TouhouQANet: Touhou QA Knowledge Graph Annotation with GPT-4

Letian Peng
University of California, San Diego
lepeng@ucsd.edu

Abstract

In this paper, we unveil TouhouQANet, a Chinese knowledge graph for the Touhou Project characters, built using GPT-4 and data from ThbWiki. We cover key characters across nine games from "Embodiment of Scarlet Devil" to "Subterranean Animism." Constructing a Touhou Project dataset has been challenging due to the complexity and subjectivity of the game's universe. To overcome this, we've created guidelines for GPT-4 to annotate new edges in the graph, with each edge being associated with a question to minimize ambiguity. We've annotated 1208 sentences, establishing 2657 edges with question-answer pairs. Despite GPT-4's capabilities, it has limitations like hallucination, leading to the generation of non-factual information. With TouhouQANet, we've demonstrated the ability to fill these knowledge gaps in large language models, pointing towards a promising approach for enhancing the accuracy of AI systems in handling specific domains. Our knowledge graphs, datasets, and codes are released at github.com/KomeijiForce/TouhouQANet

1 Introduction

In this research paper, we present TouhouQANet, a novel Chinese knowledge graph developed for characters from the Touhou Project, a popular Japanese "bullet hell" shooter game series. Our knowledge graph is constructed by extracting, refining, and annotating information from ThbWiki[1], a dedicated wiki for the Touhou Project. Focusing on the key characters from nine games, starting from "Embodiment of Scarlet Devil" and extending up to "Subterranean Animism," we aim to create a comprehensive knowledge base for these characters.

Developing a dataset for the Touhou Project has traditionally posed several challenges, primarily due to its vast and intricate universe with characters possessing complex histories and relationships.

[1] https://thwiki.cc/

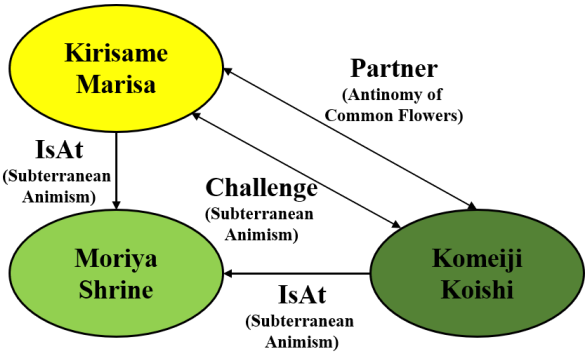

Figure 1: An instance shows the complexity of the knowledge graph of the Touhou Project. Multiple edges are involved since characters have different relationships in different stories.

As the series grew over the years, the depth of its lore increased, leading to issues in properly classifying and organizing the information. Besides, the interpretation of characters and their interactions often involve a certain degree of subjectivity, leading to ambiguity and potential discrepancies in the dataset as shown in Figure 1.

We address these challenges by manually curating guidelines and providing few-shot examples to guide GPT-4 in annotating new edges in the knowledge graph. Each edge we generate is associated with a specific question, designed to both enrich the knowledge contained within the graph and to resolve any ambiguity that may exist. Through this methodology, we have annotated 1208 sentences from ThbWiki, establishing 2657 edges with corresponding question-answer pairs.

Large language models (LLMs) like GPT-4 (OpenAI, 2023), despite their impressive capabilities, do have certain limitations when it comes to remembering specific domain knowledge. One notable weakness is their propensity to hallucinate, i.e., to generate information that may seem plausible but is not based on real data. This can result in the propagation of incorrect or misleading infor-

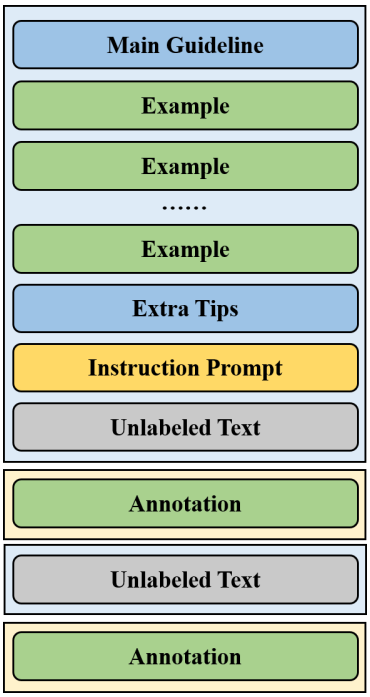

Figure 2: An overview of our annotation pipeline.

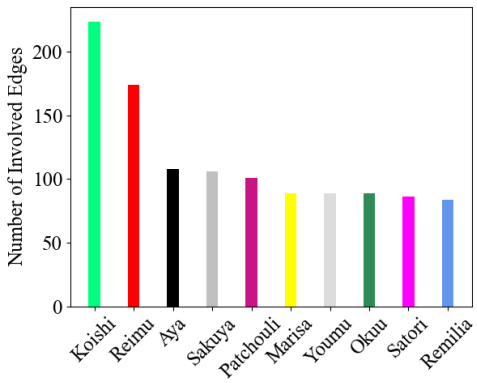

Figure 3: Statistics of character appearance in edges of TouhouQANet.

mation, particularly when dealing with niche areas like the Touhou Project, where the model's training data may be limited.

By introducing the TouhouQANet, we aim to address this shortcoming. Our work demonstrates that our knowledge graph is capable of patching these gaps in knowledge that exist in large language models. We believe that this approach, combining the power of large language models with well-structured and expertly curated knowledge graphs, represents a promising direction for enhancing the reliability and accuracy of AI systems in handling specific domains.

2 TouhouQANet

2.1 Annotation with LLMs

To automatize the annotation, we provide guidelines and few-shot examples to prompt GPT-4 as presented in Figure 2.

Guidelines for annotation consist of specific instructions regarding the format and structure of the annotations. In this case, the annotations are represented in JSON format, utilizing the "input" and "output" keys. The "input" key contains the original text that requires annotation, while the "output" key holds a list of annotated edges derived from the input. Each element in the output list contains four keys: "subject", "object", "predicate", and "question". The first three elements correspond to the traditional knowledge graph's triplet, while the question element serves to specify the predicate and resolve any potential ambiguity. Additionally, to prevent misunderstandings, explanations for certain predicates are included in the annotated edges. Moreover, if the object in an edge consists of a conjunction of different entities, it should be split into separate components.

Few-shot Examples are an invaluable resource in guiding the annotation process for GPT-4. These examples are human-annotated results that aim to provide a reference for the annotation of GPT-4. As the input text is already incorporated within the JSON data, we provide a list of several JSON data instances to assist the model in comprehending various annotation intricacies. Throughout the annotation process, annotations generated by GPT-4 of high quality are incorporated into the examples. This augmentation helps enhance the diversity and coverage of the few-shot examples, ensuring a more comprehensive and robust annotation process for GPT-4.

Unlabeled Texts refer to manually selected texts that contain valuable descriptions of Touhou characters. These texts are typically sourced from three different resources. Firstly, contributors to ThbWiki provide character attribute summaries. Secondly, character description documents from the games themselves are utilized. Lastly, the comprehensive and detailed information found in *Perfect Memento in Strict Sense* also contributes to these texts.

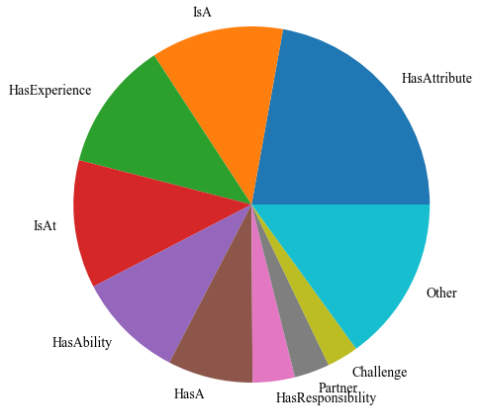

Figure 4: Statistics of edge predicates in TouhouQANet.

Property	Value
Number of Texts/Edges	1208/2657
Number of Elements/Predicates	2571/90
Average Length of Q/A	15.59/7.36

Table 1: More statistics of TouhouQANet.

2.2 Statisitcs

Figures 3 and 4 present the statistics of TouhouQANet regarding character appearances and edge predicates, respectively.

Character The top 10 characters are all immensely popular among fans of the Touhou Project. Consequently, it's unsurprising that they have received a higher number of annotations in the knowledge graph. The character with the most annotations is *Komeiji Koishi*, which has been annotated 224 times. Generally, popular characters receive around 100 annotations.

Predicate Among the top 6 predicates in TouhouQANet, 5 of them (*HasAttribute*, *IsA*, *IsAt*, *HasAbility*, *HasA*) are common in other knowledge graphs such as ConceptNet (Speer and Havasi, 2012; Speer et al., 2017). This indicates that TouhouQANet shares common relationships with traditional knowledge graphs. In addition, TouhouQANet includes popular predicates like *HasExperience*, *HasResponsibility*, *Partner*, and *Challenge*, which reflect the characteristic properties of a knowledge graph focused on character documents.

Other statistics of TouhouQANet are presented in Table 1.

3 Experiments

3.1 Touhou Knowledge-125

We conducted performance tests on LLM-based question-answering systems specifically for the Touhou Project. To evaluate their capabilities, we created a test dataset called "Touhou Knowledge-125" (THK-125) comprising 5 groups of 25 questions each, with varying difficulty levels.

Here are the different types of questions included in the dataset:

- Easy questions: These queries pertain to well-known attributes of Touhou that can be answered by individuals who are not necessarily fans of the series. For example, *What is Hakurei Reimu's occupation?*
- Normal questions: These inquiries delve into common knowledge shared among Touhou fans but may not be widely known by those unfamiliar with the series. An example would be, *What connects Gensokyo with the outside world?*
- Hard questions: This category involves more intricate details within Touhou, which might be unfamiliar to fans unless they have memorized them. These questions may also require multi-hop inference. An example of a hard question is, *What is the mechanism used by Flandre Scarlet to destroy objects?*
- Lunatic questions: These queries are highly specific and involve complex chains of thought. They often necessitate a comprehensive understanding of multiple sources of knowledge within the Touhou Project. For instance, *What contradiction exists between the records in Touhou Keijihan and Perfect Memento in Strict Sense regarding Sakuya's magic show?*
- Extra questions: Falling between the difficulty levels of Normal and Hard, these questions pertain to information after the events of "Subterranean Animism." They cover content not yet addressed in the current version of TouhouQANet. An example question could be *Where does the story in Wily Beast and Weakest Creature happen?*

It is important to note that while the Lunatic questions may not pose significant challenges to

	Method	Easy	Normal	Hard	Lunatic	Extra	Average
GPT-3.5	Vanilla	44%/24%	20%/8%	12%/4%	0%/0%	0%/0%	15.2%/7.2%
	w/ BM25	72%/68%	52%/52%	36%/28%	20%/20%	0%/0%	36.0%/33.6%
	w/ SBERT	72%/68%	56%/56%	32%/28%	20%/20%	0%/0%	36.0%/34.0%
GPT-4	Vanilla	**88%**/84%	40%/32%	12%/12%	12%/12%	**8%/8%**	32.0%/32.0%
	w/ BM25	**88%/88%**	**76%/68%**	**52%/48%**	**24%/24%**	**8%/8%**	**49.6%/47.2%**
	w/ SBERT	**88%/88%**	64%/60%	36%/36%	**24%/24%**	**8%/8%**	44.0%/43.2%

Table 2: Test results of vanilla and patched LLMs on THK-125. The metrics are Correctness/Factuality. GPT here refers to the ChatGPT version.

professional Touhou fans, they remain difficult for GPT-4 due to their requirement for a comprehensive understanding of the Touhou universe.

Metrics used in the experiment are Correctness and Factuality. **Correctness** measures the ratio of correct answers, while **Factuality** is a stricter metric that evaluates the ratio of correct answers accompanied by factual explanations. Due to the varied outputs produced by LLMs, we manually examined the metrics for the answers provided by the models.

QA Performance of vanilla LLMs is presented in Table 2. GPT-4 significantly outperforms its predecessor, GPT-3.5, on the Easy and Extra categories, and also exhibits a marginal improvement in the Hard and Lunatic categories. This indicates that the model's capacity to handle both basic and more complex Touhou-related queries has been enhanced in the newer iteration. Nevertheless, even with these improvements, the overall performance of the vanilla GPT-4 model remains relatively low, especially in the higher difficulty categories. This highlights the inherent limitations of these models when dealing with domain-specific knowledge without any form of external enhancement or information retrieval system.

3.2 Patching Results

TouhouQANet also serves as a valuable resource for supplementing the knowledge about the Touhou Project that is unknown to the LLms. To achieve this, we utilize two retrievers, BM25 and Sentence BERT (Reimers and Gurevych, 2019), for patching.

BM25 is a statistical algorithm commonly used in information retrieval systems. It measures the relevance of documents to a given query by taking into account factors such as term frequency, document length, and term weighting. Based on these metrics, BM25 ranks documents according to their level of relevance. We use the cosine similarity between pooled representations to rank the similarity between sentences.

Sentence BERT [2] is a technique designed to enhance the representation of sentences by generating dense vector embeddings. It leverages the Transformer architecture and a Siamese network structure to optimize the embeddings for capturing semantic meaning. Sentence BERT proves highly effective in tasks like text retrieval, sentence similarity evaluation, and clustering.

To patch the unknown knowledge of LLMs, we retrieve the five most relevant question-answer pairs using the aforementioned retrieval mechanisms. These pairs are then attached before the question, enabling us to enrich the LLMs' understanding.

QA Performance w/ Patching When looking at the performance of the models with BM25 and SBERT enhancements, it is clear that these information retrieval mechanisms significantly improve the models' performance across all difficulty levels. GPT-4's performance with both BM25 and SBERT is superior to that of GPT-3.5 with the same enhancements, which is consistent with the overall trend. GPT-4 combined with BM25 produces the best performance across all categories, demonstrating the value of this retrieval mechanism in improving the performance of large language models in a domain-specific context. The results also indicate that BM25 appears to be a more effective and efficient complement to the GPT-4 model in this particular context.

[2]We use sbert-base-chinese-nli as the retriever.

4 Conclusion

In conclusion, our paper presents TouhouQANet, a Chinese knowledge graph for the Touhou Project characters, constructed using GPT-4 and data from ThbWiki. By leveraging guidelines for annotation and question-answer pairs, we successfully addressed the challenges of complexity and subjectivity in the Touhou universe. The results demonstrate the potential of knowledge graphs in enhancing the accuracy of AI systems within specific domains and provide a valuable resource for the Touhou community.

References

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 3679–3686. European Language Resources Association (ELRA).