Explanation of Revisions

We sincerely thank all reviewers for their constructive feedback. Below we provide a summary of revisions and point-by-point response indicating how we have addressed each concern in our revised manuscript. The key points raised in the meta-review are addressed comprehensively in the following individual reviewer responses.

Summary of Major Revisions

Better Presentation: Comprehensive formatting revision, clearer explanations, and consolidated implementation details.
Stronger Motivation: Added concrete biomedical examples and clearer justification for technical contributions.
Enhanced Experimental Rigor: Added multiple runs with statistical validation, expanded RAG baseline comparisons (RAPTOR, KGP, MedRAG) (Section 4.3), and comprehensive case studies (Section 4.7).
Improved Generalizability: Demonstrated effectiveness across four domain-specific biomedical models (Section 4.4).
Methodological Details: Added computational efficiency analysis, acknowledged certain limitations, and provided detailed error analysis in case studies.

Reviewer xZcS:

Concern: No code/data release

Response: We are preparing a cleaned codebase for public release with the final publication, including complete implementation and documentation.

Concern: Limited to multiple-choice QA

Response: We have added free-form reasoning as an explicit limitation (Section 6).

Concern: Single base model, no alternative comparisons

Response: We have incorporated four domain-specific biomedical models (BioMistral-7B, Meditron3-8B, Llama3-Med42-8B, MMed-Llama-3-8B) (Section 4.4, Table 2) demonstrating CLAIMS' generalizability across different model architectures and training paradigms. We utilized the latest versions of PMC_LLAMA (MMed-Llama-3-8B) and meditron (Meditron3-8B).

Concern: No qualitative examples or error analysis

Response: We have added comprehensive case studies (Section 4.7, Appendix B) that include both success and failure examples with graph visualizations (Figures 4-7) and analysis of why methods succeed or fail.

Concern: No human evaluation is provided

Response: We have acknowledged this limitation (Section 6) and discussed the importance of expert validation while noting the logistical challenges of recruiting domain experts.

Concern: No discussion of computational efficiency provided

Response: We have added computational efficiency analysis for key components (claim extraction, triple extraction, graph construction, summarization) (Appendix A.6 Table 7) and discussed computational costs and trade-offs (Section 6).

Concern: Mixed sources for retrieval lack justification and ablation analysis

Response: We have provided better justification for our multi-source retrieval approach (Appendix A.1), explaining how each source contributes unique value. We did not conduct retrieval source ablation due to computational constraints.

Concern: No multiple runs, robustness testing, or statistical significance testing is provided

Response: We have conducted multiple runs for our QA experiments (5 runs with different answer shuffling) and report standard errors in all QA results tables (Tables 1-3), providing statistical validation for our key findings.

Concern: Reproducibility is limited by missing implementation details

Response: We have consolidated implementation details into a dedicated "Experimental Setup" section (Section 4.2) in the main paper, with additional model settings provided in Appendix C. We have specified decoding strategies (greedy decoding), answer evaluation procedures (JSON extraction with lm-format-enforcer), and moved key methodological details (model identification) to the main text.

Concern: The overall presentation lacks polish with formatting issues and poor structure

Response: We have conducted comprehensive formatting revision throughout the manuscript, fixed formatting issues including reference breaks using non-breaking spaces, improved appendix structure, corrected typographical errors, removed repetition in the methods, ensured consistent naming conventions, and ensured relevant appendix sections are properly referenced in the main text. We have also streamlined explanations and clarified the claims of interest identification process (Section 3.3) to better explain how test summaries are generated and relevance is computed.

Concern: The related work section would benefit from discussing prior research on semantic parsing into graph structures

Response: We have added additional references on graph construction approaches (Yang et al., 2025; Mo et al., 2025) while noting that existing work in our related work section (Edge et al., 2024; Wu et al., 2024) did use KG construction. Due to space constraints, we limited the expansion to these key additions.

Concern: Unclear how claims are handled during entity deduplication in graph construction

Response: We have clarified in the paper (Section 3.2) that multiple edges can exist between nodes, each representing a claim. During deduplication, we preserve unique claims as separate edges, allowing effective aggregation of related claims during summarization.

Reviewer wehe:

Concern: The choice of baselines is unclear given the references to other RAG methods in related work

Response: We have expanded our baseline comparison to include RAPTOR, KGP, and MedRAG (Section 4.3, Table 1), representing hierarchical summarization, dynamic knowledge graph generation, and direct retrieval approaches respectively. This provides more comprehensive evaluation against RAG methods.

Concern: Experimental details are missing from the main section

Response: We have created a dedicated "Experimental Setup" section (Section 4.2) in the main paper that includes model configurations, answer extraction protocols, and evaluation procedures, with additional detailed model settings provided in Appendix C "Model Settings", making the core experimental design clear.

Concern: There is a lack of error analysis showing what examples the model got incorrect and why

Response: We have added detailed case studies (Section 4.7, Appendix B) examining both successful predictions (showing effective cross-document reasoning) and failure cases (sparse graph construction, knowledge gaps), with specific examples and analysis.

Concern: The results from Component Level Analysis Tables appear unusual with scores close to 1, making it unclear whether conclusions can be drawn or if results fall within experimental variation

Response: We have provided additional analysis of component-level results (Section 4.6, Appendix H), offering better context for score interpretation.

Concern: The term "propositional claim" seems redundant since proposition and claim are traditionally synonyms

Response: We have moved the citation for propositional claims to the first mention of the term (Introduction). We use this terminology following Chen et al. (2024) who established this usage in the context of chunking strategies for retrieval systems.

Concern: The paper requires proofreading after the methodology section

Response: We have conducted thorough proofreading throughout the manuscript, standardized significant figures in all tables, and ensured consistent formatting.

Concern: Over-reliance on ArXiv papers over peer-reviewed publications

Response: We have rebalanced our citations to prioritize peer-reviewed publications where available, using ArXiv citations only when necessary.

Reviewer tbfu:

Concern: The scientific problem needs to be more clearly defined

Response: We have added a concrete biomedical reasoning example (Figure 1) in the Introduction showing how multi-document relationships are essential for biomedical QA, providing clearer motivation for our technical innovation.

Concern: The use of propositional claims for KG construction seems relatively engineering. The authors should be more specific about stating the novelty of this study

Response: We have enhanced our contribution statements to more clearly articulate our specific innovations: dynamic local knowledge graph construction with propositional claims without requiring prebuilt knowledge graphs, and layerwise topological graph summarization for LLM contexts. We maintain that our current specification sufficiently distinguishes our work from prior approaches.

Concern: In Table 5, why is the faithfulness score of CLAIMS lower than the semantic approach?

Response: We have added explanation in the component analysis results (Section 4.6) that the slightly lower faithfulness score results from multiple LLM calls in our pipeline, each introducing potential for hallucination. The small difference suggests this is an acceptable trade-off for improved reasoning capabilities.

Concern: Are there other methods to compare with CLAIMS rather than what appears to be mainly ablation studies?

Response: We have added comparisons with additional RAG baselines (RAPTOR, KGP, MedRAG) (Section 4.3, Table 1) and demonstrated generalizability across four domain-specific biomedical models (Section 4.4, Table 2), moving beyond ablation studies to comprehensive method comparison.

Concern: Is there a risk of data leakage during the retrieval and construction of the local knowledge graph?

Response: We have acknowledged data leakage as a limitation (Section 6), noting that this is a methodological challenge shared with other RAG approaches and discussing transparency about this limitation.