Revisions for HopWeaver

We thank the reviewers for their feedback on our previous submission. We have thoroughly revised the manuscript to address the points raised and have incorporated significant additions to enhance the paper's clarity, depth, and empirical validation. The following is a summary of the key changes.

1. Expanded Appendix

We have significantly expanded the appendix to provide more detailed context, supplementary results, and deeper analysis, strengthening the empirical claims made in the paper. The main text has been updated with clear references to these new sections. The new appendices include:

- **Appendix A: Detailed Dataset Statistics:** We have added a comprehensive statistical comparison (**Table 6**) of the questions synthesized by HopWeaver against three widely-used, human-annotated benchmarks (HotpotQA, 2WikiMultiHopQA, and MuSiQue). This contextualizes our work and clarifies the experimental design for a fair comparison.
- **Appendix B.1: Large-Scale Validation (500 Samples):** To address potential concerns about sample size and to verify the statistical stability of our results, we conducted a large-scale validation with 500 samples. The results, presented in **Table 7**, closely mirror our main findings and demonstrate the reliability of our evaluation methodology.
- **Appendix B.2: Human Validation of LLM-as-Judge:** To substantiate the reliability of our LLM-as-judge framework, we performed a human validation study. The results (**Table 8**) show a high level of agreement between the LLM's assessment and human expert judgment, with the LLM's ranking aligning with the human consensus in **94% of cases**.
- **Appendix B.3: Manual Evaluation of Reasoning Paths:** To directly verify that our synthesized questions require authentic multi-hop reasoning, we manually evaluated 100 reasoning paths. The analysis, detailed in **Table 9**, confirms that **92%** of the generated paths are correct and logically sound, validating the structural integrity of our synthesis process.
- **Appendix B.4: Error Analysis of QA Failures:** We conducted a detailed case-by-case analysis of question-answering failures. This analysis reveals that the majority of errors stem from the inherent challenges of multi-hop reasoning for current LLMs (59.2% logical reasoning errors) or minor format variations (36.7%), rather than flaws in the synthesized questions themselves.

2. Correction of Minor Typo and Phrasing

We have carefully proofread the manuscript to correct typos.

3. Formatting and Presentation

In accordance with ACL formatting, all table captions have been moved to appear below their corresponding tables throughout the manuscript.

4. Inclusion of Recent Related Work

We have updated our literature review to include recent and relevant work. We now cite **Chen et al. (2024)**, a study on LLM-based multi-hop question answering synthesizing.

5. Refinement of Evaluation Terminology and Prompts

The terminology used for our evaluation dimensions has been refined to be more precise and professional. The prompts used in our framework have also been included and polished for clarity.

We are confident that these revisions have substantially improved the paper and addressed the key points from the previous review cycle.