Evidence-Constrained NLI-Based Fact Verification with LLMs with Routed Knowledge-Graph Retrieval

Tala Aljibbaoui[1,*], Maximos Skandalis[1] and Michael Sioutis[1]

[1]LIRMM, CNRS and University of Montpellier, Montpellier, France

Abstract

Large language models can assign fact-verification labels from parametric knowledge even when no admissible evidence is available. We study an evidence-constrained pipeline that routes claims to heterogeneous structured sources, verbalises retrieved graph relations, ranks candidate premises, and asks an LLM to predict *Entailment*, *Contradiction*, or *Neutral* using only those premises. Encyclopedic, lexical, and event knowledge graphs are separated from indexed professional fact checks because they cover different claim regimes. On a 30-claim FEVER subset, the best observed retrieval configuration raises macro-F1 over an LLM-only baseline for LLaMA-4-Scout (+12.2%), Command-R (+9.2%), and LLaMA-70B (+5.1%), but degrades LLaMA-8B (−24.3%). On a separate 195-claim multilingual news stress test without gold labels, retrieval lowers the Entailment rate and increases Neutral predictions for all evaluated models. This distributional shift indicates stronger conditioning on explicit evidence, but cannot by itself establish improved factual accuracy or reduced hallucination. Error analysis identifies entity-linking noise and weak rejection of irrelevant premises as central failure modes. The results support routed structured retrieval as a promising control mechanism while delimiting the evaluation and provenance requirements needed for reliable automated fact checking.

Keywords

automated fact checking, natural language inference, knowledge graphs, retrieval-augmented generation, evidence grounding, large language models

1. Introduction

Automated fact checking is not merely a binary classification problem: a defensible verdict must be conditioned on relevant evidence and should remain undecided when the available evidence is insufficient [1]. This requirement is especially important for large language models (LLMs), whose fluent outputs may conceal whether a decision follows from supplied evidence or from untraceable parametric associations. Retrieval-augmented generation (RAG) addresses part of this problem by coupling a language model to external memory [2]; nevertheless, retrieval can introduce irrelevant or incomplete context, and the model may still override it.

This paper studies a structured, evidence-constrained NLI-based verifier. Given a claim, the system retrieves facts from routed knowledge sources, turns graph relations into textual premises, selects premises by semantic similarity or lexical overlap, and instructs an LLM to return one of three natural-language-inference (NLI) labels. The contribution is empirical: (i) a modular pipeline combining source routing, graph construction, premise ranking, and evidence-only inference; (ii) a controlled comparison of LLM-only and retrieval-conditioned decisions on small French and FEVER samples; and (iii) a stress-test analysis that distinguishes desirable abstention from unverified claims of accuracy improvement.

We use *GraphRAG* in the broad sense of retrieval over graph-structured evidence. This differs from the system of Edge et al. [3], which constructs hierarchical entity graphs and community summaries

Graph-enhanced LLMs for trustwOrthy Web data management (GLOW), in ISWC 2026 Workshops Joint Proceedings, October 25–26, 2026, Bari, Italy

*Corresponding author.

tala.aljibbaoui@etu.umontpellier.fr (T. Aljibbaoui); maximos.skandalis@cnrs.fr (M. Skandalis); michael.sioutis@umontpellier.fr (M. Sioutis)

https://mskandalis.github.io (M. Skandalis); https://msioutis.gitlab.io (M. Sioutis)

0000-0003-2361-5115 (M. Skandalis); 0000-0001-7562-2443 (M. Sioutis)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

CEUR-WS.org/Vol-4313/glow-paper3.pdf

over a text corpus for global question answering. The present method is more precisely a *knowledge-graph-backed, routed RAG* architecture for claim verification.

2. Related Work

FEVER established large-scale three-way verification against Wikipedia, with labels Supported, Refuted, and NotEnoughInfo [4]. AVeriTeC subsequently shifted attention toward real-world claims and web evidence, where evidence availability, temporal validity, and provenance are harder to control [5]. These benchmarks expose a key distinction: the verifier must assess whether a claim follows from admissible evidence, not whether it appears plausible in isolation.

Graph-based fact verification has primarily represented relations among retrieved evidence sentences or semantic units. Semantic-level evidence graphs improve multi-evidence reasoning [6], whereas FactKG directly evaluates linguistic claims against knowledge-graph reasoning patterns [7]. Pratapa et al. [8] explicitly impose closed-world reliance on extracted FEVER evidence, closely matching our evidence-relative label semantics, although their constraints target text retrieval rather than heterogeneous knowledge sources. The present system is complementary: it does not train a graph neural verifier, but exposes verbalised graph facts to general-purpose LLMs under an evidence-only decision rule.

The retrieval layer combines sources with different semantics. Wikidata provides collaboratively maintained entity facts [9]; YAGO 4 aligns a knowledge base with ontological constraints [10]; BabelNet supplies multilingual lexical and encyclopedic relations [11]; and EventKG represents events and temporal relations [12]. The Google Fact Check Tools API indexes published fact-checks rather than general world knowledge [13]. Their coverage is therefore non-substitutable, motivating source routing. Sentence-level semantic ranking follows the dense similarity paradigm exemplified by Sentence-BERT [14].

3. Routed Evidence-Constrained Verification

Let c be a claim and $\mathcal{Y} = \{E, C, N\}$ the labels Entailment, Contradiction, and Neutral. The intended semantics are evidence-relative:

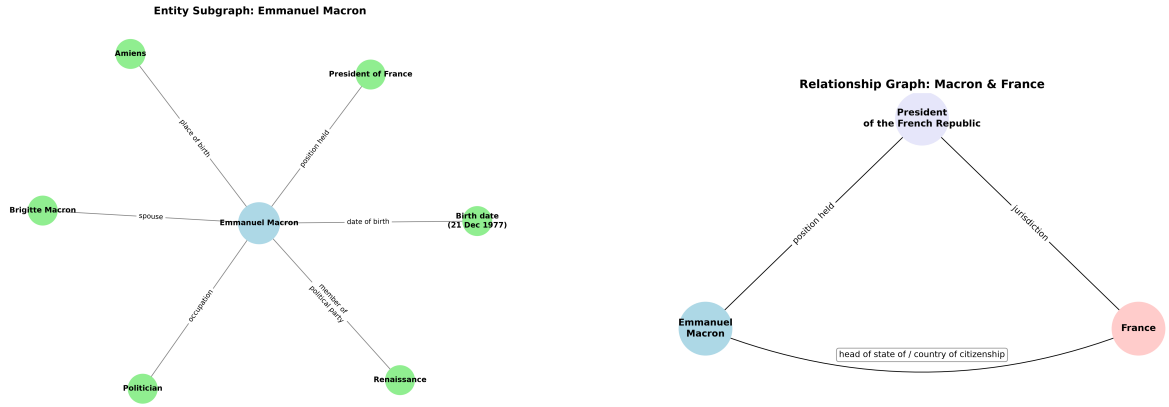
$$E : P \models c, \qquad C : P \models \neg c, \qquad N : P \not\models c \ \wedge \ P \not\models \neg c,$$

where P is the retrieved premise set. Thus, Neutral denotes insufficient retrieved support, not proof that the claim is unknowable or false. This distinction is necessary under the open-world assumption of public knowledge graphs: absence of a triple is not a negative fact.

Extraction and routing. A spaCy-based front end extracts salient entities and nominal concepts and attempts to link them to Wikidata identifiers. A semantic router then selects one of two source families. *KG1* comprises Wikidata, YAGO, BabelNet, and EventKG and targets stable encyclopedic, lexical, and event-centred claims. *KG2* queries indexed professional fact checks (PolitiFact, Snopes, FactCheck.org) through the Google API and targets rumours or previously investigated claims. The router is intended to avoid querying every source indiscriminately.

To visualise the decision process, a complete execution flow example is provided in Figure 1. For the input claim stating that Emmanuel Macron is the president of France, the system extracts the entities Macron and France. It then queries Wikidata to construct a graph connecting these two nodes by edges such as presidential function or nationality. These edges are translated into textual premises, which are finally submitted to the generative model to obtain the final verdict validating Entailment.

Graph construction and premise selection. Retrieved entities and relations are assembled into a local NetworkX graph. Each selected relation is verbalised as a short premise. Candidate premises



(a) Subgraph for the entity (Emmanuel Macron).

(b) Relationship graph (Macron and France).

Figure 1: An example of knowledge graph visualisations for Emmanuel Macron.

are ranked either by keyword intersection with the claim or by cosine similarity between sentence embeddings. The top-ranked premises form P. Semantic ranking is expected to improve recall beyond exact lexical overlap, but similarity alone does not guarantee logical relevance.

Evidence-only inference. The claim and P are supplied to an open LLM with an instruction to use no external or parametric knowledge and to output one label. The implementation evaluated Command-R, LLaMA variants, and Qwen through hosted inference. This prompt is a behavioural constraint rather than a formal guarantee: a model can still import memorised facts, overfit lexical cues, or rationalise irrelevant context; LLM distractibility under irrelevant input is independently documented [15]. Consequently, retrieval quality and the verifier's context-rejection ability must be evaluated separately.

4. Experimental Design

Three evaluations were conducted. First, a 25-claim French pilot compared Wikidata and BabelNet, keyword and semantic selection, and several LLMs. Second, 30 English FEVER claims enabled comparison with LLM-only baselines. Macro-F1 is emphasised because label imbalance and model-specific collapse toward Neutral can make accuracy misleading.

Third, more than 260 recent articles from English (FoxNews, DailyMail, The Sun, The Mirror) and French (CNews) news sources, a French parody outlet (Le Gorafi), and recent AVeriTeC material were processed. LLaMA-3.3-70B was used to extract three to seven short claims per article; deduplication yielded 195 claims. No human gold labels were reported for this set. It is therefore treated as a *stress test of prediction behaviour*, not as an accuracy benchmark.

5. Results and Analysis

5.1. Controlled samples

Before confronting real-time news data, the system was calibrated on French and English samples from reference corpora. A first series of tests on 25 French claims allowed us to isolate the most effective premise selection method between semantic similarity and simple keyword intersection. The use of semantic similarity on the Wikidata base proved superior. The evaluation was then extended to the standard FEVER corpus through 30 English claims, offering a direct comparison between isolated models and the GraphRAG solution. Table 1 shows the results for two LLMs (Command-R and three different

Base	Method	LLM	French		English	
			Accuracy	F1	Accuracy	F1
wikidata	similarity	command-r	**64%**	0.47	56.7%	0.578
wikidata	similarity	llama-70b	60%	**0.61**	50.0%	0.471
wikidata	similarity	llama-8b	56%	0.38	53.3%	0.422
wikidata	keyword	command-r	52%	0.39	**60.0%**	**0.614**
wikidata	keyword	llama-70b	48%	0.51	56.7%	0.566
wikidata	keyword	llama-8b	52%	0.42	53.3%	0.424
wikidata	similarity	llama4-scout	36%	0.40	63.3%	0.579
wikidata	keyword	llama4-scout	40%	0.44	63.3%	0.576
babelnet	similarity	llama-70b	40%	0.44	53.3%	0.455
babelnet	keyword	llama-70b	44%	0.47	50.0%	0.434

Table 1
Performance on the French (25 claims) and English (30 claims) datasets, per candidate premise ranking method.

Model	Base Accuracy	Base F1	RAG F1	ΔF1
LLaMA-4-Scout	46.7%	0.457	0.579	+0.122
Command-R	53.3%	0.522	0.614	+0.092
LLaMA-70B	56.7%	0.515	0.566	+0.051
LLaMA-8B	66.7%	0.667	0.424	-0.243

Table 2
FEVER subset: LLM-only baseline versus best observed retrieval configuration. Scores are reproduced unchanged.

Configuration (Wikidata)	Entailment	Contradiction	Neutral	Comment
command-r keyword	8	14	8	most balanced
command-r similarity	7	14	9	balanced
llama-70b keyword	10	8	12	well distributed
llama-70b similarity	10	5	15	correct
llama4-scout keyword	7	3	20	too much Neutral
llama4-scout similarity	7	2	21	too much Neutral
llama-8b keyword	10	0	20	never predicts Contradiction
llama-8b similarity	13	0	17	never predicts Contradiction

Table 3
Performance per label on the English claim sample.

versions of LLaMA) on these samples of French and English claims, per candidate premise ranking method as mentioned in Section 3.

In the French pilot, the highest accuracy was 64% for Wikidata similarity with Command-R, whereas the highest macro-F1 was 0.61 for Wikidata similarity with LLaMA-70B. The mismatch between the two metrics already cautions against identifying a single universally best configuration.

Table 2 compares each model's LLM-only baseline with its best observed GraphRAG macro-F1 on the 30 FEVER claims. Retrieval improves three models, most strongly LLaMA-4-Scout, but substantially harms LLaMA-8B. The latter never predicts Contradiction in the reported Wikidata configurations and shifts heavily toward Neutral (see Table 3, indicating label collapse rather than calibrated abstention. The result supports a conditional conclusion: structured evidence can improve verification when the underlying model can discriminate relevant from irrelevant premises; retrieval is not a substitute for that capability.

Wikidata generally outperforms BabelNet in these tests, plausibly because the evaluated claims require entity-level factual relations rather than lexical senses. This interpretation remains tentative: retrieval recall and premise precision were not measured, so the downstream difference cannot be uniquely attributed to source quality.

Model	Mode	Entailment	Neutral	Contradiction
LLaMA-4-Scout	LLM only	77%	19%	4%
	+ retrieval	61%	26%	13%
Qwen-32B	LLM only	67%	15%	17%
	+ retrieval	57%	28%	15%
LLaMA-8B	LLM only	93%	5%	1%
	+ retrieval	90%	9%	0%

Table 4
Verdict distributions on the 195-claim stress test. The set has no reported gold labels.

5.2. Unlabelled news stress test

For the evaluation on the 195 extracted claims, Table 4 reports verdict distributions before and after retrieval. Entailment falls for all models, while Neutral rises by 7, 13, and 4 percentage points for LLaMA-4-Scout, Qwen-32B, and LLaMA-8B, respectively. This is consistent with the evidence-only instruction: when a routed source returns no probative premise, a compliant model should select Neutral rather than validate the claim from memory. Yet the distribution alone cannot distinguish correct abstention from excessive abstention. Likewise, LLaMA-4-Scout's increase in Contradiction may reflect useful counter-evidence, retrieval noise, or a changed decision prior. Gold labels and evidence annotations are required to decide among these explanations.

5.3. Failure mechanism: entity-linking noise

The clearest observed failure originates before reasoning. For the AVeriTeC claim "In a letter to Steve Jobs, Sean Connery refused to appear in an apple commercial", the extractor selected the token "In" and linked it to indium. Wikidata then returned chemically valid but claim-irrelevant properties. A smaller model generated an Entailment judgement from this noise, whereas larger models returned Neutral. The example shows that graph provenance does not imply relevance: every premise may be factually correct while the premise set is logically useless. Robust verification therefore requires entity-link confidence, type constraints, relation-level relevance tests, and an explicit *no usable evidence* path before LLM inference.

6. Conclusion, Limitations and Future Work

The study establishes feasibility, not a definitive accuracy gain. The labelled evaluations contain only 25 and 30 claims; configuration selection is not separated from evaluation; model versions, decoding settings, prompts, retrieval depth, and similarity thresholds are incompletely documented; and FEVER claims may occur in model pretraining data. Public SPARQL and inference APIs introduced timeouts, rate limits, and changing availability, weakening reproducibility. The 195-claim set mixes sources and automatically extracted claims without human labels, and its verdict shift cannot be equated with hallucination reduction. Finally, the system does not report retrieval recall, evidence precision, calibration, or provenance completeness.

Within these limits, the central result is informative: routed structured retrieval changes LLM decision policies, often replacing unsupported Entailment with Neutral, but its benefit is strongly model-dependent and can reverse under noisy context. A rigorous next evaluation should freeze source snapshots, publish prompts and routing rules, annotate evidence and verdicts, separate development from testing, and report both retrieval and classification metrics. A locally hosted graph would remove API instability; a larger French NLI resource such as DACCORD [16, 17] could support larger evaluation. The appropriate claim is therefore not that the system is a "cognitive firewall", but that evidence-constrained, routed retrieval is a testable mechanism for making unsupported decisions more observable and, under suitable retrieval and reasoning conditions, less frequent.

Acknowledgements

Maximos Skandalis was supported by the French Ministry of the Armed Forces through the Defence Innovation Agency (AID-DGA), and by the ICO, *Institut Cybersécurité d'Occitanie*, funded by *Région Occitanie*, France. Michael Sioutis holds a Tenure-track Junior Professor (CPJ) Chair in Hybrid AI and is supported by the ANR and the I-SITE program of excellence of the University of Montpellier.

Declaration on Generative AI

Given the topic of the paper, generative AI models (LLaMA, Command-R, Qwen) were used as part of the claim extraction pipeline and the experimental evaluation in this research work.

During the preparation of this manuscript, OpenAI ChatGPT (GPT-5.6 Sol) was used to identify potential grammatical, syntactic, spelling, and/or phrasing issues. Suggested changes were then manually incorporated, after review and where appropriate, by the authors.

References

[1] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206. URL: https://aclanthology.org/2022.tacl-1.11. doi:10.1162/tacl_a_00454.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[3] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph rag approach to query-focused summarization, 2025. URL: https://arxiv.org/abs/2404.16130. arXiv:2404.16130.

[4] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074. doi:10.18653/v1/N18-1074.

[5] M. Schlichtkrull, Z. Guo, A. Vlachos, Averitec: A dataset for real-world claim verification with evidence from the web, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 65128–65167. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/cd86a30526cd1aff61d6f89f107634e4-Paper-Datasets_and_Benchmarks.pdf.

[6] W. Zhong, J. Xu, D. Tang, Z. Xu, N. Duan, M. Zhou, J. Wang, J. Yin, Reasoning over semantic-level graph for fact checking, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6170–6180. URL: https://aclanthology.org/2020.acl-main.549. doi:10.18653/v1/2020.acl-main.549.

[7] J. Kim, S. Park, Y. Kwon, Y. Jo, J. Thorne, E. Choi, FactKG: Fact verification via reasoning on knowledge graphs, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 16190–16206. URL: https://aclanthology.org/2023.acl-long.895. doi:10.18653/v1/2023.acl-long.895.

[8] A. Pratapa, S. M. Jayanthi, K. Nerella, Constrained Fact Verification for FEVER, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7826–7832. URL: https://aclanthology.org/2020.emnlp-main.629. doi:10.18653/v1/2020.emnlp-main.629.

[9] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85. URL: https://doi.org/10.1145/2629489. doi:10.1145/2629489.

[10] T. Pellissier Tanon, G. Weikum, F. Suchanek, Yago 4: A reason-able knowledge base, in: A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, M. Cochez (Eds.), The Semantic Web, Springer International Publishing, Cham, 2020, pp. 583–596.

[11] R. Navigli, S. P. Ponzetto, Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial Intelligence 193 (2012) 217–250. URL: https://www.sciencedirect.com/science/article/pii/S0004370212000793. doi:https://doi.org/10.1016/j.artint.2012.07.001.

[12] S. Gottschalk, E. Demidova, Eventkg: A multilingual event-centric temporal knowledge graph, in: A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, M. Alam (Eds.), The Semantic Web, Springer International Publishing, Cham, 2018, pp. 272–287.

[13] Google, Fact check tools api, Developer documentation, 2025. URL: https://developers.google.com/fact-check/tools/api, accessed 2026-07-12.

[14] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[15] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, D. Zhou, Large language models can be easily distracted by irrelevant context, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 31210–31227. URL: https://proceedings.mlr.press/v202/shi23a.html.

[16] M. Skandalis, R. Moot, C. Retoré, S. Robillard, New datasets for automatic detection of textual entailment and of contradictions between sentences in French, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 12173–12186. URL: https://aclanthology.org/2024.lrec-main.1065.

[17] M. Skandalis, R. Moot, S. Robillard, DACCORD : un jeu de données pour la détection automatique d'énonCés COntRaDictoires en français, in: C. Servan, A. Vilnat (Eds.), Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, ATALA, Paris, France, 2023, pp. 285–297. URL: https://aclanthology.org/2023.jeptalnrecital-long.22.