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

Arnaud Ferré[1]

1 University of Paris-Saclay

Potential competing interests: No potential competing interests to declare.

The article is clear and easy to read, providing valuable insights into the use of general LLMs to explore advanced knowledge in expert domains. However, the educational relevance of the study could be articulated more effectively. While it evaluates large language models (LLMs) in explaining intrinsically disordered proteins (IDPs), it does not establish the necessity of these “inaccurate, incorrect” AI-driven tools in educational contexts. Specifically, what is the intended educational context? Are learners expected to begin studying IDPs by conversing with a chatbot, or are you testing the alignment between your vision and the knowledge contained in the assistant? A more detailed discussion on how LLMs address gaps in understanding IDPs, which appears to be your main hypothesis, would enhance the paper's significance.

Indeed, in your introduction, you state: “*However, the broader scientific community has not fully embraced IDPs. Many researchers still adhering to the classical structure-function dogma and often times dismissing the importance and relevance of IDPs. Therefore, it is imperative to utilize available tools to the community that can help educate the scientific community about IDPs, their physiological importance, and their role in human diseases*”. This assertion lacks supporting arguments and may be perceived as overly generalized. It would benefit from specific examples and references to strengthen your claims. Additionally, the statement regarding researchers adhering to the classical dogma could be further nuanced, particularly in light of the ongoing research and increasing recognition of IDPs in various fields.

I also have concerns about the direction of the questions posed to the LLMs. They seem to encourage responses that support the authors' views on the importance of IDPs. For instance, asking “*What are some misconceptions about IDPs?*” strongly prompts the assistant to generate misconceptions, whether they exist or not. Conversely, when I asked ChatGPT (free GPT-4o-mini) a more open question, “*What are the most significant findings or breakthroughs in the study of protein structure and function?*”, the response did not include IDPs. However, asking, “*What are some misconceptions in the study of protein structure and function?*” seemed to be able to address IDPs, even without explicitly mentioning them. My point is that a learner may not share the same questions as those evaluated, as they might not possess the same knowledge and preconceived notions as the authors.

Additionally, I have concerns about the evaluation of the responses. A conversational assistant is likely to provide different answers to the same questions in a given context, and some could be hallucinated. Have you studied the variability of the responses to determine if there are completely different answers?

The exclusive focus on proprietary models raises concerns about reproducibility and accessibility. The lack of open-

source/open-weight models (e.g., Llama 3, Mistral models, …) limits the applicability of the findings, as proprietary tools often lack transparency and can hinder broader scientific engagement. Highlighting the advantages of open-source alternatives and addressing their relevance to educational settings would strengthen the study. Furthermore, be cautious not to adopt OpenAI's terminology as your own. OpenAI has not disclosed the precise data used for the pre-training of GPT-4. Your phrasing may suggest that the technical report you cite provides this list and may even imply that this report is a peer-reviewed scientific article. I recommend reformulating to explicitly present these excerpts as quotations from the cited report and clarify that it is a non-reproducible technical report from a private company.

Additionally, the Python code you used for statistical analysis and data visualization should be shared as open-source. Please explain in the article that you have supplementary data and specify what it includes.

Additional remarks:

- I understand the use of metonymy to refer to both the pre-trained LLM and the instruction-tuned model (+ RLHF), but perhaps you could clarify this for non-expert readers.
- The prompt model you used is defined in the body of the article, and I recommend separating it into a figure for clarity.
- While citing Dr. Uversky as an expert on IDPs is understandable, I would prefer more general phrasing in most of the article, such as “the expert on IDPs.”
- The explanation of what “use cases” are does not seem easily understandable.
- In the discussion, 'GPT-4.0' appears, while only 'GPT-4' has been referenced thus far. As there is now a ‘GPT-4o’ (mini or not) for most of the ChatGPT options, it is really ambiguous.
- In the discussion, the section “*We posit that the underlying architecture and extensive training data of GPT-4 [...]*” seems mainly more appropriate for the literature review.
- There appears to be no evaluation of the quality of the models (the ratings are relative), with only comparisons between models. For instance, are the responses of the best model good for the expert? Are the responses of the worst model poor? A comment on this aspect, even if informal, would enhance the understanding of the results.