The reviewer recognizes that this manuscript, “Natural Language Querying of Biological Databases with Large Language Models,” presents a novel and timely contribution to the intersection of artificial intelligence, bioinformatics, and pharmaceutical data mining. The topic is highly relevant, given the rapid emergence of “AI co-scientist” paradigms and the growing need for accurate, interpretable natural language querying of complex biomedical databases. The manuscript’s systematic evaluation of multiple NL2QL (Natural Language to Query Language) strategies and its emphasis on multi-agent LLM frameworks mark it as an innovative and forward-looking effort.

However, while the conceptual framework and experimental design are sound, several methodological and statistical issues limit the paper’s reproducibility and interpretability. These concerns warrant a major revision before acceptance.

**Key Strengths**

Novelty and Timeliness:

The manuscript addresses an emerging challenge at the interface of AI and biomedicine: how to make biological databases accessible through natural language querying. It provides one of the first systematic evaluations of 21 NL2QL strategies using multiple LLM architectures, including agentic and retrieval-augmented models. Its forward-looking recommendations on FAIR AI agents and standardization of APIs are both relevant and impactful for the research community.

**Major Concerns: Statistical and Methodological Issues**

Lack of Statistical Detail and Replicability

Reported “success percentages” lack measures of variance (e.g., standard deviation, confidence intervals) or repeated-trial analysis.

Without error estimates, the reproducibility and robustness of the reported accuracies (e.g., 83–98%) are uncertain.

No Significance Testing

The paper does not apply any statistical tests (e.g., t-tests, ANOVA, or bootstrap resampling) to determine whether differences between strategies or models are significant.

Absence of Error Metrics

The evaluation could benefit from established performance metrics such as precision, recall, F1-score, and accuracy confidence intervals, which are standard in computational method comparison.

Inadequate Quantification of Task Avoidance and Hallucination

Although the authors note “task avoidance” and “background knowledge contamination,” these behaviors are not quantitatively measured or benchmarked, leaving the assessment largely qualitative.

Limited Baseline Comparison

No direct statistical comparison is made between LLM-generated and manually created (expert) Cypher queries, which would serve as a necessary performance baseline.

Unclear Replication Framework

It is not stated how many trials were conducted per question or per LLM strategy, which makes reproducibility evaluation difficult.

**Minor Suggestions for Improvement**

Figures over Tables for Visual Clarity:

Replace or complement key tables (e.g., Table 3 and the supplemental error distributions) with visual figures such as bar charts or heatmaps.

Visual summaries of model performance and error types will enhance accessibility and allow readers to identify trends more easily.

Condense Supplementary Data Summaries:

Include brief quantitative summaries of supplemental materials (e.g., Table 4) within the main text for better context.

Clarify Schema Complexity Impact:

Add a short discussion or visual correlation between schema complexity and model accuracy to support claims made in the Results section.

**Overall Assessment**

The paper represents an important and innovative contribution to AI-driven biomedical informatics. However, major revisions are necessary to improve statistical transparency, provide quantitative rigor, and enhance result visualization. Addressing these points will significantly strengthen the manuscript’s credibility and scientific impact.