Qeios

Peer Review

Review of: "A Comparative Study of Large Language Models in Explaining Intrinsically Disordered Proteins"

Hany Akeel Al-hussaniy[1]

1. University of Baghdad, Iraq

This manuscript addresses a timely and important question—whether large language models (LLMs) can effectively explain complex biochemical concepts such as intrinsically disordered proteins (IDPs)—and does so using a structured expert-based evaluation. The study is well-organized, clearly describes the prompt standardization process, employs blinded assessment, and uses appropriate statistical methods (ANOVA with Tukey HSD) to compare models. The inclusion of an internationally recognized IDP expert as an evaluator strengthens content validity, and the finding that GPT-4 outperformed other models is consistent with broader benchmarking literature. The work contributes meaningfully to the emerging field of AI-assisted scientific education.

However, several methodological and interpretative limitations should be addressed to strengthen the study. First, reliance on a single expert evaluator substantially limits generalizability and prevents assessment of inter-rater reliability; inclusion of multiple blinded experts and reporting of agreement statistics (e.g., ICC or Cohen's κ) would greatly enhance robustness. Second, the sample size (10 questions + 5 use cases) is relatively small for inferential statistics, and the Likert-scale data are ordinal; justification for parametric ANOVA (normality assumptions) should be clarified, or non-parametric alternatives considered. Third, the statistical power is not reported. Fourth, model versions are time-sensitive (June 2023 access), and LLMs evolve rapidly; the discussion should emphasize temporal reproducibility limitations. Fifth, the browsing models' lower performance may partly reflect interface or plugin instability rather than intrinsic architectural limitations; this distinction should be made more cautiously. Sixth, although the manuscript discusses hallucination risk and plagiarism concerns, it does not systematically evaluate citation accuracy or factual error rates—adding such an analysis would

doi.org/10.32388/VDJBBP

strengthen educational implications. Finally, more detailed reporting of prompt wording variability and response length normalization would improve methodological transparency.

Declarations

Potential competing interests: No potential competing interests to declare.