Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

What Do Current Evaluation Practices Evaluate About LLMs in Ontology-Oriented Generation? A Focused Review for Specialized-Domain Settings

Dhaker Iyad Nejah[1,*]
dhaker.nejah@etu.univ-nantes.fr

Mounira Harzallah[1,*]
mounira.harzallah@univ-nantes.fr

Giuseppe Berio[2,*]
Giuseppe.Berio@univ-ubs.fr

Yasamin Eslami[3]
yasamin.eslami@ec-nantes.fr

[1] Polytech Nantes, Nantes Université, LS2N, Nantes, France
[2] Université Bretagne Sud, France
[3] École Centrale de Nantes, LS2N, Nantes, France
[*] Equal contribution

RÉSUMÉ

Cette contribution réexamine l'évaluation des systèmes à base de grands modèles de langue (LLM) pour la construction, la population et l'enrichissement d'ontologies dans des domaines spécialisés. À partir d'un corpus PRISMA(Page *et al.*, 2021) de 20 études, elle propose un déplacement de perspective : au lieu de partir principalement de la qualité de l'artefact ontologique produit, elle demande ce que les protocoles d'évaluation caractérisent effectivement à propos du LLM. Trois vues analytiques sont distinguées : l'évaluation au niveau de la tâche, où le LLM est évalué comme prédicteur local ; l'évaluation au niveau de l'artefact, où le LLM agit dans un cadre contraint par un schéma ou un métamodèle ; et l'évaluation au niveau du pipeline, où le LLM intervient comme composant d'un système plus large, souvent avec RAG. L'analyse montre que les protocoles caractérisent surtout soit des sorties locales, soit la qualité de l'artefact final, soit le comportement de bout en bout du pipeline, tandis que la contribution propre du LLM reste plus rarement spécifiée avec précision. Trois limites ressortent alors : une faible caractérisation des capacités de structuration du modèle, une évaluation du grounding souvent médiée par la qualité des sorties dans les systèmes avec RAG, et une reproductibilité encore inégale.

ABSTRACT

What Do Current Evaluation Practices Evaluate About LLMs in Ontology-Oriented Generation? A Focused Review for Specialized-Domain Settings

Large language models (LLMs) are increasingly used for ontology construction, enrichment or population. In this setting, evaluation is not only a matter of whether an ontology-oriented output appears correct. It also concerns what the evaluation protocol actually reveals about the LLM. Based on a PRISMA-style review of 20 studies(Page *et al.*, 2021), this paper reframes the literature from that perspective. It distinguishes three analytical views: task-level evaluation, where the LLM is assessed as a local predictor; artifact-level evaluation, where the LLM generates within a schema- or metamodel-constrained setup; and pipeline-level evaluation, where the LLM is one component inside a broader multi-stage or retrieval-augmented system. These views are not proposed as a

Actes de l'atelier sur l'évaluation des modèles génératifs (LLM) et challenges (EvalLLM 2026), pages 207–215.

rigid typology of systems, but as different ways of identifying what evaluation primarily targets. Across the corpus, current protocols mainly characterize local prediction quality, assembled artifact quality, or end-to-end pipeline behavior, while the specific contribution of the LLM is described less directly. The review highlights three recurring gaps: limited evaluation of cross-output structural behavior, indirect evaluation of grounding in retrieval-based systems, and uneven reproducibility due to incomplete reporting of prompts and interaction protocols. This reframing helps make evaluation claims more explicit and more comparable across ontology-oriented LLM studies.

MOTS-CLÉS : grands modèles de langue, génération augmentée par recherche, ontologies, évaluation, domaines spécialisés.

KEYWORDS: large language models, retrieval-augmented generation, ontologies, evaluation, specialized domains.

1 Introduction

Recent work shows that LLMs can effectively support ontology learning, enrichment, and population from unstructured and semi-structured data sources. However, in ontology engineering, output quality goes beyond the correctness of isolated predictions. Generated ontologies must be logically consistent, structurally coherent (so that the organization of concepts and relations remains meaningful and compliant with modeling constraints (Garijo *et al.*, 2024). These challenges are further amplified in retrieval-augmented generation (RAG), where outputs depend not only on the model's internal knowledge but also on external information retrieved at inference time (Lewis *et al.*, 2020). As a result, multiple aspects must be considered jointly, including generation quality, structural validity, grounding, and practical usefulness. This is particularly important in specialized domains, where strong performance on individual subtasks does not guarantee reliable ontology construction or population.

This paper revisits a PRISMA-style review of 20 studies to analyze how LLM-based ontology generation approaches are evaluated across LLM-only, schema-guided, and retrieval-augmented settings, and to clarify what current evaluation protocols actually reveal about the capabilities of LLMs.

2 From ontology-oriented evaluation to what is being evaluated about the LLM

Across the reviewed studies, evaluation protocols vary with task scope, modeling guidance, and the use of retrieval. This variation highlights a deeper issue: evaluation does not consistently target the LLM itself.

In some cases, evaluation focuses on the correct prediction of ontology-related elements (e.g., types or relations). In others, it assesses whether a generated ontology satisfies formal or modeling requirements, or whether a full pipeline performs effectively end-to-end.

To capture these differences, we introduce three analytical views: task-level, artifact-level, and pipeline-level evaluation. These views do not define strict categories of LLM-based approaches,

but rather reflect different ways evaluation is operationalized across the literature, independently of specific design choices such as schema guidance or retrieval.

2.1 Task-level evaluation: the LLM as predictor

At the task level, the LLM is primarily evaluated as a predictor of ontology components, such as term type assignment, taxonomy extraction, or relation prediction. These tasks are assessed against a ground truth using standard metrics such as precision, recall, or F1-score (Giglou *et al.*, 2023; Lo *et al.*, 2024). Such measures capture the ability to generate individual elements in isolation. In this setting, the LLM is not required to construct a full ontology and does not adhere to any explicit global structure.

However, this evaluation does not assess the LLM's ability to perform structured reasoning over generated knowledge. In particular, it does not test whether predictions remain globally coherent when combined. For example, a model may correctly generate Dog is-a Mammal and Mammal is-a Animal, while also producing inconsistent relations such as Animal is-a Dog.

Because predictions are evaluated independently, such inconsistencies are not necessarily penalized. As a result, while task-level evaluation provides reliable insights into local performance, it does not capture whether the LLM can support ontology construction as a coherent process.

The approach of (Kabal *et al.*, 2024) points in the same direction. Evaluation mainly focuses on generated triplets, using automatic measures such as precision, recall, and F1, alongside manual matching by annotators. This goes beyond a purely automatic benchmark, but it still evaluates the output mainly at the level of relations and triples rather than as a complete ontology artifact.

2.2 Artifact-level evaluation: the LLM as generator inside a constrained setup

At the artifact level, the LLM is evaluated under explicit structural guidance, typically provided by a schema or metamodel. Outputs vary, including ontology population under predefined classes, ontology modules derived from requirements or competency questions, and initial ontology drafts in formats such as OWL or Turtle (Ciatto *et al.*, 2025; Lippolis *et al.*, 2025, 2024; Fathallah *et al.*, 2025, 2024). Some approaches extend this to ontology-to-knowledge-graph construction workflows (Tupayachi *et al.*, 2024).

In these settings, task-level metrics capture only part of the result. The generated artifact is further evaluated through syntax validation, logical consistency checking, structural analysis, competency-question testing, or expert review (Cappelli & Di Marzo Serugendo, 2025; Lippolis *et al.*, 2024; Fathallah *et al.*, 2025; Ciatto *et al.*, 2025).

Evaluation therefore shifts from assessing the LLM as a predictor of isolated outputs to assessing the quality of a structured artifact produced within a constrained setup. This setup includes not only the LLM, but also the schema that guides generation and the tools used to validate or filter the results.

As a result, the evaluation primarily reflects the quality of the artifact within this environment, rather than the standalone capabilities of the LLM. In some cases, additional validation is performed through downstream applications such as question answering or decision support (Tupayachi *et al.*, 2024; Oarga *et al.*, 2026), further emphasizing system-level performance.

2.3 Pipeline-level evaluation: the LLM as one component in a broader system

At the pipeline level, the LLM is evaluated as one component within a larger pipeline, often involving retrieval. Retrieval-augmented approaches are diverse. Some are still assessed on specific ontology-learning tasks using reference data, with standard metrics such as precision, recall, F1, or ranking-based scores. For example, Phoenixes was evaluated in the LLMs4OL 2024 challenge on term typing, taxonomy discovery, and relation extraction (Sanaei *et al.*, 2024). Other approaches integrate retrieval into broader workflows for ontology completion, population, enrichment, or multi-stage construction (Toro *et al.*, 2024; Norouzi *et al.*, 2024; Kollapally *et al.*, 2025; Nayyeri *et al.*, 2025; Lobo-Santos & Borrego-Díaz, 2025; Xie *et al.*, 2026). In these cases, evaluation becomes more composite, combining local metrics with consistency checking, structural analysis, expert review, and sometimes downstream validation.

From this perspective, evaluation primarily characterizes end-to-end pipeline performance. While it can show improvements in generated outputs or final artifacts, attribution becomes more difficult, as results depend on multiple interacting components, including retrieval, prompting strategies, schema guidance, and validation tools.

This is particularly evident for retrieval: most studies evaluate its impact indirectly through output quality or downstream performance, rather than through explicit IR metrics or source attribution. As a result, these protocols provide a clearer evaluation of overall system behavior than of how the LLM itself uses retrieved information.

3 Evaluation protocols and operational implications

Table 1 summarizes the main evaluation protocols observed in the reviewed corpus. The table does not imply that each study uses only one protocol; rather, it makes explicit which object is primarily evaluated and what each protocol can and cannot reveal about the LLM.

Table 1: Main evaluation protocols observed in the reviewed corpus.

Protocol type	**Typical target**	**Common measures**	**Main limitation**
Benchmark-style ontology-learning tasks	Term typing, taxonomy discovery, relation extraction	Precision, recall, F1, ranking scores, exact or partial match against a reference	Captures local prediction quality, but not full ontology coherence.
Reference-ontology comparison	Generated taxonomy, ontology fragment, or KG structure	Graph similarity, structural overlap, comparison with gold ontology or curated annotations	Sensitive to the reference model; alternative valid modeling choices may be penalized.
Formal and structural validation	OWL/Turtle/JSON artifact, schema-constrained output	Syntax validity, reasoner consistency, cycle detection, ontology pitfall detection, metamodel compliance	Shows whether the artifact is formally acceptable, but not necessarily whether the LLM itself reasoned correctly.
Expert or qualitative review	Conceptual adequacy, domain relevance, modeling choices	Manual inspection, expert scoring, qualitative comparison	Important for ontology quality, but less reproducible if criteria are not explicit.
Pipeline or RAG-based evaluation	Multi-stage system output, retrieved evidence, populated ontology/KG	Output quality, downstream QA, ablation, grounding checks, sometimes retrieval measures	Often evaluates the whole pipeline more directly than the isolated LLM contribution.

3.1 Operational implications for evaluation design

Although this paper does not introduce a new benchmark or experimental protocol, the proposed distinction between task-level, artifact-level, and pipeline-level evaluation can be translated into a practical evaluation grid. For each evaluated system, an evaluation protocol should specify: (i) the unit being evaluated, such as a local prediction, an ontology fragment, or a complete pipeline output; (ii) the role assigned to the LLM, such as classifier, extractor, generator, repair component, or validation aid; (iii) the type of control used, such as prompting constraints, schema guidance, metamodel constraints, retrieval, or post-generation validation; and (iv) the evaluation measures used at each level.

At the task level, suitable measures include precision, recall, F1, ranking scores, and exact or partial match against reference annotations. At the artifact level, these measures should be complemented by syntactic validity, logical consistency, ontology pitfall detection, duplicate or cyclic structure detection, and compliance with the intended schema or metamodel. At the pipeline level, especially in RAG-based systems, the evaluation should separate the quality of retrieval, the use of retrieved evidence by the LLM, and the quality of the final generated ontology or knowledge graph. Without this separation, improvements may be attributed to the LLM even when they result from retrieval, schema constraints, filtering, or expert correction.

The grid can also be read as a minimal evaluation design for future ontology-oriented LLM studies: comparing LLM-only, schema-guided, RAG-only, and schema-guided RAG configurations on the same input documents and target ontology requirements. Such an ablation design would make it easier to determine whether observed gains come from the model itself, from structural guidance, from retrieved evidence, or from their combination.

4 Main evaluation gaps

Across the reviewed studies, evaluation protocols focus on different targets, including local outputs, structured artifacts, and full pipelines. From an LLM-centered perspective, these variations reveal three main gaps.

Gap 1 – Lack of structural reasoning evaluation. Current evaluation protocols mainly assess the correctness of individual outputs or the acceptability of final artifacts. However, they do not evaluate how LLMs construct and maintain structured knowledge. In particular, they do not test whether the model produces globally coherent structures, respects modeling constraints, or makes consistent decisions across multiple generation steps.

Gap 2 – Weak grounding evaluation in retrieval-based settings. Retrieval-augmented approaches are often motivated by grounding, but retrieval quality is rarely evaluated directly. Instead, grounding is inferred from improvements in generated outputs or overall system performance. As a result, it remains unclear whether these improvements come from better retrieval, better use of retrieved information by the LLM, or other components of the pipeline.

Gap 3 – Limited reproducibility of LLM behavior. LLM performance depends strongly on prompts, interaction design, and decomposition strategies. However, these elements are often only partially reported and not standardized. This makes it difficult to reproduce results or to compare systems in a consistent way.

Taken together, these gaps highlight a common limitation: current evaluation protocols capture what systems produce, but not how LLMs behave.

5 Ethical, linguistic, and reproducibility considerations

Ontology-oriented LLM systems raise specific ethical and methodological issues. First, generated ontology content may appear formally structured while still containing unsupported, biased, or domain-inaccurate modeling choices. In specialized domains, such errors can affect downstream decision-support systems if generated classes, relations, or instances are reused without validation. Second, retrieval-augmented systems may give an impression of grounding even when the retrieved evidence is incomplete, irrelevant, or only weakly connected to the generated output. For this reason, provenance, source attribution, and expert validation are important safeguards. Third, reproducibility is an ethical issue in itself: when prompts, retrieval settings, model versions, and validation procedures are not reported, results become difficult to verify or compare. Finally, the language scope of the reviewed corpus is limited, since the search and included studies are mainly English-based; conclusions may therefore not transfer directly to multilingual or low-resource settings.

6 Conclusion

This paper revisits a focused review of ontology-oriented LLM studies by asking what current evaluation practices actually measure about the LLM. The answer depends on the level of evaluation. At the task level, the LLM is assessed as a predictor of local outputs. At the artifact level, evaluation focuses on the quality of a structured ontology. At the pipeline level, it reflects the performance of a larger system in which the LLM is only one component.

This perspective clarifies what current protocols capture: local correctness, artifact quality, or end-to-end system performance. However, it remains unclear what these results say about the LLM itself. In particular, key aspects of LLM behavior are only partially evaluated, including structured reasoning, grounding in retrieval-based settings, and reproducibility.

The proposed evaluation grid helps make these targets explicit by linking each level to typical objects of evaluation, possible measures, and limits of interpretation. As a result, studies presented as evaluations of ontology generation with LLMs may in fact evaluate different things, making results difficult to compare and interpret. Making evaluation targets more explicit would help better understand LLM capabilities.

References

CAPPELLI M. A. & DI MARZO SERUGENDO G. (2025). Methodological exploration of ontology

generation with a dedicated large language model. *Electronics*, **14**(14). DOI : 10.3390/electronics14142863.

CIATTO G., AGIOLLO A., MAGNINI M. & OMICINI A. (2025). Large language models as oracles for instantiating ontologies with domain-specific knowledge. *Knowledge-Based Systems*, **310**, 112940. DOI : https://doi.org/10.1016/j.knosys.2024.112940.

FATHALLAH N., DAS A., GIORGIS S. D., POLTRONIERI A., HAASE P. & KOVRIGUINA L. (2025). Neon-gpt: A large language model-powered pipeline for ontology learning. In *The Semantic Web: ESWC 2024 Satellite Events*, p. 36–50, Cham: Springer Nature Switzerland.

FATHALLAH N., STAAB S. & ALGERGAWY A. (2024). Llms4life: Large language models for ontology learning in life sciences.

GARIJO D., POVEDA-VILLALÓN M., AMADOR-DOMÍNGUEZ E., WANG Z., GARCÍA-CASTRO R. & CORCHO O. (2024). Llms for ontology engineering: A landscape of tasks and benchmarking challenges. In *Proceedings of the Special Session on Harmonising Generative AI and Semantic Web Technologies (HGAIS 2024), co-located with the 23rd International Semantic Web Conference (ISWC 2024)*.

GIGLOU H. B., D'SOUZA J. & AUER S. (2023). Llms4ol: Large language models for ontology learning. In *The Semantic Web – ISWC 2023*, p. 408–427. DOI : 10.1007/978-3-031-47240-5_22.

KABAL O., HARZALLAH M., GUILLET F. & ICHISE R. (2024). Enhancing domain-independent knowledge graph construction through OpenIE cleaning and LLMs validation. In *Proceedings of the 28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)*, p. 2617–2626. DOI : 10.1016/j.procs.2024.09.436.

KOLLAPALLY N. M., GELLER J., KELOTH V. K., HE Z. & XU J. (2025). Ontology enrichment using a large language model: Applying lexical, semantic, and knowledge network-based similarity for concept placement. *Journal of Biomedical Informatics*, **168**, 104865. DOI : https://doi.org/10.1016/j.jbi.2025.104865.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, p. 9459–9474.

LIPPOLIS A. S., CERIANI M., ZUPPIROLI S. & NUZZOLESE A. G. (2024). Ontogenia: Ontology generation with metacognitive prompting in large language models. In *ESWC 2024 Satellite Events*, p. 259–265.

LIPPOLIS A. S., SAEEDIZADE M. J., KESKISÄRKKÄ R., ZUPPIROLI S., CERIANI M., GANGEMI A., BLOMQVIST E. & NUZZOLESE A. G. (2025). Ontology generation using large language models.

LO A., JIANG A. Q., LI W. & JAMNIK M. (2024). End-to-end ontology learning with large language models.

LOBO-SANTOS A. & BORREGO-DÍAZ J. (2025). Enhancing mathematical knowledge graphs with large language models. *Modelling*, **6**(3). DOI : 10.3390/modelling6030053.

NAYYERI M., YOGI A. A., FATHALLAH N., THAPA R. B., TAUTENHAHN H.-M., SCHNURPEL A. & STAAB S. (2025). Retrieval-augmented generation of ontologies from relational databases. DOI : 10.48550/arXiv.2506.01232.

NOROUZI S. S., BARUA A., CHRISTOU A., GAUTAM N., EELLS A., HITZLER P. & SHIMIZU C. (2024). Ontology population using large language models. DOI : 10.48550/arXiv.2411.01612.

OARGA A., HART M., BRAN A. M., LEDERBAUER M. & SCHWALLER P. (2026). Scientific knowledge graph and ontology generation using open large language models. *Digital Discovery*, **5**(3), 1269–1279. DOI : 10.1039/d5dd00275c.

PAGE M. J., MCKENZIE J. E., BOSSUYT P. M., BOUTRON I., HOFFMANN T. C., MULROW C. D., SHAMSEER L., TETZLAFF J. M., AKL E. A., BRENNAN S. E., CHOU R., GLANVILLE J., GRIMSHAW J. M., HRÓBJARTSSON A., LALU M. M., LI T., LODER E. W., MAYO-WILSON E., MCDONALD S., MCGUINNESS L. A., STEWART L. A., THOMAS J., TRICCO A. C., WELCH V. A., WHITING P. & MOHER D. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, **372**, n71. DOI : 10.1136/bmj.n71.

SANAEI M., AZIZI F. & GIGLOU H. B. (2024). Phoenixes at LLMs4OL 2024 tasks a, b, and c: Retrieval-augmented generation for ontology learning. In *Proceedings of the 1st Large Language Models for Ontology Learning Workshop at ISWC 2024*, p. 39–47.

TORO S., ANAGNOSTOPOULOS A. V., BELLO S. M., BLUMBERG K., CAMERON R., CARMODY L., DIEHL A. D., DOOLEY D., DUNCAN W. D., FEY P., GAUDET P., HARRIS N. L., JOACHIMIAK M. P., KIANI L., LUBIANA T., MUÑOZ-TORRES M. C., O'NEIL S. T., OSUMI-SUTHERLAND D., PUIG A. & MUNGALL C. J. (2024). Dynamic retrieval augmented generation of ontologies using artificial intelligence (DRAGON-AI). *Journal of Biomedical Semantics*, **15**(1), 19. DOI : 10.1186/s13326-024-00320-3.

TUPAYACHI J., XU H., OMITAOMU O. A., CAMUR M. C., SHARMIN A. & LI X. (2024). Towards next-generation urban decision support systems through ai-powered construction of scientific ontology using large language models—a case in optimizing intermodal freight transportation. *Smart Cities*, **7**(5), 2392–2421. DOI : 10.3390/smartcities7050094.

XIE S., YANG T., XIE Y., YING H. & WU Z. (2026). Llm-driven multimodal knowledge graph construction for industrial process with prompt optimization and fuzzy rag. *IEEE Transactions on Fuzzy Systems*, p. 1–14. DOI : 10.1109/TFUZZ.2026.3665172.

A Corpus map of reviewed system configurations

Figure 1 provides a descriptive overview of the dominant configurations represented in the reviewed corpus. It groups papers into broad system configurations: LLM-only, LLM with schema/metamodel guidance, and LLM with schema/metamodel guidance plus retrieval. This grouping is retained here as a corpus overview, but it is not the main analytical structure used in the paper.

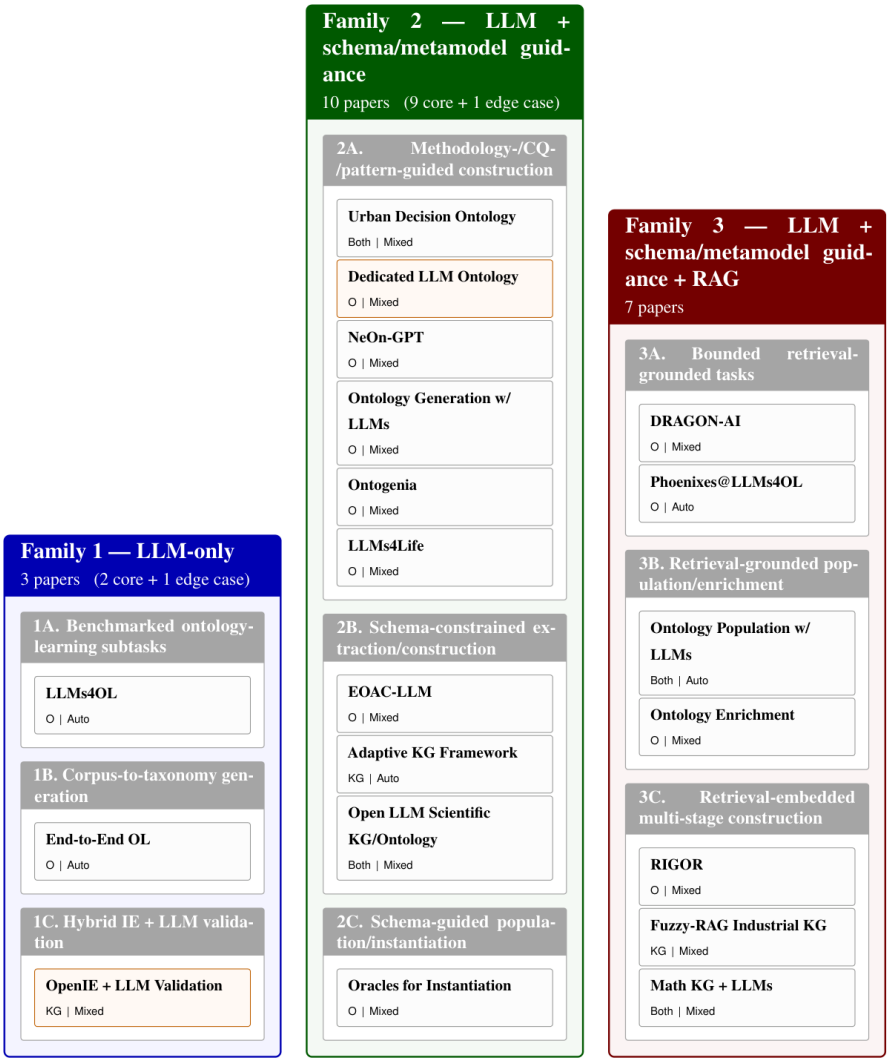

Figure 1: Corpus map of the dominant configurations represented in the reviewed studies. The figure groups papers by broad system configuration (LLM-only, LLM with schema/metamodel guidance, and LLM with schema/metamodel guidance plus retrieval). These families serve as a descriptive overview of the corpus, not as the main analytical categories of this paper. In the main analysis, we instead examine what evaluation primarily targets: local task outputs, generated artifacts, or broader pipeline behavior.