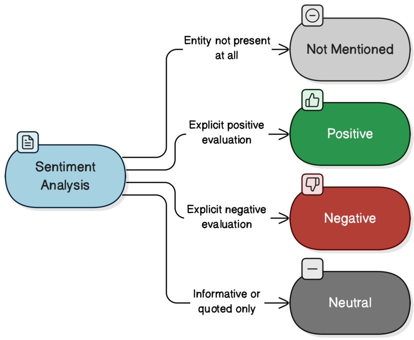

Sentiment Analysis
Entity not present at all
Not Mentioned
Explicit positive evaluation
Positive
Explicit negative evaluation
Negative
Informative or quoted only
Neutral

Figure 4: Decision tree in sentiment analysis

3.6 Claims Analysis

Claim analysis starts after the extraction step returns a claims list. It consists of multiple parallel LLM calls, each analyzing a single article against the claims list, using 4 categorizations for whether the article confirms the claim: Yes, Partially, No and Not mentioned, as depicted in Figure 3.

False negatives were the biggest problem we faced with claim analysis. Originally, there were only 3 claim categories; however, due to too many "not mentioned" results, we added a fourth partial classification that led to significant improvements. To further reduce false negatives without adding false positives, we tightened the categorization rules for the Not mentioned category, to default to Partial instead when answer is unclear.

Portion of the rule-set that helped improve results:

```
Before selecting "Not mentioned", you MUST check the following
transformations/hints:
- paraphrases/synonyms; hypernyms/hyponyms; abbreviations/acronyms;
  coreferences (pronouns, descriptive references)
- numbers/units/conversions; relative dates -> absolute;
  geographic hypernyms (e.g. EU -> country)
- sections: title, introduction, body, subtitles, captions,
     tables/graphs,
  quotes/indirect statements
- negations, questions, conditionals, predictions/hypotheses
Rule to reduce false negatives:
- If in doubt between "Partial" and "Not mentioned", choose "Partial"
```

3.7 Sentiment Analysis

The sentiment analysis proceeds in parallel with claims analysis after receiving the entity list (Figure 1) from the extraction. It is structured in a manner very similar to the claims analysis, it calls the LLM once per publisher, and it has 4 categorizations (Figure 4): Positive, Negative, Neutral, and Not Mentioned. Accuracy assessment is harder due to subjective interpretation. The module uses gemini-2.5-flash-lite [14] due to empirical observation of better results, every other LLM call uses gpt-4o-mini [13].

LLMs struggle with implicit criticism conveyed through selective quoting. For instance, when Mladina [15] quoted Trump praising himself as "smart" and suggesting people want a dictator, the LLM classified sentiment as positive, missing the article's critical intent to portray authoritarianism.

To account for this weakness, we added more constraints and rules in the prompts:

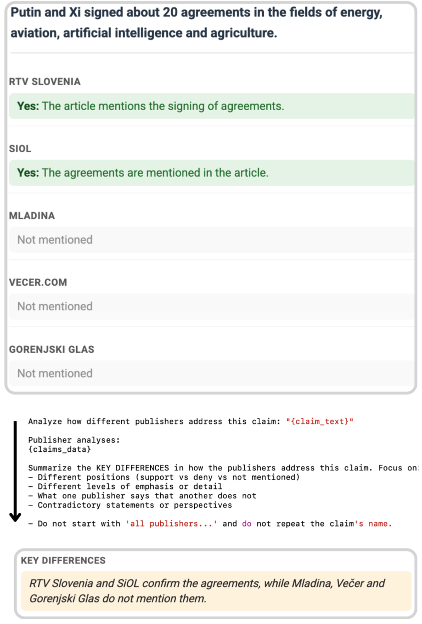
```
Important: OUTCOME ≠ SENTIMENT

- Do not mark "Positive" because the entity wins/makes a profit,
     without explicit value judgement of the entity.
- Do not mark "Negative" because the entity loses/has a bad result
     , without explicit value judgement of the entity.
Mandatory decision steps (before choosing a label):
- First identify the role of the entity's mention: SPEAKER /
       TARGET / MENTIONED WITHOUT ROLE
- Then look for META-EVALUATION of the entity (adjectives,
      evaluative verbs, framing before/after the quote, editorial
       tone).
- If the entity is only a SPEAKER without meta-evaluation, choose
       "Neutral".
```

This resulted in false negatives and positives reducing significantly, however it also came with the tradeoff of having a much higher incidence of neutral classification, even when it is slightly positive or negative.

3.8 Key Differences

The final step of the pipeline is the generation of the key differences (Figure 5). It uses the claims/sentiment categorizations from the previous step as input. It works for both Claims and Sentiment analysis in an almost identical manner; we will use claims for explanation in this example. A parallel LLM call is made once per every claim in the analysis, containing all claim-publisher pairs of the claim.



Putin and Xi signed about 20 agreements in the fields of energy, aviation, artificial intelligence and agriculture.

RTV SLOVENIA

Yes: The article mentions the signing of agreements.

SIOL

Yes: The agreements are mentioned in the article.

MLADINA

Not mentioned

VECER.COM

Not mentioned

GORENJSKI GLAS

Not mentioned

```
Analyze how different publishers address this claim: "{claim_text}"

Publisher analyses:
{claims_data}

Summarize the KEY DIFFERENCES in how the publishers address this claim. Focus on:
- Different positions (support vs deny vs not mentioned)
- Different levels of emphasis or detail
- What one publisher says that another does not
- Contradictory statements or perspectives

- Do not start with 'all publishers...' and do not repeat the claim's name.
```

KEY DIFFERENCES

RTV Slovenia and SiOL confirm the agreements, while Mladina, Večer and Gorenjski Glas do not mention them.

Figure 5: Key difference generation for claim from Russia-China Summit

	Hvar snakebite		**Putin prepared to meet Zelenski**		**Carpaccio's Mary Returns to Piran**		**Giorgio Armani dies**		**Russia–China summit**		**Weighted avg**	
	Single	**Multi**	**Single**	**Multi**	**Single**	**Multi**	**Single**	**Multi**	**Single**	**Multi**	**Single**	**Multi**
Publishers	7		7		9		7		5		—	
Claims	9	15	9	14	8	15	8	15	8	12	—	—
Error rate	25.4%	3.80%	30.15%	3.06%	38.9%	6.3%	37.5%	7.62%	32.5%	0%	**33.0%**	**4.67%**
Major errors	25.4%	1.90%	14.28%	0%	37.5%	0%	30.4%	1.91%	32.5%	0%	**28.0%**	**0.79%**
Rows affected	100%	20%	88.8%	7.14%	100%	33.3%	87.5%	33.3%	100%	0%	**95.3%**	**21.5%**

Table 2: Single-stage (left) vs. multi-stage (right) per event. Final column shows weighted averages. For error rates and major errors, weights = number of claim-publisher pairs tested per pipeline. For rows affected, weights = number of claims (rows) per pipeline. Note that weights differ between pipelines due to different extraction results.

4 Evaluation

4.1 Manual Testing

To test our hypothesis that the multi-stage pipeline is superior to a single-stage pipeline (where all articles and instructions are included in a single one prompt LLM call), we conducted a comparison of claim analysis results spanning 797 claim-publisher pairs, of which 294 are from single-stage pipeline and 503 from multi-stage pipeline. Both single and multi-stage results were generated across the same 5 control news events.

Quantitative testing was not done for sentiment due to time constraints, combined with increased difficulty due to level of subjectiveness.

Each claim-publisher pair was manually reviewed for correctness. We classified errors into two categories: minor errors (positive or not mentioned classified as partial) and major errors (false positives/negatives). Results were grouped by event to enable direct comparison between the two architectures on identical data. Weighted averages were calculated, using claim-publisher pair counts for error rates, and distinct claim counts for rows affected (Row refers to a distinct claim, and it's corresponding claim-publisher pairs).

4.2 Results

The multi-stage pipeline achieved 4.67% error rate versus the 33.0% error rate of the single-stage pipeline.2

The results table 2 shows results across the five test news events. Each percentage represents the proportion of claim-publisher pairs that were incorrectly classified. For example, in "Russia-China summit" with 5 publishers, single-stage misclassified 32.5% of all claim-publisher pairs while multi-stage achieved 0% error.

Major errors (false positives/negatives) are critical misclassifications where claims are marked "confirmed" when absent or "not mentioned" when present. Minor errors involve "partial" misclassifications. The multi-stage pipeline reduced major errors from 28.0% to 0.79%.

Rows affected shows the percentage of claims with at least one error across publishers. Single-stage produced errors in 95.3% of claims versus 21.5% for multi-stage, demonstrating more localized error patterns.

The improvement was consistent across all five news events. The most dramatic gain was the 35-fold reduction in major errors.

5 Discussion

Our results demonstrate that LLM prompt architecture fundamentally impacts LLM classification accuracy in cross-source news analysis. Significant error reduction validates task decomposition as a critical design principle for complex NLP pipelines.

While the multi-stage pipeline (Figure 2) requires more API calls (8+ versus one), costs remain manageable at $0.008-0.015 per event with both modules enabled. The accuracy improvement justifies this modest cost increase, especially considering manual verification would require expensive human labor. Considering that an event only needs to be analyzed once with no variable cost, this offers a lot of potential for analysis at scale.

Sentiment analysis struggles with irony and implicit criticism, as shown in the Mladina [15] example where selective quoting conveyed negativity despite positive surface language.

Future work includes comprehensive user testing with journalists and researchers, optimization of current modules, and expansion to other languages. We plan structured evaluations to understand how different user groups interpret and act upon cross-source comparisons.

Acknowledgments

The research described in this paper was supported by the TWON project, funded by the European Union under Horizon Europe, grant agreement No 101095095.

References

[1] Yushi Bai et al. 2023. Longbench: a bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

[2] [n. d.] Ground news - breaking news headlines and media bias. Ground News. Retrieved Sept. 7, 2025 from https://ground.news/.

[3] [n. d.] Event registry api documentation. Event Registry. Retrieved Sept. 7, 2025 from https://eventregistry.org/documentation.

[4] [n. d.] Case study: ground news at harford community college - a collaborative mission to modernize media literacy. Library Up. Retrieved Sept. 7, 2025 from https://www.libraryup.org/news-1/case-study-ground-news-at-harford-community-college.

[5] [n. d.] Ground news - media bias and news comparison. West Virginia University Libraries. Retrieved Sept. 7, 2025 from https://libguides.wvu.edu/c.php?g=1204801&p=8818927.

[6] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web* (WWW '14 Companion). ACM, Seoul, Korea, 107–110. ISBN: 978-1-4503-2745-9. doi:10.1145/2567948.2577024.

[7] C.J. Hutto and Eric Gilbert. 2014. Vader: a parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. AAAI Press, 216–225.

[8] [n. d.] Claude code. Anthropic. Retrieved Sept. 7, 2025 from https://claude.ai/code.

[9] Armin Ronacher. [n. d.] Flask. Retrieved Sept. 7, 2025 from https://flask.palletsprojects.com/.

[10] [n. d.] Postgresql. PostgreSQL Global Development Group. Retrieved Sept. 7, 2025 from https://www.postgresql.org/.

[11] [n. d.] Gemini api. Google. Retrieved Sept. 7, 2025 from https://ai.google.dev/.

[12] [n. d.] Openai. OpenAI. Retrieved Sept. 7, 2025 from https://openai.com/.

[13] [n. d.] Gpt-4o mini. OpenAI. Retrieved Sept. 7, 2025 from https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

[14] [n. d.] Gemini 2.5 flash lite. Google. Retrieved Sept. 7, 2025 from https://openrouter.ai/google/gemini-2.5-flash-lite-preview-06-17.

[15] [n. d.] Trump bi ministrstvo za obrambo preimenoval v ministrstvo za vojno. Mladina. Retrieved Sept. 7, 2025 from https://www.mladina.si/243046/trump-bi-ministrstvo-za-obrambo-preimenoval-v-ministrstvo-za-vojno/.