Research Interest & Proposal

Xinyuan Tu

July 2025

Background

Large Language Models (LLMs), including the GPT family [Ach+23] and the Llama family [Tou+23], have demonstrated substantial success across various Natural Language Processing (NLP) tasks, such as text summarization [Rei+24] and question answering [Liu+23; Ras+24]. Despite these advancements, LLMs frequently suffer from hallucinations and inefficiencies in complex logical reasoning. These limitations pose significant risks in systems requiring high reliability and stringent safety standards, as hallucinations and logical errors may lead to severe consequences.

Traditional symbolic AI approaches rely on deterministic algorithms and structured data representations, such as ontologies or Knowledge Graphs (KGs), offering inherent transparency and explainability. However, while ontologies provide clarity, their construction traditionally requires manual curation by domain experts, a process that is labor-intensive and time-consuming. Recent research efforts have aimed to mitigate this limitation by automating ontology construction using LLMs. For example, Babaei et al. [BDA23] have proposed an approach using LLMs for ontology creation, albeit still dependent on pre-defined concepts and relations. Other studies [Fun+23; Oar+24] have attempted to automatically generate concepts and relationships but encountered challenges, including hallucinated conceptual hierarchies and context-window limitations. Moreover, a notable gap remains regarding benchmark datasets for systematically evaluating these LLM-based ontology construction methods.

Once constructed, ontologies typically require continuous updates to incorporate new knowledge (e.g., integrating additional concepts), namely the ontology enrichment. Embedding based models [Che+21; YCS25; JCH24] are not capable for continuous knowledge as embedding models have limited memories. Recent works [Cru+24; MG23; Ouy+24] have introduced LLM-based pipelines for ontology enrichment. Nonetheless, the existing literature is incomplete, as they do not incorporate with contextual information. Moreover, the latest techniques such as LLM agents and GraphRag [Pen+24], which might boost the benchmark, have yet to be thoroughly explored in these publications.

Also, downstream tasks in ontology are interesting. One representative downstream task is reasoning. Several studies [Sun+21; He+23; Pet+19; Luo+23; Ye+22] have utilized language models as knowledge bases, training the models reasoning on the fixed ontology. However, they fail to admit updated knowledge without updating weights. Alternative methodologies, such as retrieval-augmented techniques (e.g., KG-enhanced LLM inference [Pan+24]), integrate ontology-based knowledge directly into the inference context without necessitating retraining. Given these considerations, significant research opportunities remain in KG-enhanced LLM inference field.

Research Questions

Base on the background & motivation, here three research questions are proposed:

- RQ1: Can LLMs automatically construct an ontology from unstructured data like text corpus and semi-structured data (e.g., tables) ?
- RQ2: Can LLMs enrich an ontology given unstructured data and semi-structured data?
- RQ3: Can ontology support LLMs to behave less hallucinated and more explainable for tasks like question answering?

Research Methodology

RQ1: Can LLMs automatically construct an ontology from unstructured data and semi-structured data?

Problem Definition This research problem can be divided into two sub-problems: **(i)** Given unstructured or semi-structured data, can LLM extract TBoxes and the taxonomy relations? **(ii)** Given the data and TBoxes, can LLM extract the ABoxes, term-typing and the non-taxonomy relations?

Dataset One benchmark [Mih+23] prepares two dataset for subproblem (ii) of shape <corpus, ontology>. Moreover, I could also construct this type of dataset. The construction idea is to look for existing ontologies (e.g., Wordnet [Mil+90], GeoNames[1]) and obtain the corpus via tools like KG-to-Text [ZZY23; Wu+23].

[1] http://www.geonames.org/

Metrics Traditional metrics like accuracy, F1 score will be applied. Also, hallucination metrics [Mih+23] will be considered. Moreover, downstream task evaluation via GraphRag [Pen+24] will be conducted.

Research Plan Research plan will first conduct literature review on ontology construction + LLMs. Papers related with subproblem (i) [Fun+23; Oar+24] and one paper [BDA23] will be considered for preliminary study. Finetuning LLMs and combining LLM agents will options.

RQ2: Can LLMs enrich an ontology given unstructured data and semi-structured data?

Problem Definition Given a corpus consisting of unstructured data or semi-structured, can LLMs extra non-existing individuals, concepts from the corpus and insert them into the ontology? The problem can be further divided into several sub-problems: (i) Ontology population: Insert ABoxes and non-taxonomy relations. (ii) Concept discovery: Insert TBOxes and taxnoomy relations.

Dataset The dataset should consist of a corpus describing out-of-KB individuals or concepts, an ontology and its ground truth insertion. Dataset [Don+23; Don+24]. Moreover, the dataset from RQ1 can be adapted for this task.

Metrics For ontology completion tasks, metrics such as Precision@K, Recall@K, and F1@K will be employed. Standard classification metrics, including accuracy and F1-score, will be used for ontology population tasks. Furthermore, expanded ontologies will be evaluated in downstream tasks (e.g., question answering), applying the corresponding downstream task metrics.

Research Plan Initial research will include a literature review focused on ontology enrichment via LLMs, followed by preliminary studies using direct prompting methods. Approaches from RQ1, such as LLM4OL [BDA23], will be reused and adapted, alongside exploration of fine-tuning LLMs, deploying multi-agent systems, and employing retrieval techniques like GraphRag [Pen+24].

RQ3: Can ontology support LLMs to behave less hallucinated and more explainable for tasks like question answering?

Problem Definition Given an ontology O and a text prompt S, we ask whether there exists evidence E such that $O \models E$ and, when E is joined to S, an LLM generates an

answer with fewer hallucinations than it would produce from S alone.

Dataset Existing Question Answering over Knowledge Graphs (KGQA) datasets such as Jiang et al. [JU22], consisting of tuples <natural language question, KG>, will be utilized.

Metrics Evaluation will depend on the question-answering task type: yes/no questions will be assessed using classification metrics (e.g., F1-score), information retrieval tasks using metrics such as Precision@K, and natural language answers using structural metrics (e.g., ROUGE score [Lin04]) and semantic metrics (e.g., BERTScore [Zha+19]). Moreover, hallucination metrics (e.g factual@k) will be applied.

Research Plan The study will begin with a comprehensive literature review on integrating LLMs, ontologies, and question-answering tasks. Preliminary studies will employ simple prompts and retrieval-augmented generation. Exploration of GraphRAG [Pen+24], combing LLMs with symbolic ontology reasoners, and RAG with KG-to-Text methodologies [ZZY23; Wu+23] as baselines will be conducted. Finetuning the retrieval and generation components of GraphRAG, as well as integrating agent-based LLMs, will also be considered.

Risks & Mitigate

Risk 1 What if in RQ1, the way to construct the corpus from ontology fails.
Mitigation 1 (i) Manually convert around 200 ontology statements to text per person. (ii) Try native representation, e.g., converting triplets (h,r,t) to natural language "h r t".

Risk 2 What if in RQ3, the retrieval performance from GraphRAG is suboptimal.
Mitigation 2 Experiment with alternative GraphRAG configurations, increase the retrieved subgraph size, or switch to simpler retrieval methods with fewer triplets.

Risk 3 What if in RQ1 and RQ2, the ontology generated is not self-consistent (i.e., containing conflicts)?
Mitigation 3 Execute a validation loop before generating ontology.

Potential Output

All of the research questions above will produce open-source code. Self created dataset from RQ1 and its variants adapted for RQ2 and RQ3 will be released. Several related

papers might be published. The methodologies can be immigrated to industrial domain (e.g., healthcare), where high explainability is required, to obtain an ontology.

Research Statement

Background During my Master's studies in Artificial Intelligence, my primary focus was ontology learning. My Master's thesis, completed at Robert Bosch Reutlingen, involved extracting ontologies from industrial literature and meeting transcripts. Additionally, I gained significant research experience, notably at the IDEA Lab, FAU, where I conducted research on generative AI techniques applied to medical imaging. Furthermore, I gained teaching experience by instructing symbolic AI courses (including ontology) for two semesters at FAU.

Motivation This PhD project aligns closely with my academic interests, serving as a natural extension of my Master's thesis. I am particularly motivated by the opportunity to delve deeper into the theoretical aspects of ontology generation, and I aim to undertake a comprehensive and rigorous investigation into this area.

References

[Mil+90] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. "Introduction to WordNet: An on-line lexical database". In: *International journal of lexicography* 3.4 (1990), pp. 235–244.

[Lin04] Chin-Yew Lin. "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 2004, pp. 74–81.

[Pet+19] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. "Language models as knowledge bases?" In: *arXiv preprint arXiv:1909.01066* (2019).

[Zha+19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. "Bertscore: Evaluating text generation with bert". In: *arXiv preprint arXiv:1904.09675* (2019).

[Che+21] Jiaoyan Chen, Pan Hu, Ernesto Jimenez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks. "OWL2Vec*: embedding of OWL ontologies". In: *Machine Learning* 110.7 (2021), pp. 1813–1845.

[Sun+21] Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. "Can language models be biomedical knowledge bases?" In: *arXiv preprint arXiv:2109.07154* (2021).

[JU22] Longquan Jiang and Ricardo Usbeck. "Knowledge Graph Question Answering Datasets and Their Generalizability: Are They Enough for Future Research?" In: *arXiv preprint arXiv:2205.06573* (2022).

[Ye+22] Hongbin Ye, Ningyu Zhang, Shumin Deng, Xiang Chen, Hui Chen, Feiyu Xiong, Xi Chen, and Huajun Chen. "Ontology-enhanced Prompt-tuning for Few-shot Learning". In: *Proceedings of the ACM web conference 2022*. 2022, pp. 778–787.

[Ach+23] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[BDA23] Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. "LLMs4OL: Large Language Models for Ontology Learning". In: *The Semantic Web – ISWC 2023*. Ed. by Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li. Cham: Springer Nature Switzerland, 2023, pp. 408–427. ISBN: 978-3-031-47240-4.

[Don+23] Hang Dong, Jiaoyan Chen, Yuan He, and Ian Horrocks. "Ontology enrichment from texts: A biomedical dataset for concept discovery and placement". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2023, pp. 5316–5320.

[Fun+23] Maurice Funk, Simon Hosemann, Jean Christoph Jung, and Carsten Lutz. "Towards ontology construction with language models". In: *arXiv preprint arXiv:2309.09898* (2023).

[He+23] Yuan He, Jiaoyan Chen, Ernesto Jimenez-Ruiz, Hang Dong, and Ian Horrocks. "Language model analysis for ontology subsumption inference". In: *arXiv preprint arXiv:2302.06761* (2023).

[Liu+23] Yixin Liu, Alexander R Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. "Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization". In: *arXiv preprint arXiv:2311.09184* (2023).

[Luo+23] Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. "Reasoning on graphs: Faithful and interpretable large language model reasoning". In: *arXiv preprint arXiv:2310.01061* (2023).

[MG23] Patricia Mateiu and Adrian Groza. “Ontology engineering with large language models”. In: *2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. IEEE. 2023, pp. 226–229.

[Mih+23] Nandana Mihindukulasooriya, Sanju Tiwari, Carlos F. Enguix, and Kusum Lata. “Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text”. In: *The Semantic Web – ISWC 2023*. Ed. by Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li. Cham: Springer Nature Switzerland, 2023, pp. 247–265. ISBN: 978-3-031-47243-5.

[Tou+23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).

[Wu+23] Yike Wu, Nan Hu, Sheng Bi, Guilin Qi, Jie Ren, Anhuan Xie, and Wei Song. “Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering”. In: *arXiv preprint arXiv:2309.11206* (2023).

[ZZY23] Feng Zhao, Hongzhi Zou, and Cheng Yan. “Structure-aware knowledge graph-to-text generation with planning selection and similarity distinction”. In: *Proceedings of the 2023 Conference on empirical methods in natural language processing*. 2023, pp. 8693–8703.

[Cru+24] Elias Crum, Antonio De Santis, Manon Ovide, Jiaxin Pan, Alessia Pisu, Nicolas Lazzari, and Sebastian Rudolph. “Enriching Ontologies with Disjointness Axioms using Large Language Models”. In: *arXiv preprint arXiv:2410.03235* (2024).

[Don+24] Hang Dong, Jiaoyan Chen, Yuan He, Yongsheng Gao, and Ian Horrocks. “A language model based framework for new concept placement in ontologies”. In: *European Semantic Web Conference*. Springer. 2024, pp. 79–99.

[JCH24] Mathias Jackermeier, Jiaoyan Chen, and Ian Horrocks. “Dual box embeddings for the description logic EL++”. In: *Proceedings of the ACM Web Conference 2024*. 2024, pp. 2250–2258.

[Oar+24] Alexandru Oarga, Matthew Hart, Andres M Bran, Magdalena Lederbauer, and Philippe Schwaller. “Scientific knowledge graph and ontology generation using open large language models”. In: *AI for Accelerated Materials Design-NeurIPS 2024*. 2024.

[Ouy+24] Siru Ouyang, Jiaxin Huang, Pranav Pillai, Yunyi Zhang, Yu Zhang, and Jiawei Han. “Ontology enrichment for effective fine-grained entity typing”. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024, pp. 2318–2327.

[Pan+24] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. “Unifying large language models and knowledge graphs: A roadmap”. In: *IEEE Transactions on Knowledge and Data Engineering* 36.7 (2024), pp. 3580–3599.

[Pen+24] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. “Graph retrieval-augmented generation: A survey”. In: *arXiv preprint arXiv:2408.08921* (2024).

[Ras+24] Zafaryab Rasool, Stefanus Kurniawan, Sherwin Balugo, Scott Barnett, Rajesh Vasa, Courtney Chesser, Benjamin M Hampstead, Sylvie Belleville, Kon Mouzakis, and Alex Bahar-Fuchs. “Evaluating llms on document-based qa: Exact answer selection and numerical extraction using cogtale dataset”. In: *Natural Language Processing Journal* 8 (2024), p. 100083.

[Rei+24] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. “Gpqa: A graduate-level google-proof q&a benchmark”. In: *First Conference on Language Modeling*. 2024.

[YCS25] Hui Yang, Jiaoyan Chen, and Uli Sattler. “TransBox: EL++-closed Ontology Embedding”. In: *Proceedings of the ACM on Web Conference 2025*. 2025, pp. 22–34.