LLM with Local Knowledge Base Q&A

Shengjun NIU
University of Wisconsin–Madison

Jingcheng HOU
The Hong Kong University of Science and Technology

Introduction

In recent years, with the rapid development of artificial intelligence, Large Language Models (LLMs) such as GPT-3, LLaMA, ChatGPT, and GPT-4 have shown excellent performance in many fields. These models not only have the ability of context learning and thought chain reasoning, but also solve a variety of tasks in a zero-shot or few-shot manner, including machine translation, summary generation, sentiment analysis, and intelligent question answering. However, while LLMs excel at understanding instructions, reasoning, and problem-solving, they still face many challenges when dealing with complex and changing real-world scenarios. In particular, real-time knowledge update, professional skills display, autonomous decision-making ability, and cross-field collaboration still need to be improved. The lack of real-time access to the latest information and the use of specialized tools further restrict the effectiveness of LLMs in practical applications.

To overcome these limitations and realize the full potential of LLMs, researchers began to explore ways to combine LLMs with Agent technology. The LLM-based AI Agent system provides a promising direction to solve the above challenges. This combination can not only utilize the powerful language understanding and generation ability of LLMs, but also enhance the real-time interaction, tool use, and task planning ability of the model through the design of the Agent. LLM-based Agents can receive natural language task instructions provided by users and work out detailed plans to solve complex tasks through their own reasoning ability and by calling external resources and tools. This approach shows great potential in tasks that require a combination of skills, such as the creation of art on a specific topic or the development of personalized travel plans

Literature Review

Knowledge Graph Embedding

Table 3. Link prediction results on WN18RR, FB15k-237 and YAGO3-10. Best results are in **bold** and second best results are underlined. [†]: Results are taken from (Nguyen et al., 2018); [◦]: Results are taken from (Dettmers et al., 2018). Other results are taken from the corresponding original papers.

	WN18RR					FB15k-237					YAGO3-10				
Model	MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10
TransE†	3384	.226	-	-	.501	357	.294	-	-	.465	-	-	-	-	-
DistMult◦	5110	.43	.39	.44	.49	254	.241	.155	.263	.419	5926	.34	.24	.38	.54
ComplEx◦	5261	.44	.41	.46	.51	339	.247	.158	.275	.428	6351	.36	.26	.4	.55
ConvE◦	4187	.43	.40	.44	.52	224	.325	.237	.356	.501	1671	.44	.35	.49	.62
RotatE	3340	.476	.428	.492	.571	177	.338	.241	.375	.533	1767	.495	.402	.55	.67
Rotate3D	3328	.489	.442	.505	.579	165	.347	.250	.385	.543	-	-	-	-	-
QuatE	3472	.481	.436	.500	.564	176	.311	.221	.342	.495	-	-	-	-	-
DualE	-	.482	.440	.500	.561	-	.330	.237	.363	.518	-	-	-	-	-
Rot-Pro	2815	.457	.397	.482	.577	201	.344	.246	.383	.540	1797	.542	.443	.596	.669
HousE-r	1885	.496	.452	.511	.585	165	.348	.254	.384	.534	1449	.565	.487	.616	.703
HousE	**1303**	**.511**	**.465**	**.528**	**.602**	**153**	**.361**	**.266**	**.399**	**.551**	**1415**	**.571**	**.491**	**.620**	**.714**

Table 4. MRR for the models tested on each relation of WN18RR.

Relation Name	RotatE	QuatE	HousE-r	HousE
hypernym	0.154	0.172	0.182	**0.207**
instance_hypernym	0.324	0.362	0.395	**0.440**
member_meronym	0.255	0.236	0.275	**0.312**
synset_domain_topic_of	0.334	0.395	0.396	**0.428**
has_part	0.205	0.210	0.217	**0.232**
member_of_domain_usage	0.277	0.372	0.415	**0.453**
member_of_domain_region	0.243	0.140	0.281	**0.395**
derivationally_related_form	0.957	0.952	0.958	**0.958**
also_see	0.627	0.607	0.638	**0.640**
verb_group	0.968	0.930	0.968	**0.968**
similar_to	1.000	1.000	1.000	**1.000**

MRR: Mean Reciprocal Rank,

Performance of HousE

Table 5. MRR for the models tested on RMPs in FB15k-237.

Task	RMPs	RotatE	HousE
Predicting Head (MRR)	1-to-1	0.498	**0.514**
	1-to-N	0.475	**0.479**
	N-to-1	0.088	**0.114**
	N-to-N	0.260	**0.286**
Predicting Tail (MRR)	1-to-1	0.490	**0.502**
	1-to-N	0.071	**0.086**
	N-to-1	0.747	**0.778**
	N-to-N	0.367	**0.392**

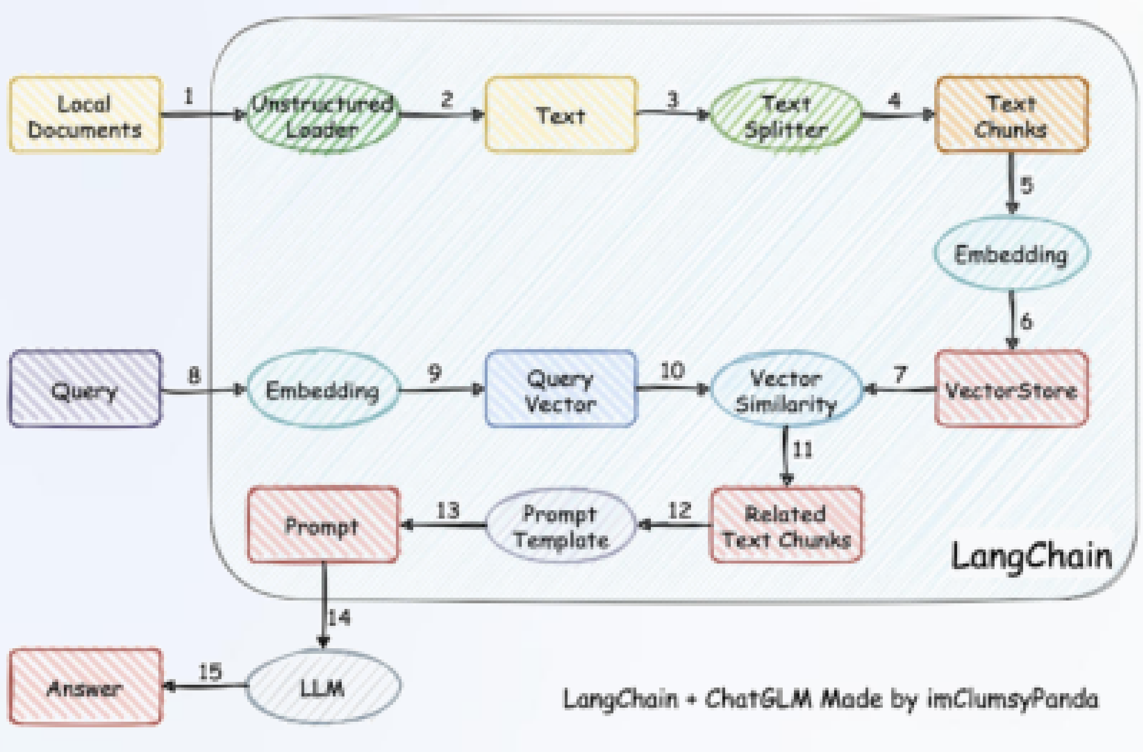
1-to-N and N-to-1 relations. For example, HousE outperforms RotatE on 1-to-N relation *member_of_domain_region* and N-to-1 relation *instance_hypernym* with 62.55% and 35.80% relative gains, respectively.

RMP: Relation Mapping Properties

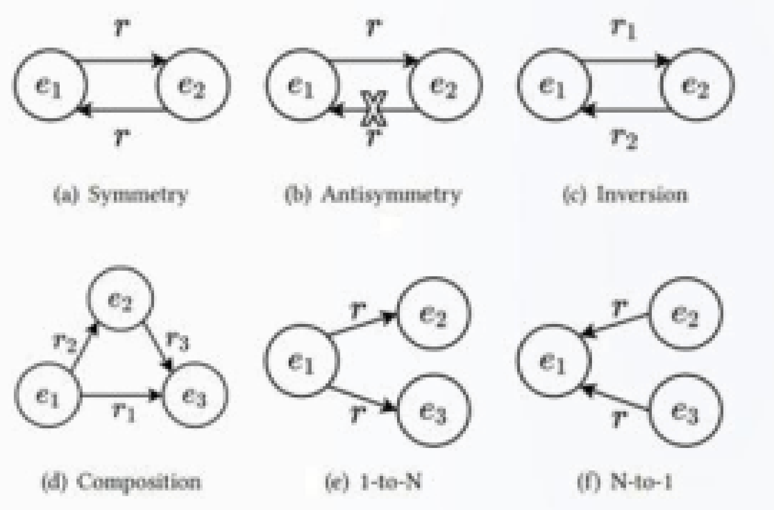
Neural modular and compositional approaches have been explored to automatically perform desired sub-task decomposition, enhancing interpretability and adaptability across various reasoning tasks. Early work posits that complex reasoning tasks are fundamentally compositional and proposes neural module networks (NMN) to decompose them into subtasks. However, these methods rely on brittle off-the-shelf parsers and are limited by module configurations. Some later work, takes a step further by predicting instance-specific network layouts in an end-to-end manner, without relying on parsers, using reinforcement learning [58] and weak supervised learning. In visual reasoning, models comprising a program generator and an execution
engine have been proposed to combine deep representation learning and symbolic program execution. In the domain of mathematical reasoning, an interpretable solver has been developed to incorporate theorem knowledge as conditional rules and perform symbolic reasoning step by step. Our work takes inspiration from neural module networks, yet it offers several distinct advantages.

Knowledge base core flowchart

Langchain-based Q&A application based on local knowledge base. The process is as follows:
load document → read document → split text → embed text → embed query → Match the top k of the text vectors that are most similar to the stationery vectors → The matched text is added to the prompt as context along with the question → Submit to the llm to generate a Q&A



Relation Categories



Relation categories in knowledge graphs are fundamental for accurately representing and reasoning about the complex web of connections in real-world data. They enhance semantic understanding, improve data integrity, optimize query performance, and enable advanced machine learning applications.

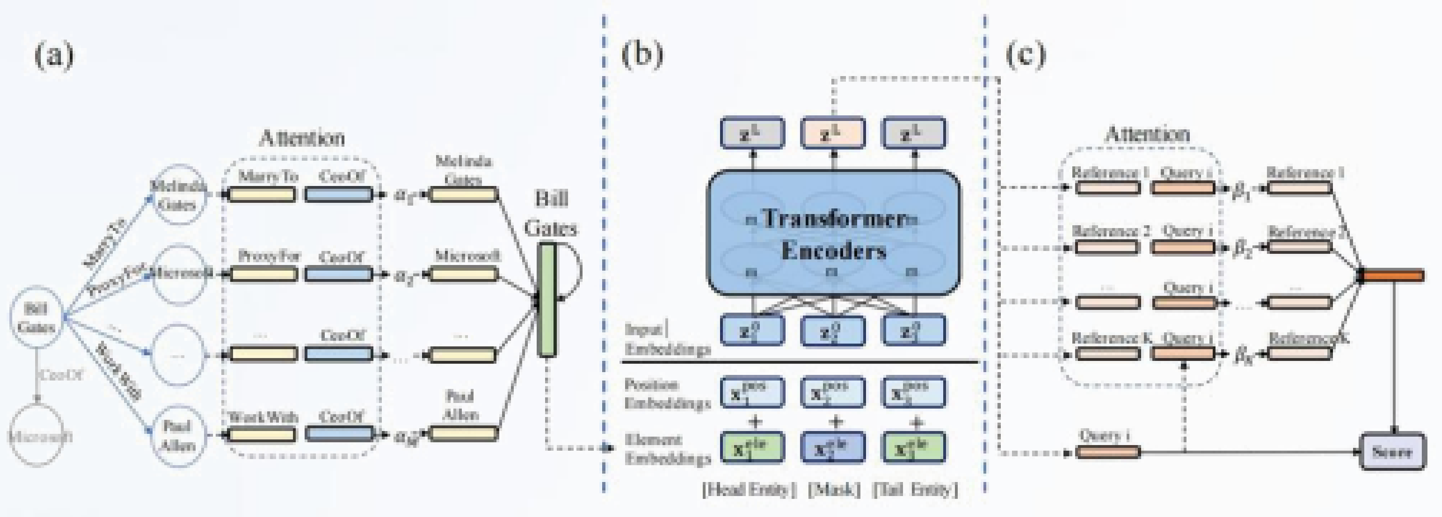


Figure 2: The framework of FAAN: (a) Adaptive neighbor encoder for entities; (b) Transformer encoder for entity pairs; (c) Adaptive matching processor to match K-shot references and the query.

Three Major Parts of FAAN

(1) Adaptive neighbor encoder to learn adaptive entity representations;

(2) Transformer encoder to learn relational representations for entity pairs;

(3) Adaptive matching processor to compare the query to the given references.

Future study

Althoughh large language models (LLMs) have achieved excellent performance in a variety of evaluation benchmarks, they still struggle in complex reasoning tasks which require specific knowledge and multi-hop reasoning. We will try more models that can improve llm accuracy. Through improved prompt strategies (such as CoT and ChatCoT), different toolkits and logical chain thinking model frameworks are invoked to make large language models perform better in different areas of expertise.

Reference list

[1]Sheng, J., Guo, S., Chen, Z., et al. (2020). Adaptive Attentional Network for Few-Shot Knowledge Graph Completion. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1681-1691). 【with code】
[2]Zhang, C., Yao, H., Huang, C., et al. (2020). Few-shot knowledge graph completion. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(03), 3041-3048. 【with code】
[3]Shomer, H., Jin, W., Wang, W., et al. (2023). Toward degree bias in embedding-based knowledge graph completion. In Proceedings of the ACM Web Conference 2023 (pp. 705-715). 【with code】
[4]Zhao, W. X., Zhou, K., Li, J., et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.
[5]Xi, Z., Chen, W., Guo, X., et al. (2023). The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864.
[6]Lu, P., Peng, B., Cheng, H., et al. (2024). Chameleon: Plug-and-play compositional reasoning with large language models. Advances in Neural Information Processing Systems, 36.
[7]Zhao, A., Huang, D., Xu, Q., et al. (2024). Expel: LLM agents are experiential learners. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 19632-19642.