Explanation Of Revisions

1. One reviewer raised a concern about potential latency introduced by Operator–Supervisor interactions. To address this, we measured per-query inference time for both the single-agent and dual-agent settings and confirmed that the observed difference is negligible. Results are reported in § 6.4 and detailed in Appendix K.

2. We conducted a thorough comparison of R2-KG's token consumption and corresponding API cost when using a mixed low-/high-capacity LLM setup versus a single high-capacity LLM. Our analysis demonstrates that the dual-agent configuration maintains strong performance while significantly reducing cost. Full details are provided in Appendix L.

3. Some reviewers perceived the iteration limit in R2-KG as a critical performance-determining factor. However, this is a misinterpretation—rather than directly affecting performance, the iteration limit is intended to serve as a reliability knob. A low iteration limit results in lower coverage and higher accuracy, while a high iteration limit yields high coverage and slightly lower accuracy. We have clarified this point by adding an explanation in §4.3.

4. To address concerns raised by some reviewers regarding the use of strong LLMs exclusively for R2-KG, we clarify that all baseline methods were evaluated under both low- and high-capacity LLM settings, including GPT-4o. To ensure this is clearly communicated, we have revised §5.2 and §6.1 to explicitly describe the LLM configurations used across all methods.