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

Shokhan M. Al-Barzinji[1]

1 University of Anbar, Iraq

Potential competing interests: No potential competing interests to declare.

Evaluation: MAJOR REVISION

Dear Editor, please send the research for evaluation again after modifications are corrected by the author

Comments

This paper provides a well-structured and insightful evaluation of LLMs in explaining IDPs, with GPT-4 emerging as a clear leader. Its findings are significant for both AI and scientific education communities, highlighting the potential of AI to bridge gaps in niche domains like IDPs. Addressing the aforementioned limitations in future studies could further solidify the role of LLMs as transformative educational tools. This paper provides a fascinating exploration of the use of AI, particularly large language models (LLMs), in the domain of scientific education with a focus on intrinsically disordered proteins (IDPs). Below is a detailed review of the paper:

Strengths

1.The study addresses a timely and relevant topic—leveraging AI to advance education in the specialized domain of IDPs. The emphasis on IDPs is particularly valuable given their role in challenging traditional paradigms of protein structure-function relationships.

2. The paper meticulously evaluates four advanced LLMs (GPT-3.5, GPT-4, GPT-4 with Browsing, and Google Bard) using a rigorous framework:

Accuracy, relevance, depth of understanding, clarity, and overall quality.

Leveraging the expertise of Dr. Vladimir Uversky ensures a credible and authoritative evaluation.

3. The results strongly establish GPT-4 as the leading AI model in this context, reinforcing its effectiveness across various educational dimensions. The finding that GPT-4 consistently outperformed its counterparts provides actionable insights for future AI-driven educational tools.

4. The inclusion of a blinded evaluation process prevents bias, lending credibility to the findings.

The consistent format of engaging the LLMs and the use of descriptive statistics, ANOVA, and Tukey's HSD test enhance the robustness of the analysis.

5. The discussion section connects the findings to broader educational contexts, emphasizing the potential of GPT-4 and similar tools to revolutionize niche scientific education domains like IDPs.

6. The authors provide a nuanced discussion on the limitations of AI models with browsing capabilities, such as GPT-4 with Browsing and Google Bard, highlighting how these tools may paradoxically hinder performance in specific contexts.

Suggestions for improvement:

1. Incorporating evaluations from educators and students who directly use these models could yield more practical insights.
2. Delving into how LLMs handle subtopics within IDPs, such as specific diseases or mechanisms, could provide targeted recommendations.
3. A focused study on integrating LLMs into education while minimizing dependency and fostering critical thinking.
4. Benchmarking LLMs against human educators in IDP education would clarify their practical utility.