The Second Bridge on Artificial Intelligence for Scholarly Communication (AAAI-26)

Contributions

https://doi.org/10.52825/ocp.v8i.3173

Published: 18 Dec. 2025

Echo-LLM

Evidence-Checked Hierarchical Ontology

Aryan Singh Dalal[1] and Hande McGinty[2,*]

[1,2]Department of Computer Science, Kansas State University

*Correspondence: Hande McGinty, hande@ksu.edu

Abstract. Large language models can draft ontologies, but unverified extraction yields hallucinated triples—producing plausible yet incorrect facts. EchoLLM is a text-only, evidence-grounded pipeline for ontology construction. Candidate triples are first extracted with an instruction-following LLM. A hybrid retriever (BM25 + dense) gathers sentence-level evidence for each triple. Natural language inference then tests whether the evidence entails the triple; only entailed, lexically consistent hypotheses are accepted, and all decisions are logged. Accepted entities are embedded and clustered to induce classes and a lightweight hierarchy; rdfs:comment is generated from supporting text. The result is a validated triple set and an initial ontology suitable for bootstrapping domain knowledge graphs. The construction design favors high precision which requires no domain-specific rules, and surfaces failure modes (extraction, retrieval, verification). This enables authors and subject-matter experts to build trustworthy knowledge graphs quickly while keeping model and cost choices flexible.

Keywords: Knowledge Graphs, Ontology Induction, Large Language Models, Retrieval Augmented Generation, Natural Language Inference

1. Introduction and Related Work

Large language models (LLMs) have transformed knowledge graph (KG) construction by automating the extraction of structured triples from text [1], [2], [3], [4], [5], [6], [7]. Yet their propensity for factual hallucination [8]—producing fluent but false assertions—renders them unreliable for scientific and enterprise applications [9]. Addressing hallucination requires explicit validation beyond linguistic coherence.

Retrieval-Augmented Generation (RAG) mitigates hallucination by grounding generation in evidence corpora [10]. Yet, even with relevant documents, RAG can misinterpret or overlook context [11]. Sparse retrievers such as BM25 [12], [13] excel at lexical precision but falter on semantic variation, while dense retrievers based on Sentence Transformers [14] capture meaning but often miss domain-specific terms. Combining both improves recall and precision [15], [16]. Reciprocal Rank Fusion (RRF) [17] merges ranked lists without supervision, outperforming complex learning-to-rank methods in low-data scenarios [18].

While hybrid retrieval strengthens grounding, verification remains incomplete. Logical validation via Natural Language Inference (NLI) provides a principled mechanism to test whether evidence entails a candidate triple [19]. In this formulation, retrieved sentences act as premises, and triples verbalized as hypotheses are accepted only if entailed. NLI-based verification has proven effective in fact-checking [20], though domain shift [21], [22] can degrade accuracy in specialized contexts. Despite progress, prior frameworks [23], [24], [25] treat retrieval and verification as isolated modules.

EchoLLM integrates these components into a unified pipeline for ontology generation. It extracts triples via an LLM, retrieves hybrid evidence, and applies NLI-based validation before embedding and clustering entities into a hierarchical ontology. This coupling of retrieval, inference, and clustering creates a transparent and auditable system emphasizing factual precision and structural consistency.

2. Methodology

2.1 Model Selection and Evaluation

Four candidate LLMs—DeepSeek-7B, Mistral-7B, Llama3-8B, and ChatGPT-4o-mini—were evaluated under identical hyperparameters (temperature 0.5, max tokens 500) and a fixed prompt. Each was tasked with converting 35 expert-annotated research abstracts ($\approx$65-70 words each) from the domains spanning diverse disciplines including emerging technologies , sustainability (Renewable Energy, Climate Change), and socio-cultural studies and tech into subject–predicate–object triples. Relaxed matching rules allowed minor predicate or object variations to count as correct.Temperature was set to 0.5 to balance structural adherence with extraction variety. Llama3-8B was selected as it achieved the highest Precision (0.47) and Recall (0.44) compared to Mistral-7B and ChatGPT-4o-mini, whereas DeepSeek-7B yielded incomplete triples. This balance of instruction following and extraction accuracy justified its selection for the pipeline.

Llama3-8B achieved the best balance of precision, structure, and interpretability, and was selected for all subsequent stages.

2.2 Preprocessing

Input text undergoes normalization (removal of boilerplate and whitespace standardization via regular expressions) followed by segmentation using the `spaCy` dependency parser (`en_core_web_sm`). This parser-based segmentation automates boundary detection, splitting abstracts into atomic sentences to isolate local context. The resulting set of sentences is then passed to *Llama3-8B*, which is prompted to extract triples from each sentence independently, thereby reducing cross-sentence hallucination and ensuring tighter contextual grounding.

2.3 Triple Extraction

Processed text is fed to *Llama3-8B* with the instruction. The model's structured output (e.g., $[\,\text{Urban agriculture}, \text{benefits}, \text{biodiversity}\,]$) functions as an implicit entity–boundary detector, treating multi-word expressions as atomic units. This generative formulation removes the need for separate BIO-tagging or span-prediction modules and enables the extraction of complex, discontinuous, or nested entities that conventional NER pipelines frequently fail to capture.

```
Convert each numbered sentence into [Subject, Predicate, Object]
triples. Return only triples under a header 'Triples:'
```

Example output for the sample above:

```
Triples:
[Urban agriculture, provides, food]
[Urban agriculture, benefits, biodiversity]
```

2.4 Hybrid Retrieval and Verification

Each triple is verified against its source text using a two-stage hybrid search and logical validation.

Hybrid retrieval:

BM25 (lexical) and all-MiniLM-L6-v2 (semantic) results are fused by Reciprocal Rank Fusion (k = 60). For example, the query "Urban agriculture benefits biodiversity" retrieves top 3 sentences combining both keyword and semantic matches.

Entailment Verification:

To minimize hallucinations, candidates undergo a two-step validation. First, lexical verification enforces strict grounding: subjects must appear as substrings, while object lemmas must form a subset of the sentence lemmas (e.g., matching 'leaves' to 'leaf'). Matches are assigned a high confidence ($\geq$ 0.95). Second, NLI verification (BART-Large-MNLI) validates the logical relationship; the entailment threshold ($\geq$ 0.7) was empirically tuned on a held-out set to maximize F1. Context expansion resolves coreference by prepending the preceding sentence when the subject is absent.

Thresholds were tuned on a held-out development set (10%): a lexical confidence of ≥ 0.95 enforces strict entity grounding, while an NLI threshold of ≥ 0.7 maximizes F_1 by filtering neutral or weakly supported inferences. For example, the sentence *"Urban farming increases diversity"* entails the triple *"Urban agriculture benefits biodiversity"* with an NLI score of 0.82, satisfying the threshold and validating the relation.

2.5 Entity Clustering and Ontology Construction

Validated subjects and objects are embedded using `bert-base-uncased` and ℓ_2-normalized. Spectral Clustering (SC) was selected over Affinity Propagation because it yielded tighter, semantically coherent groupings (Silhouette 0.56 vs. 0.41). To induce the ontology structure, cluster centroids are formalized as superclasses (`owl:Class`), while the constituent subjects and objects are assigned as subclasses (`rdfs:subClassOf`). This constructs a subsumption hierarchy (e.g., *anti-inflammatory* $\sqsubseteq$ *Bioactive Properties*), treating extracted terms as atomic domain concepts rather than flat instances.

SC produced tighter, more interpretable clusters. Cluster centroids become `owl:Class`; members become subclasses. Example:

```
Class: AntioxidantProperties
SubclassOf: [anti-carcinogenic properties, anti-inflammatory properties]
```

Each entity receives an `rdfs:comment` generated from source sentences, reviewed interactively by the user, and optional phonetic labels for accessibility.

2.6 Runtime and Reproducibility

Processing 35 abstracts (≈2,500 words) required ∼11 minutes on a MacBook M3 Pro using CPU for NLI. Retrieval and verification consumed 80% of runtime. The pipeline logs every step with timestamps and confidence scores for full auditability.

3. Results and Conclusion

Comprehensive evaluation of EchoLLM on the CaRB (Comprehensive Assessment of Relation Extraction Benchmark) [26], [27], [28] dataset demonstrates its effectiveness in reducing hallucinations while maintaining semantic integrity in ontology generation. EchoLLM's integration of retrieval-based verification and Natural Language Inference (NLI) filtering leads to statistically validated precision improvements and substantial error reduction compared with a direct LLM-only baseline.

Quantitatively, The pipeline achieved a 55% reduction in False Positives (dropping from 1,229 to 558). This drastic cut in noise drove the Precision score up from 0.64 to 0.76 (rejection breakdown: 65% lexical, 30% NLI, 5% retrieval). Although recall dropped from 0.74 to 0.57, this represents a controlled tradeoff favoring correctness. The F1-score showed less volatility (0.69 $\rightarrow$ 0.65), showing balance between precision gains and recall reduction. The induced taxonomy achieved a Silhouette score of 0.56 (compared to 0.41 for Affinity Propagation), reflecting distinct and well-separated concept clusters. Importantly, a McNemar's test validated the significance of this precision improvement ($\chi^2(1) = 319.07$, $p < 0.001$), confirming that EchoLLM's advantage is not due to random variance but reflects a genuine methodological benefit.

Qualitatively, EchoLLM-generated ontologies exhibited tighter class hierarchies, coherent relationships, and accurate contextual annotations. Most of auto-generated `rdfs:comment` and `rdfs:label` entries were accepted by domain experts, verifying their contextual correctness. These structural and semantic gains, combined with empirically validated precision improvements, demonstrate EchoLLM's capability to produce verifiable, low-hallucination knowledge graphs at scale. EchoLLM achieves a statistically validated reduction in hallucinated triples, producing more factually reliable and semantically coherent ontologies. Its retrieval–NLI verification pipeline prioritizes correctness over completeness, ensuring that only well-supported relations are retained—an essential property in domains where incorrect assertions can distort downstream insights. A paired t-test on per-document precision scores ($t = 14.82$, $p < 0.001$) confirms that these improvements are significant across the corpus, establishing EchoLLM as a robust framework for verifiable, automated ontology construction.

Additional documentation, and detailed results can be found at our GitHub.

Author contributions

Aryan Singh Dalal: Conceptualization, Methodology, Software, Investigation, Writing.
Hande McGinty: Conceptualization, Supervision, Review and Editing.

Competing interests

The authors declare no competing interests.

Funding

No external funding was acquired for this work.

References

[1] T. Brown et al., "Language models are few-shot learners", *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] S. Sadruddin et al., "Llms4schemadiscovery: A human-in-the-loop workflow for scientific schema mining with large language models", in *European Semantic Web Conference*, Springer, 2025, pp. 244–261.

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

[4] T. Aggarwal, A. Salatino, F. Osborne, and E. Motta, "Large language models for scholarly ontology generation: An extensive analysis in the engineering field", *Information Processing & Management*, vol. 63, no. 1, p. 104 262, 2026.

[5] A. S. Dalal, Y. Zhang, D. Doğan, A. M. İleri, and H. K. McGinty, "Flavonoid fusion: Creating a knowledge graph to unveil the interplay between food and health", *arXiv preprint arXiv:2510.06433*, 2025.

[6] A. S. Dalal, S. Abadifard, and H. K. McGinty, "Gliide: Global-local image integration via descriptive extraction", in *Proceedings of the 13th Knowledge Capture Conference 2025*, 2025, pp. 194–197.

[7] Y. Zhang, A. S. Dalal, C. Martin, S. R. Gadusu, and H. K. McGinty, "Olive: Ontology learning with integrated vector embeddings", *Applied Ontology*, vol. 20, no. 1, pp. 36–53, 2025.

[8] Z. Ji et al., "Survey of hallucination in natural language generation", *ACM computing surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[9] R. Bommasani et al., "On the opportunities and risks of foundation models", *arXiv preprint arXiv:2108.07258*, 2021.

[10] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks", *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[11] T. Bruckhaus, "Rag does not work for enterprises", *arXiv preprint arXiv:2406.04369*, 2024.

[12] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al., "Okapi at trec-3", *Nist Special Publication Sp*, vol. 109, p. 109, 1995.

[13] B. J. Chan, C.-T. Chen, J.-H. Cheng, and H.-H. Huang, "Don't do rag: When cache-augmented generation is all you need for knowledge tasks", in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 893–897.

[14] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks", *arXiv preprint arXiv:1908.10084*, 2019.

[15] P. Mandikal and R. Mooney, "Sparse meets dense: A hybrid approach to enhance scientific document retrieval", *arXiv preprint arXiv:2401.04055*, 2024.

[16] D. Lee, S.-w. Hwang, K. Lee, S. Choi, and S. Park, "On complementarity objectives for hybrid retrieval", in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 13 357–13 368.

[17] G. V. Cormack, C. L. Clarke, and 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.

[18] T. Chen et al., "Dense x retrieval: What retrieval granularity should we use?", in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 15 159–15 177.

[19] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference", *arXiv preprint arXiv:1508.05326*, 2015.

[20] M. Pàmies et al., "A weakly supervised textual entailment approach to zero-shot text classification", in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 286–296.

[21] D. Tam, A. Mascarenhas, S. Zhang, S. Kwan, M. Bansal, and C. Raffel, "Evaluating the factual consistency of large language models through news summarization", *arXiv preprint arXiv:2211.08412*, 2022.

[22] D. Hendrycks et al., "Measuring massive multitask language understanding", *arXiv preprint arXiv:2009.03300*, 2020.

[23] M. J. Saeedizade and E. Blomqvist, "Navigating ontology development with large language models", in *European Semantic Web Conference*, Springer, 2024, pp. 143–161.

[24] N. Mihindukulasooriya, S. Tiwari, C. F. Enguix, and K. Lata, "Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text", in *International semantic web conference*, Springer, 2023, pp. 247–265.

[25] H. Yang, L. Xiao, R. Zhu, Z. Liu, and J. Chen, "An llm supported approach to ontology and knowledge graph construction", in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2024, pp. 5240–5246.

[26] G. Stanovsky and I. Dagan, "Creating a large benchmark for open information extraction", in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2300–2305.

[27] R. Schneider, T. Oberhauser, T. Klatt, F. A. Gers, and A. Löser, "Analysing errors of open information extraction systems", *arXiv preprint arXiv:1707.07499*, 2017.

[28] W. Lechelle, F. Gotti, and P. Langlais, "Wire57: A fine-grained benchmark for open information extraction", in *Proceedings of the 13th Linguistic Annotation Workshop*, 2019, pp. 6–15.