Same Labels, Smaller Bill: An Ablation Study of LLM Claim Classification in Cross-Source News

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

Blaž Škrlj
blaz.skrlj@ijs.si
Jožef Stefan Institute
Jožef Stefan Postgraduate School
Ljubljana, Slovenia

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

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute
Jožef Stefan Postgraduate School
Ljubljana, Slovenia

Abstract

Comparing how news outlets cover the same event requires labelling each claim per outlet; we ask which LLM set-up does this accurately and cheaply. Our prior multi-stage prompt architecture reduced classification errors 7-fold but was evaluated with a single model (gpt-4o-mini), prompt and decoding setting. Here we evaluate 30 configurations in 78 runs, spanning two classification architectures, eight LLMs from five providers (OpenAI, Google, Anthropic, xAI, DeepSeek), prompt variants for both architectures and a temperature sweep, all with reasoning off, minimal or low (grok-4.1-fast at its default reasoning). Using 1,079 manually annotated claim–publisher pairs from 10 Slovenian news events, we report exact and relaxed accuracy, per-class F1 and cost, and compare against a non-LLM sentence-embedding baseline. The best configurations reach 82% exact and 93–95% relaxed accuracy (gemini-3-flash and gemini-2.5-flash-lite with the reduced per-publisher prompt, gpt-5-nano with the label_only_no_rules per-claim prompt); the best budget-tier configurations match or exceed the tested frontier configurations with reasoning off or at low effort; the per-publisher architecture is 2.5–5.2× cheaper at equal or higher accuracy; and every LLM configuration clearly outperforms the embedding baseline (57.5%).

Keywords

News Analysis, Stance Detection, Claim Verification

1 Introduction

Comparing how publishers cover the same event is essential for understanding media bias and information diversity; we ask which LLM set-up does it best at low cost. In our prior work [1], we presented BetweenTheLines, a system that extracts claims from news articles and classifies whether each publisher confirms, partially confirms, contradicts, or omits each claim. The key finding was that a multi-stage LLM architecture (separate extraction and classification prompts) reduced the rate of reviewer-flagged errors from 33.0% to 4.67% over 797 claim–publisher pairs, compared with a single prompt that reads all articles at once.

Those experiments, however, used a fixed configuration (gpt-4o-mini as the sole model, a single detailed prompt, default temperature), compared the multi-stage design only with a single prompt rather than with the two ways its own classification stage can be organized, and had no non-LLM baseline. How far the results generalize is therefore open, and this study asks five research questions:

RQ1 (Architecture): Does the classification architecture (per_publisher vs per_claim_publisher, Figure 1) affect accuracy and cost?

RQ2 (Models): Do different LLM providers and models perform differently on this task?

RQ3 (Prompts): Do prompt length and explicitly stated classification rules affect accuracy?

RQ4 (Temperature): Does temperature affect classification consistency and accuracy?

RQ5 (Baseline): Can non-LLM baselines compete with LLMs for claim classification?

We contribute a systematic evaluation of 30 configurations in 78 runs across architectures, models, prompts and temperature; an annotated dataset of 1,079 claim–publisher pairs from 10 events; a non-LLM embedding baseline; and practical cost–accuracy guidance. The dataset, evaluation scripts and full per-class metrics are public.[1]

2 Related Work

Our work intersects claim verification and stance detection; Dmonte et al. [2] review LLM-based claim verification, including chain-of-thought prompting and retrieval-augmented methods. Automated fact-checking asks whether a claim is true or supported by evidence [3]; claim comparison asks instead whether each outlet includes, omits or presents an equivalent claim differently, with the news event as the unit of comparison [4].

LLMs exhibit notable sensitivity to prompt formulation. Errica et al. [5] show that minor wording changes produce significant performance variation. Regarding decoding settings, Renze and Guven [6] find that sampling temperature between 0 and 1 has no statistically significant effect on multiple-choice problem-solving accuracy across nine LLMs.

[1] https://github.com/GeorgiGeor/datamining-ablations-newsdiff-data

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SiKDD 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.113

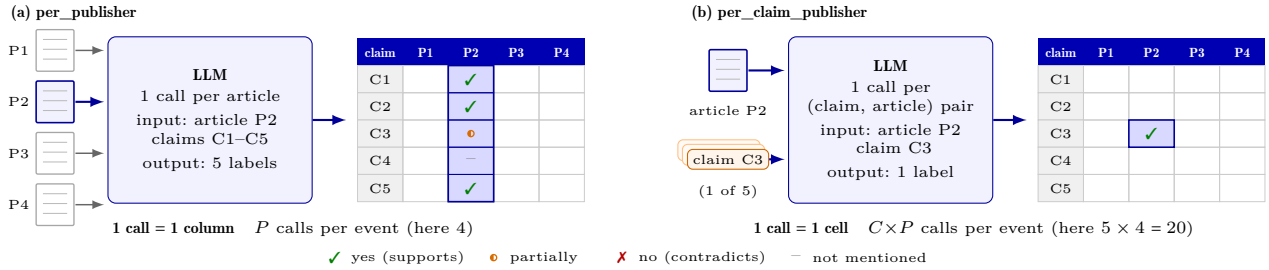


Figure 1: Illustrative classification architectures on a hypothetical event with 4 articles and 5 claims: (a) `per_publisher`, one call per article for all claims; (b) `per_claim_publisher`, one call per (claim, article) pair.

These findings motivate our comparison of prompt variants (RQ3) and of whether the temperature result transfers to our classification task (RQ4). Intermediate labels are hard elsewhere too: in natural language inference the *neutral* class has the lowest annotator agreement [7].

3 Methodology

3.1 Task and Label Definitions

Claim. A claim is a short, self-contained declarative statement drawn from an event's coverage—a factual assertion, an accusation, or a statement attributed to an actor—that can be checked against the text of an individual article. Claims are defined at the *event* level: the extraction stage reads all articles of an event jointly and returns 8–15 such statements, favoring those that are central to the story or reported by several outlets; opinions of the reporting outlet are not claims. For example, one of the events in our dataset (Section 3.2) concerns the EU investigation of the platform X; one of its extracted claims reads "X was already fined 120 million euros in December for transparency violations".

Claim–publisher pair and labels. The unit of evaluation is the pair (claim, article). Its label states whether the article *asserts or supports* the claim (`yes`), *mentions it only indirectly*—in paraphrase, in part, or in a broader or narrower form (`partially`), *contradicts or denies* it (`no`), or *does not address it at all* (`not_mentioned`). A claim may thus be absent from most articles and still be a valid claim; the example above is labelled `yes` for seven of the event's nine publishers and `not_mentioned` for two. A single annotator labelled all 1,079 pairs; the distribution is 52.8% `yes`, 37.5% `not_mentioned`, 9.5% `partially` and 0.1% `no` (n=1). The `no` class is rare because news articles rarely contradict each other directly.

3.2 Dataset

We constructed an evaluation dataset from Slovenian news coverage collected with Event Registry [8], a news aggregator that clusters articles from different outlets into *events* (one news story each): 10 events from January 2026, 87 articles, 124 distinct claims and 1,079 annotated claim–publisher pairs. Claims were extracted once per event with gemini-3-flash-preview from all articles concatenated (about 6,800 input tokens per call); the resulting claim list was annotated and is shared by every configuration, which isolates classification from extraction variability.

The events span diverse topics to test generalization: EU/technology regulation, sports, domestic politics, international relations, and economic policy. Each event contains between 7–10 articles from different Slovenian publishers and 12–13 claims. We denote by P the number of articles (publishers) and by C the number of claims of an event.

3.3 Classification Architectures

Classification is the second, ablated stage: it receives an event's fixed claim list and its articles and labels every (claim, article) pair. Figure 1 contrasts the two ways of doing so.

per_publisher: One LLM call per article classifies all C claims simultaneously. Minimizes API calls and cost (P calls per event).

per_claim_publisher: One LLM call per (claim, article) pair. More granular but requires $C \times P$ calls per event instead of P.

A per_publisher call averages about 1,800 input tokens, a per_claim_publisher call 1,100.

3.4 Models

We evaluate eight LLMs across five providers, organized by the providers' own tiering—flagship, mid-size, small (Table 1), plus a non-LLM embedding baseline. Hereafter we abbreviate gemini-3-flash-preview as gemini-3-flash.

Tier	Model	Provider	Reasoning	T
Frontier	gpt-5.2	OpenAI	low	n/a
Frontier	deepseek-v3.2	DeepSeek	off	0.3
Mid	gpt-5-mini	OpenAI	low	n/a
Mid	claude-sonnet-4.5	Anthropic	off	0.3
Mid	grok-4.1-fast	xAI	on	1
Mid	gemini-3-flash-preview	Google	minimal	1
Budget	gpt-5-nano	OpenAI	low	n/a
Budget	gemini-2.5-flash-lite	Google	off	0/1/2
Baseline	embedding	Local	—	—

Table 1: Models, reasoning setting and temperature (T).

Reasoning effort. Reasoning was left off (provider default) for Claude, DeepSeek and gemini-2.5-flash-lite, set to `minimal` for gemini-3-flash and to `low` for the GPT-5 family (Table 1); the exception is grok-4.1-fast, whose provider default left reasoning enabled. It is a fixed condition of this study, not an ablated factor: all findings hold with reasoning off, minimal or low (Section 5).

Embedding baseline. We implement a non-LLM baseline using sentence-transformers [9] (model `paraphrase-multilingual-MiniLM-L12-v2`) with cosine-similarity thresholds: $>0.55 \rightarrow$ `yes`, 0.35–0.55 $\rightarrow$ `partially`, $\leq 0.35 \rightarrow$ `not_mentioned`. This baseline cannot distinguish `yes` from `no` since embeddings capture relatedness, not stance. Thresholds were fixed a priori and not tuned on the evaluation set.

3.5 Prompt Variants

We test prompt variants appropriate to each architecture. For **per_publisher**: **baseline** (label definitions, a mandatory checklist of transformations—paraphrase, hypernyms, abbreviations, coreference—to check before answering `not_mentioned`, and a tie-break preferring `partially`) and **minimal** (reduced instructions). For **per_claim_publisher**: **baseline** (the same guidance for a single (claim, article) pair), **shortened** (condensed baseline), **label_only_rules** (label-only output with rule guidance), and **label_only_no_rules** (question and labels only). All prompts written in Slovenian to match article language.

3.6 Temperature and Runs

A *configuration* is a combination of architecture, model, prompt variant and temperature; a *run* is one pass of a configuration over all 1,079 pairs. Temperature was swept at 0, 1 and 2 for gemini-2.5-flash-lite only (nine configurations: three temperatures for each of three architecture/prompt cells; Google recommends T=1 for Gemini). All other models ran at a single value (Table 1): 0.3 for Claude and DeepSeek, 1 for Grok and Gemini 3; the GPT-5 family has no temperature parameter. Every configuration was run on the same claims and articles one to three times (Table 2): the T=0 configurations, grok-4.1-fast (retired by its provider during the study) and the embedding baseline once, all other configurations three times. Runs with more than 1% unlabelled pairs were discarded, and two January runs whose exported predictions were partly substituted by embedding-baseline output (an export defect—the underlying API runs were clean) were excluded and replaced by their configurations' next runs. In total, 30 configurations and 78 runs.

3.7 Evaluation Metrics

Exact accuracy: Percentage matching ground truth exactly. **Relaxed accuracy:** Counts adjacent label confusions as acceptable—specifically `yes`↔`partially` and `partially`↔`not_mentioned`. The `no` class requires exact match. **Cost:** USD per run (10 events), computed from the logged token counts of every call and provider list prices; the embedding baseline makes no API calls. All accuracies are pooled over the 1,079 pairs (the mean over events differs by at most 1 pp); configuration-level results are means over a configuration's runs. Exact accuracy on four labels is stricter than the reviewer-judged error rate of the 2025 evaluation; relaxed accuracy is its closer counterpart. **Uncertainty:** The 1,079 pairs are clustered within 10 events and are not independent; a bootstrap that resamples whole events (10,000 resamples) puts the 95% confidence interval of a run's accuracy at about ±6 pp in the median (±4–7 pp at the accuracy levels of Table 2). Configurations are however compared on the same pairs, so we judge differences against run-to-run spread: the standard deviation of a configuration is typically below 1 pp (Table 2) and we read differences of about 2 pp or less as within run-to-run variation. Exact McNemar tests between single runs are provided in the repository.

4 Results

The 78 runs cost $24.5 in total. Table 2 lists the ten best configurations ordered by mean exact accuracy, plus the embedding baseline; they span 2.4 pp, well within the event-level uncertainty (Section 3.7) and close to the run-to-run spread, so their order should not be over-interpreted.

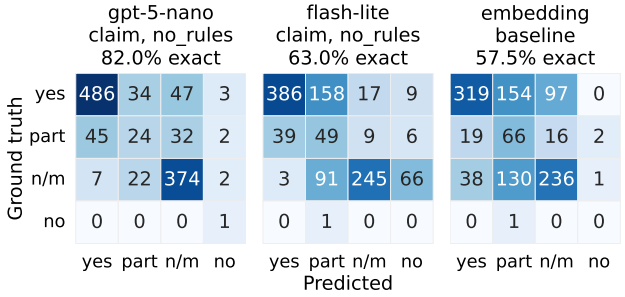


Figure 2: Confusion matrices of three single runs (not configuration means).

#	Model	Arch	Prompt	T	Exact	Relax	Cost	n
1	gemini-3-flash	pub	min	1	82.4±0.2	94.3	$0.21	3
2	gemini-2.5-flash-lite	pub	min	0	81.8	94.6	$0.03	1
3	gpt-5-nano	claim	no_rules	–	81.8±0.6	93.3	$0.14	3
4	gemini-2.5-flash-lite	pub	min	1	81.0±0.2	94.1	$0.03	3
5	gemini-2.5-flash-lite	pub	min	2	80.9±0.6	94.3	$0.03	3
6	gpt-5-mini	pub	base	–	80.5±0.1	96.4	$0.28	3
7	claude-sonnet-4.5	pub	min	0.3	80.4±0.2	95.0	$1.10	3
8	gemini-3-flash	pub	base	1	80.2±0.1	95.9	$0.28	3
9	gemini-2.5-flash-lite	pub	base	1	80.1±0.7	96.0	$0.05	3
10	gpt-5-mini	claim	base	–	80.0±0.4	96.8	$0.70	3
30	embedding	—	—	–	57.5	87.1	—	1

Table 2: The ten best of the 30 configurations by mean exact accuracy (±sd over the n runs); no_rules = label_only_no_rules; cost per run.

RQ1 (Architecture): per_publisher fills eight of the ten places in Table 2, including the two best configurations, and gpt-5-nano (81.8%) and grok-4.1-fast (79.5%) are the only models whose best result comes from per_claim_publisher. Within the same model and prompt, per_publisher is more accurate by 6.6 pp for gemini-2.5-flash-lite (80.1% vs 73.5%), 4.6 pp for gemini-3-flash and 2.8 pp for deepseek-v3.2—well beyond run-to-run variation—and within 1.4 pp for gpt-5-nano, gpt-5-mini and grok. Since per_publisher is also 2.5–5.2× cheaper for the same model (2.5× for gpt-5-mini, 5.2× for deepseek-v3.2) because of its far fewer API calls (Figure 1), it is the better default; the GPT-5 models and grok lose little or nothing when each claim is judged in isolation.

RQ2 (Models): The best observed budget-tier configurations match or exceed the tested frontier configurations with reasoning off or at low effort. The best configuration of each tier reaches 82.4% for Mid (gemini-3-flash, per_publisher minimal), 81.8% for Budget (gemini-2.5-flash-lite, per_publisher minimal, T=0, and gpt-5-nano) and 79.1% for Frontier (gpt-5.2). gpt-5.2 is thus 3.2 pp below the best configuration and 2.6 pp below gpt-5-nano label_only_no_rules, and level with gemini-2.5-flash-lite per_publisher (79.1% vs 80.1%); deepseek-v3.2 (75.9%) is 4.2 pp below the latter. The best cost–accuracy point is gemini-2.5-flash-lite per_publisher with the minimal prompt at $0.03 per run for 81.8%.

RQ3 (Prompts): Both architectures have prompt variants (Section 3.5). For per_claim_publisher, verbosity has no consistent effect but the classification rules do: for gpt-5-nano the `label_only_no_rules` prompt scores 4.2 pp above baseline (81.8% vs 77.5%), suggesting reasoning models do

well with less prescriptive instructions, while `shortened` gains 2.1 pp; for gemini-2.5-flash-lite, `label_only_rules` (78.3%) and `shortened` (77.2%) beat baseline (73.5%) by 3.7–4.8 pp, but `label_only_no_rules` collapses to 64.0%. Keeping the rule checklist matters more than verbosity, and how much depends on the model. For per_publisher the reverse holds: the minimal prompt scores 0.9–2.2 pp above the baseline mean for every Gemini configuration and 0.7 pp for Claude—at the edge of run-to-run variation—whereas for gpt-5-nano it is 2.2 pp worse (76.7% vs 78.9%).

RQ4 (Temperature): No detectable effect for gemini-2.5-flash-lite, the only model swept. With per_publisher and the baseline prompt, the T=0 run (79.6%) lies within the spread of the six runs at T≥1 (76.5–80.6%, mean 79.5%, sd 1.5); the T=1 and T=2 means differ by 1.1 pp in one direction for per_publisher (80.1% vs 79.0%) and, for per_claim_publisher, by 1.1 pp in the other (73.5% vs 74.6%), and the minimal prompt gives 81.8%, 81.0% and 80.9% at T=0, 1 and 2.

RQ5 (Baseline): LLMs substantially outperform embedding similarity. The LLM configurations reach 82% exact accuracy vs 57.5% for embeddings (+24.8 pp for the best configuration). The gap is largest for F1-not_mentioned (0.89 vs 0.63), where embeddings struggle to distinguish semantic absence from low similarity.

Per-class performance: F1-no is not meaningful because of the single `no` sample, which 35 of the 78 runs label correctly. The `partially` class remains the hardest (best F1 = 0.41 in any run), consistent with the neutral-class difficulty noted in Section 2, while `yes` and `not_mentioned` reach F1 > 0.84 in every run of the ten best configurations.

Error patterns: The gap between exact and relaxed accuracy reveals classification style (Figure 2). We identify three patterns: (1) *Decisive* models like gpt-5-nano (gap=11.6%) and grok per_publisher (78.4% exact, gap 13.7%) commit to clear labels with few yes→partial errors (27 and 67 respectively). (2) *Balanced* models like gemini-2.5-flash-lite per_publisher (gap=15.9%) show moderate hedging. (3) *Conservative* configurations such as gemini-2.5-flash-lite with the `label_only_no_rules` per_claim_publisher prompt (gap=26.5%, about 155 yes→partial per run) inflate relaxed accuracy through excessive hedging; with the rule-bearing prompts the same model hedges far less. The embedding baseline fails differently: 130 not_mentioned→partial errors from inability to distinguish semantic absence from low similarity.

5 Conclusion, Limitations and Future Work

On 1,079 annotated claim–publisher pairs, the per_publisher architecture with the reduced prompt on gemini-3-flash or gemini-2.5-flash-lite gives the best accuracy of the study (82%) at a few cents per event set; the frontier models tested do not outperform it with reasoning off or at low effort; whether a prompt states the classification rules matters more than its length, or than the temperature for the one model swept; and every LLM configuration is above the embedding baseline, the best by 25 pp.

Annotation, language and contradiction: Ground truth was created by a single annotator, so inter-annotator agreement cannot be quantified, and `partially` is subjective on both of its boundaries, with `yes` and with `not_mentioned` (36% of wrong `partially` labels fall on `not_mentioned` pairs); misclassification can falsely attribute a stance to a publisher or over-read `partially` as endorsement, so high-stakes uses need human review. The claims were extracted by gemini-3-flash (Section 3.2), itself one of the classifiers, which may favour its own phrasing; its first place should be confirmed with another extractor. Results are specific to Slovenian news. The extremely rare `no` class (0.1%, n=1) prevents meaningful evaluation of contradiction detection—the key capability distinguishing LLMs from embedding baselines; in FNC-1 the analogous *disagree* class remained the hardest even with hundreds of examples [10].

Reasoning budget and run coverage: Every model except grok-4.1-fast was evaluated with reasoning off, minimal or low; the finding that budget-tier models match frontier models therefore holds *at low deliberation*. The constraint is not neutral: models built around deliberation (the GPT-5 family, gpt-5.2 in particular) plausibly lose more from it than the others, so RQ2 is the finding most exposed to it and should be read as a statement about low-deliberation operation, not about model capability. Runs were made without seed control; the T=0 configurations and grok-4.1-fast were run only once (Section 3.6).

Three directions would strengthen these findings: multiple annotators, to measure agreement on both boundaries of `partially`; a decision on whether a contradiction (`no`) class is meaningful for news coverage at all and, if it is, events curated to contain explicit contradictions; and a sweep of reasoning effort (low/medium/high) for the top configurations, to test whether the budget-versus-frontier finding survives higher deliberation.

Acknowledgments

The research was supported by the Slovenian Research and Innovation Agency through the research program Knowledge Technologies (P2-0103) and the Gravity project AI for Science (GC-0001) and European HE projects TWON (GA 101095095) and Raido (GA 101135800). The authors acknowledge usage of LLMs in grammar refinement and overall manuscript organization and take full responsibility for the final content of this publication.

References

[1] Georgi Trajkov, Marko Grobelnik, and Adrian Mladenić Grobelnik. 2025. BetweenTheLines - cross source news analysis. In *Proceedings of the 28th International Multiconference (SiKDD)*. https://is.ijs.si/wp-content/uploads/2025/11/IS2025_Volume-C.pdf.
[2] Alphaeus Dmonte et al. 2024. Claim verification in the age of large language models: a survey. arXiv: 2408.14317 [cs.CL].
[3] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10, 178–206.
[4] Damian Trilling and Marieke van Hoof. 2020. Between article and topic: news events as level of analysis and their computational identification. *Digital Journalism*, 8, 10, 1317–1337.
[5] Federico Errica et al. 2025. What did i do wrong? quantifying LLMs' sensitivity and consistency to prompt engineering. In *Proceedings of NAACL-HLT (Volume 1: Long Papers)*, 1543–1558.
[6] Matthew Renze and Erhan Guven. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of EMNLP*, 7346–7356.
[7] Samuel R. Bowman et al. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, 632–642.
[8] Gregor Leban et al. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web* (WWW '14 Companion), 107–110.
[9] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, 3982–3992.
[10] Andreas Hanselowski et al. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of COLING*, 1859–1874.