

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

Ely Salwana¹

¹ Universiti Kebangsaan Malaysia

Potential competing interests: No potential competing interests to declare.

Abstract: Overall, it is clear, well-organized, and highlights the importance of AI models in advancing education in complex fields. However, the abstract could benefit from briefly mentioning why browsing capabilities negatively impacted performance, as this is a key finding. Additionally, a brief explanation of the **practical implications** for educators could strengthen the relevance for readers interested in educational applications.

Materials and Methods: Overall, the methodology is well-developed, but enhancing the clarity and justification of specific choices would strengthen the narrative and make it more accessible to a broader audience. It might be helpful to list the evaluation criteria (Accuracy, Relevance, etc.) earlier in the section for context before discussing the Likert scale ratings in detail. The use of randomized, anonymized responses is a solid approach to minimizing bias, but the reasoning behind prompting the LLMs to "not mention limitations" should be explained. This may influence how transparent AI responses are and thus affect their evaluation. The use of a 5-point Likert scale and detailed criteria is well thought out, but it would be useful to mention how the scales were defined for consistency in scoring. For instance, how was "profound comprehension" of IDPs measured, especially in comparison between different models?

Overall: While the study references performance differences between models, it could benefit from incorporating specific statistical outcomes (e.g., p-values, effect sizes) more clearly in the discussion. For example, how significant were the differences between GPT-4 and GPT-4 with Browsing in numerical terms?

Overall, the text provides a solid empirical analysis of AI models' performance in IDP education and highlights valuable insights into the strengths and weaknesses of browsing-enabled LLMs. However, more focus on statistical findings, a deeper exploration of the technical aspects, and an expanded discussion on future research directions would enhance the overall rigor of the study. Additionally, incorporating ethical considerations and real-world applications would add practical value for educators and researchers looking to integrate AI into their teaching strategies.