Dear editor and distinguished reviewers,

We thank you for the detailed and helpful comments about our manuscript. We have carefully revised the article based on your recommendations. Below, we provide a detailed account of the responses and revisions made to the manuscript.

Yours sincerely,

Vasile Ionut Remus Iga

Gheorghe Cosmin Silaghi

Reviewer 1

Thank you for all of your comments that helped us improve our work. Please consider our responses to all of your questions discussed below.

Question 1:

“First, on questions concerning content and overall presentation. The main issue concerning presentation is that, in my view, there is a confusing mix of discussion concerning the specific systems developed by the authors and the broad techniques under analysis. It would be important to separate what are the textual parts that discuss the tested ontologies/extractors, and the parts that are really related to LLMs in general. It is actually hard to explain exactly what to do here, but in several paragraphs I was confused as to what the authors were saying: were they describing some specific system that was only used in the specific experiments, or were they discussing the key concepts that are discussed in the experiments? This is the main issue; I suggest the authors read again their text thinking as a new reader, and try to improve the text as much as possible in this regard.”

Response:

To improve the readability and accessibility of the paper, we have made the following structural and content-related revisions:

Abstract

- The abstract was entirely rewritten to enhance clarity and better communicate the purpose and contributions of the paper.

Introduction

- The paragraph discussing TOD systems, our prior work, and CRUD operations was revised for improved clarity and divided into two paragraphs. One now provides a general definition of TOD systems, and the other focuses on our specific implementation. Additionally, we have added **Figure 1**, which illustrates an example pipeline for a TOD system.
- From the sixth paragraph onward, we consistently refer to the two datasets collectively as **TODSet**, a customized, two-part dataset.
- The second-to-last paragraph was updated to mention an additional evaluation dataset featuring a more complex ontology, an addition beyond what was included in the original conference version.

Related Work

- As per the suggestion, steps 4 and 5 have been separated from the initial three components and described as a distinct method for greater clarity.

Methodology

- The order of the subsections "Format and distribution of the datasets" and "Prompt engineering" was reversed. Introducing prompt engineering first improves the understanding of subsequent sections.
- **Definition 2** was broadened to offer a more comprehensive explanation of LLMs, and now explicitly distinguishes transformer-based models as a subcategory.

Prompt Engineering (Methodology)

- The initial paragraph, which described code-level details, was removed as it did not significantly benefit the reader.
- The sections discussing concepts such as Zero-, One-, Few-Shot learning, Direct Prompting, In-Context Learning, and Chain of Thought were expanded and rewritten to provide more background for readers less familiar with these terms.
- The paragraph describing each prompt level was revised for better clarity and now explains the rationale behind the prompt construction. Additionally, examples of all prompts have been added to the paper's public repository.

Format and Distribution of the Datasets (Methodology)

- The description of input phrase selection, classification, and content was reworded for improved clarity.
- A new paragraph introduces the **FootballSet** dataset. **Figure 4** presents the corresponding ontology, while **Tables 3 and 4** include example texts and class-wise phrase distributions.
- The section discussing alternative and false-positive triples was expanded to provide more context. **Figure 2** (now **Figure 5**) was also updated as it was difficult to grasp.

Metrics (Methodology)

- The explanation of the flexible metric was revised and extended to improve clarity.
- Descriptions of the various penalty types were reformulated for precision and consistency.

Results and Discussion

- Cost-related details from initial testing were removed, as they are no longer relevant.
- Results from tests conducted on the **FootballSet** dataset are now included. Two new tables (**Tables 11 and 12**) present model performance metrics.
- Table highlights were standardized: the best result for each model per level or class is in **bold**, the overall best result among all models is in *italics*, and the best overall model or level/class is underlined. These formatting choices enhance result readability.
- The discussions are now organized by topic:
  - "Discussion on the Influence of Prompt Engineering"
  - "Discussion on the Influence of Text and Ontology Structure"
  - "Discussion on LLM Performance"

  This reorganization makes each argument easier to follow.

Discussion on the Influence of Text and Ontology Structure (Results and Discussion)

- A new conclusion titled **"Elaborate Ontologies Are More Difficult to Grasp"** was added, emphasizing how ontology complexity affects model performance across the datasets.
- The section **"Complex Class Types Do Not Imply More Difficult Reasoning"** was revised to incorporate insights from the **FootballSet** results.

Discussion on LLM Performance (Results and Discussion)

- The conclusion titled **"GPT-4o is More Consistent and Performant, While GPT-3.5-Turbo Achieves the Best Results"** was revised to include performance details on the **FootballSet** dataset.

Conclusion

- The conclusion was entirely rewritten for improved clarity. It now includes a discussion of the study’s limitations and a more precise articulation of future research directions.

Question 2:

“- The ontology in Figure 1, used throughout the text, is quite simple; how would conclusions change with a more involved ontology? Any comments?”

Response:

To evaluate our methodology in a more complex setting, we adapted the DBpedia-WebNLG sports ontology and dataset, originally developed by Mihindukulasooriya [1]et al., to align with our format. This adaptation, inspired by their work, enabled us to assess model performance under increased ontological complexity. Throughout the paper, we refer to the resulting dataset as *FootballSet*. The final conclusions, including results from this extended evaluation, are presented in the *Results and Discussion* section, specifically in Tables 11 and 12. In summary, all models demonstrated a decline in performance, underscoring the greater challenge posed by more complex ontologies. Nonetheless, the overall conclusions remained consistent with those previously established.

Question 3:

“- The "flexible" metric is an interesting contribution, and the explanations related to it are interesting, but all of it seems to be quite ad hoc and hard to justify. I suggest more discussion is provided.”

Response:

As noted in our response to Q1, we have revised the *Methodology – Metrics* section to provide a clearer and more detailed explanation of the flexible metric and the various penalty types. This updated description highlights the relevance and practical usability of this measurement approach. We also clarify in the same section that the specific values used in our experiments were selected for our current setup, but users are free to choose alternative values based on their needs.

Additionally, in the *Conclusion* section, we acknowledge that the current system is limited to a set of predefined penalty types. However, we also outline plans for future work aimed at modularizing the flexible metric, allowing users to define and incorporate custom penalties as needed.

[1] https://github.com/cenguix/Text2KGBench

Question 4:

"- The discussion of CRUD operations at the beginning of Page 2 is quite confusing; that paragraph should be rewritten."

Response:

The paragraph that discusses TOD systems and our previous work, alongside CRUD operations, was extended, reformulated and split in two, to accommodate the TOD systems general definition and our specific system. Additionally, we have included a figure (Figure 1) showing an example of a pipeline TOD system.

Question 5:

"- End of Page 2: several issues are discussed, but it is hard to know what is the actual point of the discussion. What is intended? What are the exact points of the datasets? And so on."

Response:

We have revised the description of our contributions, as outlined at the end of the Introduction section, to enhance clarity. Our primary objective is to evaluate whether different types of LLMs are suitable for the task of Knowledge Graph Construction (KGC). To this end, we apply various prompt engineering techniques to determine which approaches yield the best performance for this specific task.

Additionally, we introduce a flexible evaluation metric designed to account for certain types of errors commonly made by LLMs—errors that can be easily corrected during post-processing. This metric allows us to focus more precisely on the models' ability to extract semantically meaningful triples.

The TODSet dataset was specifically created for KGC. All triples were manually extracted to facilitate a direct comparison between model outputs and expert annotations. We have also added two new categories of manually curated triples—*alternative* and *accepted false positives*—to support and make full use of the flexible evaluation metric.

Question 6:

"- Page 3, line 4: items 4 and 5 are not part of a pipeline, they seem to be separate modules."

Response:

As suggested, steps 4 and 5 were reformulated as a separate method from the other three components.

Qquestion 7:

“- Page 3, middle of page: the authors write "they", "their", etc, and it is often difficult to know who are the referred entities. This happens a number of times in the paper.”

Response:

All mentions of the words “they”, “their”, “them” were analyzed and, in most cases, the text was reformulated removing ambiguous references with clear mentioning of the referred subjects.

Question 8:

“- Definition 2 mixes LLMs and the specific transformer architecture, this is confusing (there are language models that are not transformers!).”

Response:

Definition 2 now includes a broader definition of LLMs, while we now specifically mention the transformer-based models as a different category.

Question 9:

“- What is the meaning of Expression (3)? It just offers an equality.”

Response:

Expression 3 is intended for readers who may be less familiar with the typical input format for a large language model (LLM) and the expected output structure for knowledge graph construction (KGC).

Question 10:

“- Table 1: "manager", not "maager".”

Response:

The phrase specified in the third line of Table 1 contains grammatical mistakes on purpose, as a representation of text types pertaining to the MS1 category. This type of samples emulates real-world scenarios where human users might write input texts in such manner, but the model should still be capable of correctly solving the task at hand.

Question 11:

"- Figure 2 is very hard to understand."

Response:

Figure 2 was changed to better show an example of an input text and associated extractable triples of different types. It is now referred to as Figure 5.

Question 12:

"- Page 2, line 10: I believe it should "the literature", as an example of a small suggestion that could be applied to several other sentences.
- Page 2, line 12: "is that ... to", seems to be incorrect (remove "that").

- Page 6, sentence "For a more comprehensive...", is very hard to parse. What does it mean?
- Page 8, "let's ask a model" seems weird.
- Page 9, line 3, "Experts [3]" misses a space.
- Table 7 and Table 8 do not have underline cells; why is it?
- Page 12, "Complex Class Types Do Not"... should be "Does".
- Page 13 mentions "Adhere" but does not seem to agree with the content of the paragraph.

- Mathematical expressions should end with period/comma, as appropriate; after a mathematical expression, no indentation is there is no new paragraph."

Responses:

We have followed all of these suggestions and corrected the texts in all cases.

Reviewer 2

Thank you for all of your comments that helped us improve our work. Please consider our responses to all of your questions discussed below.

Question 1:

"- Concepts like zero-shot, few-shot, and TOD systems are introduced but not explained in detail. Adding a bit more background would make the paper more accessible to readers who are less familiar with these topics."

Response:

These concepts were initially introduced in the Introduction section and are further detailed in the Methodology – Prompt Engineering subsection. In this version of the manuscript, we have expanded and revised the relevant paragraphs to provide additional background on concepts such as Zero-, One-, and Few-Shot learning, Direct Prompting, In-Context Learning, and Chain-of-Thought reasoning. These updates aim to make the content more accessible to readers who may be less familiar with these approaches.

Question 2:

"- The flexible evaluation approach is innovative, but the rationale for choosing specific penalty percentages isn't clear. Explaining why certain values were picked would make the methodology more transparent and credible."

Response:

As noted in our response to the Q1 of the first reviewer, we have revised the description of the flexible metric and the various penalty types in the *Methodology – Metrics* section to provide greater clarity on their relevance and practical usability. In this section, we also clarify that the specific values used in our experiments were selected for our particular use case, but users are free to choose different values to suit their needs. Additionally, in the *Conclusion* section, we acknowledge that the current system is limited to a predefined set of penalty types. However, we outline future work aimed at modularizing the flexible metric, enabling users to define and integrate custom penalty types more easily.

Question 3:

"- Considering GPT-4's widespread availability, relatively low cost, and improved performance compared to GPT-3.5, it would be great to see it included in the experiments, especially to test its capabilities in more challenging contexts."

Response:

In the initial version of the paper, we conducted experiments using GPT-4o, one of OpenAI's most advanced models. Following your suggestion, we also performed additional tests with GPT-4.1. These tests showed only minor differences in performance compared to the previously reported GPT-4o results. For this reason, we decided to retain the GPT-4o metrics in the paper to avoid re-running all experiments.

Regarding the discussion of model costs, we have removed the relevant section from the Results and Discussion, as the previous cost information is now outdated due to recent pricing changes by OpenAI. Currently, all models are relatively affordable and have comparable costs.

Question 4:

"- The experiments conducted feel a bit limited, which reduces the practical contribution of the work. It would be helpful to expand the ontologies to include more complex and realistic scenarios and test on a larger and more diverse dataset to better demonstrate the effectiveness of the flexible evaluation approach."

Response:

To evaluate our methodology on a more complex ontology, we drew inspiration from the work of Mihindukulasooriya [2], and adapted their DBpedia-WebNLG sports ontology and dataset into our format. This allowed us to test the models in a more challenging scenario. In this paper, we refer to the resulting dataset as the FootballSet. Due to the need for manual processing of each sample to ensure the inclusion of various types of extractable triples, we included only 75 examples, maintaining consistency with the size of the other two datasets.

The final results are presented in the Results and Discussion section, specifically in Tables 11 and 12. Overall, all models showed a decline in performance, which suggests that more complex ontologies pose greater challenges for these approaches. However, it is noteworthy that the metrics reported under the flexible evaluation paradigm were significantly higher than those under the strict evaluation. This indicates that while the models may struggle with adhering to a specific output format, they are nonetheless effective at extracting relevant knowledge.

[2] https://github.com/cenguix/Text2KGBench