Prompt-enhanced Large Language Models for Automated Construction of circRNA-themed Hyper-relational Knowledge Graph

Yingjie Xiao[1] and **Lei Duan**[1]

[1]College of Computer Science, Sichuan University, Chengdu, China
xiaoyingjie_@stu.scu.edu.cn

Abstract

We focus on circular RNAs (circRNAs), vital in disease progression and therapeutic targeting. To capture complex interactions, we employ hyper-relational knowledge graphs (HKGs) for circRNA-disease-drug links. Leveraging large language models (LLMs), our framework combines prompt engineering for automated HKG construction from circRNA literature. Fine-tuned with biomedical prompts, LLMs extract hyper-relational facts via tailored prompts, followed by authenticity checks, fostering reliable and research-spurring knowledge extraction.

1 Introduction

Circular RNAs (circRNAs) have been increasingly studied for their potential as disease biomarkers and therapeutic targets, due to their regulatory roles in gene expression and widespread implications, such as enhancing disease prediction, facilitating drug discovery, and advancing precision medicine, thereby reshaping the understanding and treatment of complex diseases [Liu and Chen, 2022].

While prior studies have largely relied on tabular datasets or circRNA-associated heterogeneous information networks to unravel the interconnected threads of circRNA functionality, these approaches struggle in capturing the nuanced complexity and contextuality inherent in biomedical systems. In contrast, hyper-relational knowledge graphs (HKGs) emerge as a superior paradigm, offering a higher insight capable of comprehensively storing and depicting real-world knowledge.

Extending from knowledge graphs (KGs), HKGs that hold hyper-relational facts (H-facts) involving no fewer than two entities, are much more prevalent and applicable in real-world scenarios. Based on this hyper-relational modeling, completion and reasoning over real-world HKGs are effective ways of knowledge discovery. An H-fact consists of a primary triple (*subject*, *relation*, *object*) coupled with several auxiliary (*attribute*:*value*) qualifiers. We illustrate with an example.

Example 1. {*(circPPM1F, may exacerbate, T1DM),* {*(by Promoting activation: Macrophage M1)*}}*, which represents one of the cutting-edge research progresses regarding circRNA and diabetes mellitus type 1 (T1DM) [Zhang* et al.*, 2020]: circPPM1F promotes M1 macrophage activation, thereby playing a crucial role in exacerbating the progression of T1DM.*

By this example we can see that, this hyper-relaional modeling empowers researchers to traverse intricate networks of causality and correlation, thereby clearly understanding disease mechanisms, drug actions, and the potential synergies between them in a fact-based manner. Nevertheless, traditional supervised manners for KG construction, such as named entity recognition (NER), relation extraction (RE), and event extraction (EE), pose several challenges (CHs) when it comes to circRNA-associated HKGs, e.g., the expensive consumption of manual annotation and training cost.

Addressing these challenges, we introduce an automated framework leveraging prompt-optimized LLMs to build the circRNA-themed HKG. Central to our approach is refining LLMs' capacity to discern biomedical entities, relations, alongside their general reasoning and H-fact structuring skills. We employ multi-layered prompts to boost circRNA data extraction from literature and adopt a dual-evaluation strategy combining LLMs and human expertise. Our main contributions can be summarized accordingly:

- We propose an automated construction framework for circRNA-themed HKGs, which provides a novel insight for biomedical researches, thereby inspiring and fully exploring the potential of circRNAs.
- We propose a multi-scale enhancing strategy for prompt-tuning LLMs, presenting a more efficient and resource-conscious alternative compared to conventional supervised methods for information extraction.

2 Methodology

2.1 Multi-scale Prompt Construction

For the proposed method, as illustrated in Figure 1, we design triple instructions, and then feed them into the frozen LLMs that only can infer: 1) **Fact-structure aware instructions**, which tell LLMs the background of biomedical information extraction and the representation format of H-facts; 2) **Domain-knowledge enhanced instructions**, which instruct LLMs to decompose specific domain tasks into several continuous and progressive tasks; 3) **Ground truth self-evaluation** with some real biomedical examples of inputs and outputs, which enhance the consciousness that LLMs just learned again.

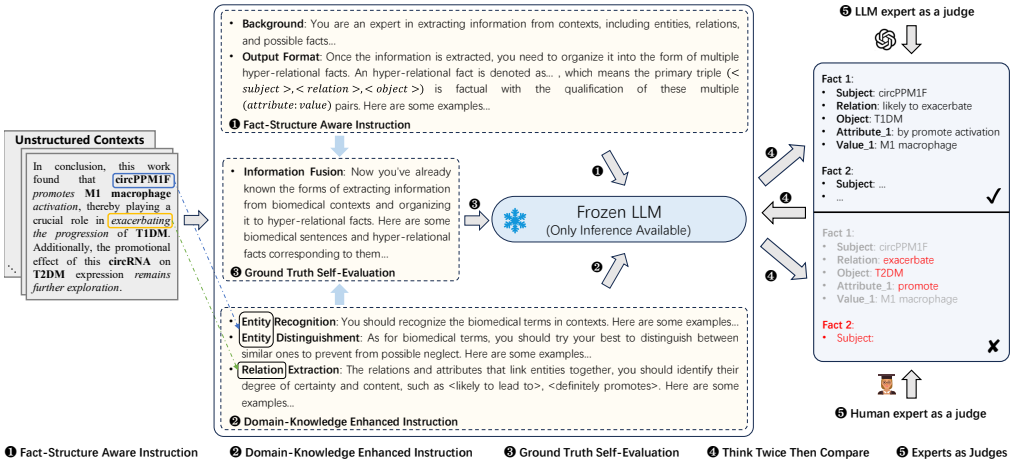

Figure 1: The overview of the proposed method.

2.2 Think-Twice-Then-Compare strategy

To counteract the randomness of large-scale models that produce inconsistent outputs for identical inputs, we implement a recursive strategy. We generate two outputs per input and re-submit them, along with the initial input, for the model to evaluate and pick the best one. This self-assessment loop enhances output consistency and reduces variability, leading to optimized results from probabilistic models.

2.3 Experts as judges

Extracting H-fact within the circRNA domain poses unique challenges due to its high level of specialization. It is challenging to assess the quality of the constructed HKG data using comparative methods like CubeRE [Chia *et al.*, 2022] for such domain-specific data. To counteract this limitation and ensure the veracity of generated content, we introduce two strategies to evaluate the data quality, involving LLMs and humans.

Our method innovatively incorporates advanced LLMs like GPT-4 or Qwen 2.5, expert-empowered models, in a rigorous evaluation process. Carefully crafted prompts assess the accuracy and completeness of H-facts from primary models, focusing on triple correctness and qualifier integrity. Each H-fact from circRNA research receives a 0-10 score, with 10 indicating perfect relevance and 0, gross error. A grading system filters high-quality facts (scores 8) for confident inclusion in the circRNA-HKG, while those below 5 trigger reevaluation. Facts scoring between 5 and 8 undergo random human expert reassessment to refine our circRNA-themed HKG further.

3 Experiments

circRNA-themed literature data. Our study's experiments entailed a thorough data gathering from PubMed[1], focusing on recent circRNA research. Using targeted keywords, we compiled 23,850 articles—rich in circRNA-themed insights. Abstracts, as concise findings summaries, are chosen for HKG construction to capture the field's cutting-edge dynamics, ensuring relevancy and depth. A content analysis validate the dataset's contemporaneity and relevance. This groundwork paves the way for our advanced language model to transform these abstracts into a comprehensive, interconnected circRNA-themed HKG, facilitating new research insights and precision medicine advancements.

LLM selections. For extracting circRNA-themed H-facts, we employ Llama-2-7B and Llama-3-8B as base models. Our strategy involved using GPT-4, a superior LLM, for evaluating extracted H-facts. The decision to leverage GPT-4 stems from its heightened capacity to discern factual accuracy crucial in biomedical contexts. Despite increased costs, its deployment ensures meticulous quality control due to its advanced comprehension capabilities.

circRNA-themed HKG construction. We've implemented a circRNA-themed HKG encompassing over 5,000 circRNAs and 40 diseases. Acknowledging its incompleteness given existing knowledge, further exploration is imperative. Computational constraints also indicate potential for advancement in our zero-shot HKG construction approach.

4 Conclusion and Future Work

This work presents a automated construction framework for circRNA-themed HKG. As future work, we plan to take more data sources into consideration, and employ stronger LLMs to construct high-quality biomedical HKGs.

References

[Chia *et al.*, 2022] Yew Ken Chia, Lidong Bing, Sharifah Mahani Aljunied, Luo Si, and Soujanya Poria. A dataset for hyper-relational extraction and a cube-filling approach. In *EMNLP*, pages 10114–10133, 2022.

[Liu and Chen, 2022] Chu-Xiao Liu and Ling-Ling Chen. Circular rnas: Characterization, cellular roles, and applications. *Cell*, page 2390, Jun 2022.

[Zhang *et al.*, 2020] Caiyan Zhang, Xiao Han, and Lan Yang et al. Circular RNA circPPM1F modulates M1 macrophage activation and pancreatic islet inflammation in type 1 diabetes mellitus. *Theranostics*, 10(24):10908–10924, 2020.

[1] https://pubmed.ncbi.nlm.nih.gov/