PERSEUS: Reducing LLM Hallucinations in Multimodal Knowledge Graph Construction and Ontology Learning

Aryan Singh Dalal[1], Hande Küçük McGinty[1,*]

[1]Department of Computer Science, Kansas State University, Manhattan, KS, USA

Abstract

Automated knowledge graph (KG) construction and ontology learning from multimodal sources are severely limited by LLM hallucinations: plausible but unsupported facts produced during extraction. We present PERSEUS, a multimodal knowledge graph construction framework that couples LLM-based open information extraction with explicit triple verification against source evidence. PERSEUS ingests text, images (via OCR and captioning), and audio (via transcription), unifies them as machine-readable text, and uses an LLM to propose candidate triples. Each candidate is validated through a two-stage gate: hybrid sparse-and-dense retrieval followed by a Natural Language Inference (NLI) entailment check against its original local context. On the CaRB open information extraction benchmark, PERSEUS improves precision by 11% over direct LLM extraction and reduces false positives by 54.6%, enabling reliable multimodal knowledge graph construction and ontology learning with less human review across text, image, and audio modalities.

Keywords

Multimodal knowledge graph construction, LLM hallucination reduction, Ontology learning, Natural Language Inference (NLI) verification, Hybrid retrieval, Open information extraction, Automated triple extraction, Retrieval-augmented fact verification

1. Introduction

Knowledge graph construction and ontology learning are central to modern AI, enabling structured reasoning over heterogeneous information [1], yet these processes have historically required slow, expert-driven curation [2]. Large Language Models (LLMs) offer a scalable route to automated triple extraction and open information extraction, but LLM hallucinations, factual errors that appear plausible yet lack supporting evidence, limit deployment in high-stakes domains [3]. This problem intensifies in multimodal knowledge graph construction, where source information is spread across text, images, and audio, and errors from any single modality can propagate into the final KG.

Retrieval-Augmented Generation (RAG) mitigates this by grounding the LLM in retrieved source passages [4]. However, RAG assumes that supplying relevant context is sufficient to guarantee factual output. That assumption does not hold: an LLM can still misread, overlook, or incorrectly synthesize information from a correct source, so retrieval alone leaves no explicit logical verification step in the pipeline.

We address this gap with PERSEUS (PERceptual Semantic Extraction & Unified System), an end-to-end framework for multimodal knowledge graph construction and ontology learning with per-triple hallucination verification. PERSEUS (i) ingests multimodal data from text, images, and audio, (ii) extracts candidate triples with an LLM using an open information extraction prompt, and (iii) validates every triple through a two-stage gate that combines hybrid retrieval with NLI-based entailment checking against the original local context. Unlike existing LLM-driven ontology learning and KG pipelines that rely on post-hoc human review or judge LLMs, PERSEUS performs logical entailment testing at the triple level at runtime. On the CaRB [5] open information extraction benchmark, this produces an 11%

LLM4KGOE'26: Workshop on LLM-driven Knowledge Graph and Ontology Engineering, Co-located with ESWC 2026
*Corresponding author.
aryand@ksu.edu (A. S. Dalal); hande@ksu.edu (H. K. McGinty)
https://aryansinghdalal.my.canva.site/ (A. S. Dalal)
0000-0003-0720-7306 (A. S. Dalal); 0000-0002-9025-5538 (H. K. McGinty)

precision gain and a 54.6% reduction in false positives, at a deliberate cost in recall that is appropriate when factual reliability matters more than coverage.

2. Related Work: LLM-Based Knowledge Graph Construction and Hallucination Mitigation

Using LLMs for automated KG construction is an active area but is challenged by hallucination [6, 3]. RAG grounds generation in retrieved evidence [4], and its effectiveness depends on retriever quality. Sparse retrievers such as BM25 excel at exact token matches but miss paraphrasing [7], while dense encoders capture meaning but can fail on rare or domain-specific terms [8]. Hybrid retrievers combine both, most commonly fused by Reciprocal Rank Fusion (RRF) [9, 10], which motivates our design. Natural Language Inference (NLI) provides a logical backbone for verification by testing whether a premise entails a hypothesis [11]. Despite this, existing LLM-based ontology engineering systems rely on informal validation, human post-hoc review, or judge LLMs rather than runtime entailment checks [12, 13]. Recent multimodal KG pipelines show the value of ingesting images and other modalities: VaLiK converts images to text via vision-language models and prunes by similarity [14]; GOSyBench induces reaction graphs from chemistry documents and verifies offline against a gold KG [15]; mKG-RAG and MR-MKG use multimodal KGs for QA but treat retrieval quality and scoring as the main safeguards [16, 17]. PERSEUS differs by coupling multimodal ingestion (text, images, audio) with hybrid retrieval and a per-triple NLI entailment gate in a single pipeline.

3. The PERSEUS Framework

3.1. Multimodal Input Processing: OCR, Captioning, and Audio Transcription

PERSEUS accepts text, image, and audio inputs and converts all to text. Audio is transcribed with OpenAI Whisper [18], selected for its robustness to noisy, multi-speaker, and accented inputs and for producing coherent long-form transcripts that preserve the sentence boundaries needed by downstream extraction. For images containing text, docTR performs OCR; it was selected over Tesseract and EasyOCR after benchmarking on CORD-v2, where docTR achieved a lower Character/Word Error Rate at higher throughput. For images without text, BLIP2-OPT [19] generates captions; in a 200-image MS-COCO 2017 evaluation against five human reference captions per image, BLIP2-OPT achieved the highest CLIPScore, SBERT semantic similarity, and ROUGE-L among six candidates (GIT-base, GIT-large, BLIP-base, BLIP-large, ViT-GPT-2, BLIP2-OPT), with BLIP-large the runner-up. The selection criterion across all three modalities was the same: highest semantic faithfulness to the source under realistic conditions, since downstream verification cannot recover information lost in the input layer.

3.2. Text Preprocessing and Unicode Normalization

Acquired text is cleaned by removing boilerplate and applying Unicode NFKC normalization (smart quotes, dashes, compatibility characters). Sentence boundaries are detected using regex rules augmented by spaCy (`en_core_web_sm`) dependency parsing. This standardizes inputs for downstream LLM extraction and addresses the sensitivity of smaller LLMs to surface noise in otherwise clean text.

3.3. LLM-Based Triple Extraction with Llama3-8B

We selected Llama3-8B [6] after comparing several 7-8B parameter open-weights LLMs on two criteria: (i) factual accuracy of extracted triples against an expert-curated ground-truth set (a quantitative score), and (ii) adherence to structured output instructions, that is, whether the model reliably emits well-formed `[entity1, relation, entity2]` triples without commentary or merged information across sentences (a qualitative score). Llama3-8B led on both criteria; smaller-instruction-tuned alternatives in the same parameter band frequently violated the structural constraint, requiring brittle post-hoc

parsing. The model is prompted to act as a knowledge-graph expert, receives preprocessed text as numbered sentences, and is instructed to extract triples per sentence without merging information across sentences. It returns triples under a "Triples:" header, each formatted as `[entity1, relation, entity2]`, which we parse into a structured list.

3.4. Triple Verification: Hybrid Retrieval and NLI Entailment

Each candidate triple is validated in two stages: hybrid retrieval and entailment checking (Figure 1).

Hybrid Retrieval. For BM25, the Subject, Predicate, and Object are concatenated, tokenized with spaCy, lemmatized, and filtered of stop words and punctuation. For the dense path, S, P, and O are individually encoded with all-MiniLM-L6-v2 and their embeddings averaged to form a unified query vector; this preserves the predicate's semantic weight, which distinguishes, for example, `[Flavonoids, are found in, Fruits]` from `[Flavonoids, prevent, CHD]`. The query is scored against precomputed sentence embeddings via cosine similarity. The two ranked lists are fused with RRF (k=60, a standard default [10]), returning the top three candidate sentences.

Entailment Checking. Candidate sentences are processed in document order. A lexical match that finds the triple's subject and the lemmatized object within a candidate yields a baseline confidence of 0.95. Otherwise the triple is rewritten as a hypothesis and checked with BART-Large-MNLI [20]; entailment probability above 0.7 is accepted [11]. Because real-world facts often span more than one sentence and rely on pronouns or anaphora, an insufficient score triggers a contextual NLI pass that prepends the preceding candidate sentence to the premise. For each candidate, the final confidence is the maximum of the lexical and NLI scores. The candidate with the highest overall confidence becomes the supporting evidence; triples meeting the support threshold are retained as validated. The full per-triple decision flow is given in Algorithm 1.

Algorithm 1 Verification of a Single Triple.

Require: Triple t, ordered candidate sentences C_t
Ensure: Best confidence c^*, method m^*, supporting sentence s^*

$c^* \leftarrow 0$, $m^* \leftarrow$ None, $s^* \leftarrow$ None
for each sentence $s \in C_t$ in document order **do**
  $c_{\text{lex}} \leftarrow 0$; **if** subject of t and lemma(object) co-occur in s **then** $c_{\text{lex}} \leftarrow 0.95$
  Convert t to hypothesis h; set premise $p \leftarrow s$
  $p_{\text{nli}} \leftarrow \text{NLI}(p, h)$
  if $p_{\text{nli}} \leq 0.7$ **and** s has a preceding candidate s' **then**
    $p_{\text{nli}} \leftarrow \text{NLI}(s' \oplus s,\ h)$ ▷ contextual pass
  end if
  $c_{\text{nli}} \leftarrow p_{\text{nli}}$ if $p_{\text{nli}} > 0.7$ else 0
  if $\max(c_{\text{lex}}, c_{\text{nli}}) > c^*$ **then**
    Update c^*, m^*, s^* with the winning channel and sentence
  end if
end for
return (c^*, m^*, s^*)

Thresholds and parameters. The values 0.95 (lexical), 0.7 (NLI entailment), k=60 (RRF), and top-3 candidates per triple reflect a precision-first design. The lexical 0.95 floor only fires when both the subject and the lemmatized object co-occur in a single sentence, a pattern that empirically associates with very high precision in our corpus and serves as a fast path that bypasses NLI for the easy cases. The NLI cutoff of 0.7 follows common practice in zero-shot entailment work [11, 21]: probabilities below this band carry too high a risk of false-positive entailment when the model is applied outside its MNLI training distribution. The RRF parameter k=60 is the literature default [10]; it down-weights documents ranked very high by only one of the two retrievers, so neither BM25 nor dense similarity alone can dominate fusion. Top-3 candidates per triple is a conservative recall budget: it bounds NLI compute while still giving the validator multiple chances to find supporting evidence, and was sufficient on CaRB to preserve all true positives that retrieval surfaced at all.

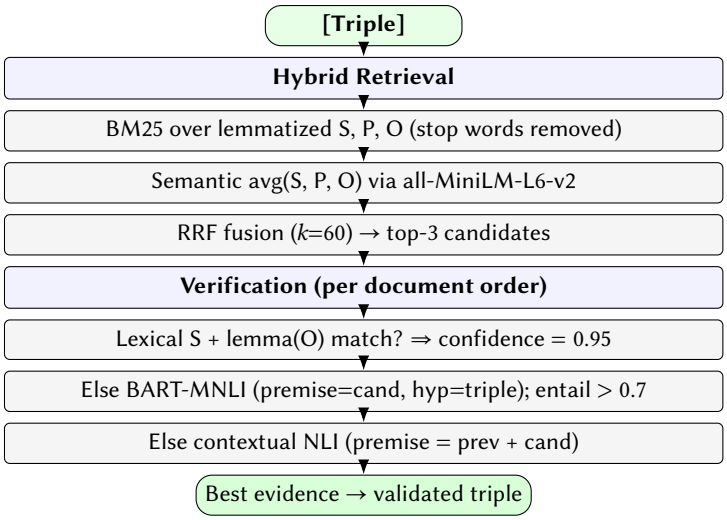

Figure 1: PERSEUS verification pipeline. A candidate triple goes through hybrid retrieval (BM25 over S, P, O lemmas plus dense similarity via all-MiniLM-L6-v2), fused with RRF. The top candidates enter a lexical/NLI cascade; if direct entailment is uncertain, a contextual NLI pass using the preceding sentence is applied. The highest-confidence candidate supports the validated triple.

3.5. Entity Clustering for OWL Class Hierarchy and Ontology Learning

Validated triples feed ontology construction. Unique entities are embedded using Bert-base-uncased via the `[CLS]` token output [22], and a pairwise cosine similarity matrix is z-score normalized with scikit-learn. We compared six clustering algorithms on semantically close phrases (four "anti-* properties" terms from a flavonoid triple set). Affinity Propagation (AP) [23] and Spectral Clustering (SC) [24] grouped all four correctly, while HDBSCAN, DBSCAN, HAC, and DPC either fragmented or marked them as noise. To choose between AP and SC without ground-truth labels, we used three complementary internal metrics: the Silhouette Score [25], the Davies-Bouldin Index, and the Calinski-Harabasz Score. AP self-determines k=4 while SC's k was empirically optimized to 3; the combined metrics favored SC, which we adopt. Each cluster's centroid entity becomes an `owl:Class`, the rest are its subclasses, and LLM-generated `rdfs:comment` annotations plus phonetic labels pass through a human-in-the-loop review step.

4. Evaluation on the CaRB Open Information Extraction Benchmark

We evaluated PERSEUS on the CaRB open-information-extraction benchmark [5] against a baseline that uses direct LLM output without verification.

Table 1
Performance comparison on CaRB: direct LLM extraction vs. PERSEUS.

Metric	Direct	PERSEUS	Δ
Precision	0.65	**0.76**	**+11%**
Recall	0.74	0.57	−17%
F1	0.69	0.65	−4%
True Positives	2232	1722	−510
False Positives	1229	**558**	−54.6%
False Negatives	777	1287	+510

PERSEUS reached 75.5% precision, an absolute 11% gain over the 64.5% baseline (Table 1), driven by a 54.6% drop in false positives (1229 → 558). As expected, stricter filtering reduces recall from 74.2% to 57.2% and F1 by four points.

Statistical significance. A McNemar's test confirmed the precision improvement is not due to chance: $\chi^2(1) = 319.07$, $p < 0.001$. The per-triple disagreement matrix (Table 2) shows where the two pipelines diverge. The off-diagonal cells are the ones that determine the test: PERSEUS overturns 661 of the baseline's correct calls (filtering, often legitimately rejecting weakly supported triples) and recovers 151 triples the baseline missed entirely (cases where verification surfaces evidence the baseline did not connect to a hypothesis). The 4:1 ratio between these counts ($\frac{661}{151} \approx 4.4$) is exactly the source of the recall trade-off: PERSEUS gives back more than it gains in raw count, but every retained triple now carries provenance, which is the property that matters for trustworthy KG construction.

Table 2
McNemar contingency on CaRB. Rows: Direct pipeline outcome (DF = found correct, DM = missed). Columns: PERSEUS outcome (PF = found correct, PM = missed). Off-diagonal cells (PERSEUS-only and Direct-only correct calls) drive the McNemar statistic.

	PF	PM
DF	1571	661
DM	151	626

For applications that prize factual reliability, the trade is net positive: a smaller but substantially more trustworthy KG.

Runtime. End-to-end execution on a representative input took 65.48 s on a workstation with an Intel i7-13700, 32 GB RAM, and an NVIDIA RTX 4080 (16 GB). Triple extraction and parsing dominated (37.47 s), followed by preprocessing (15.44 s) and entity clustering (12.47 s); Llama-3 setup was negligible. As a qualitative demonstration, we also produced an ontology from an audio track of Steve Jobs' Reed College commencement address, transcribed, extracted, verified, and clustered end-to-end.

5. Limitations and Future Work

Four directions guide the next iteration. First, our lexical checker relies on co-occurrence and can over-accept unrelated token pairs; we will integrate dependency parsing and relation-pattern matching so lexical checks respect semantic roles. Second, the static contextual-NLI window (one preceding sentence) can pull in unrelated context; upstream coreference resolution with dynamic windowing should improve long-range discourse handling. Third, fixed thresholds (0.7 for NLI, 0.95 for lexical) will be replaced with adaptive thresholds derived from ROC and PR sensitivity analyses. Fourth, the NLI step is a scalability bottleneck on large corpora; we plan ablation studies to isolate the contributions of hybrid retrieval, preprocessing, and NLI, and will benchmark against adaptive retrieval systems. We also intend to adopt the BenchIEFL protocol for fact-centered evaluation (CaRB's token-level scoring is a known limitation) and to quantify how noise from ASR and OCR propagates into extracted triples. Finally, human-in-the-loop and active-learning workflows will recover valid triples rejected by conservative verification and support schema-compliant population in specialized domains.

6. Conclusion

Factual hallucination is a core obstacle to using LLMs for automated knowledge graph construction and ontology learning. PERSEUS integrates heterogeneous multimodal ingestion (text, images, audio) with a two-stage verification gate that combines hybrid sparse-and-dense retrieval with NLI-based entailment checking, applied to every candidate triple. On the CaRB open information extraction benchmark, this yields an 11% precision gain, a 54.6% false-positive reduction, and a statistically significant overall improvement. By enforcing provenance-aware validation at extraction time, PERSEUS produces higher-fidelity multimodal knowledge graphs with less human review, offering a scalable pathway toward dependable automated KG construction and ontology learning across high-stakes domains.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4o and Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, ACM Computing Surveys (Csur) 54 (2021) 1–37.

[2] Y. Zhang, A. S. Dalal, C. Martin, S. R. Gadusu, H. K. McGinty, Olive: Ontology learning with integrated vector embeddings, Applied Ontology 20 (2025) 36–53.

[3] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM computing surveys 55 (2023) 1–38.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[5] S. Bhardwaj, S. Aggarwal, Mausam, CaRB: A crowdsourced benchmark for open IE, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6262–6267. doi:10.18653/v1/D19-1651.

[6] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[7] S. Robertson, H. Zaragoza, et al., The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.

[8] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering., in: EMNLP (1), 2020, pp. 6769–6781.

[9] P. Mandikal, R. Mooney, Sparse meets dense: A hybrid approach to enhance scientific document retrieval, arXiv preprint arXiv:2401.04055 (2024).

[10] G. V. Cormack, C. L. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 758–759.

[11] W. Yin, J. Hay, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, arXiv preprint arXiv:1909.00161 (2019).

[12] H. Babaei Giglou, J. D'Souza, S. Auer, Llms4ol: Large language models for ontology learning, in: International Semantic Web Conference, Springer, 2023, pp. 408–427.

[13] V. K. Kommineni, B. König-Ries, S. Samuel, From human experts to machines: An llm supported approach to ontology and knowledge graph construction, arXiv preprint arXiv:2403.08345 (2024).

[14] J. Liu, S. Meng, Y. Gao, S. Mao, P. Cai, G. Yan, Y. Chen, Z. Bian, D. Wang, B. Shi, Aligning vision to language: Annotation-free multimodal knowledge graph construction for enhanced llms reasoning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 981–992.

[15] A. M. Bran, Z. Jončev, P. Schwaller, Knowledge graph extraction from total synthesis documents, in: Proceedings of the 1st Workshop on Language+ Molecules (L+ M 2024), 2024, pp. 74–84.

[16] X. Yuan, L. Ning, W. Fan, Q. Li, mkg-rag: Multimodal knowledge graph-enhanced rag for visual question answering, arXiv preprint arXiv:2508.05318 (2025).

[17] J. Lee, Y. Wang, J. Li, M. Zhang, Multimodal reasoning with multimodal knowledge graph, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 10767–10782. URL: https://aclanthology.org/2024.acl-long.579/. doi:10.18653/v1/2024.acl-long.579.

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition

via large-scale weak supervision, in: International conference on machine learning, PMLR, 2023, pp. 28492–28518.
[19] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International conference on machine learning, PMLR, 2023, pp. 19730–19742.
[20] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).
[21] O. Sainz, O. L. de Lacalle, G. Labaka, A. Barrena, E. Agirre, Label verbalization and entailment for effective zero-and few-shot relation extraction, arXiv preprint arXiv:2109.03659 (2021).
[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
[23] B. J. Frey, D. Dueck, Clustering by passing messages between data points, science 315 (2007) 972–976.
[24] U. Von Luxburg, A tutorial on spectral clustering, Statistics and computing 17 (2007) 395–416.
[25] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20 (1987) 53–65.