**William James Bolton**[1]*, **Rafael Poyiadzi**[2]*, Edward Morrell[2]*, **Gabriela van Bergen Gonzalez-Bueno**[2]*, Lea Goetz[2]
*Imperial College London; [2]GSK, Artificial Intelligence and Machine Learning ; *equal contribution;* ***attending ICLR***

Creation Date: 08/05/24

# RAmBLA: A Framework for Evaluating the Reliability of LLMs as Assistants in the Biomedical Domain

## Abstract

Large Language Models (LLMs) increasingly support applications in a wide range of domains, some with potential high societal impact such as biomedicine, yet their reliability in realistic use cases is under-researched. In this work we introduce the Reliability AssessMent for Biomedical LLM Assistants (RAmBLA[1]) framework and evaluate whether four state-of-the-art foundation LLMs can serve as reliable assistants in the biomedical domain. We identify prompt robustness, high recall, and a lack of hallucinations as necessary criteria for this use case. We design shortform tasks and tasks requiring LLM freeform responses mimicking real-world user interactions. We evaluate LLM performance using semantic similarity with a ground truth response, through an evaluator LLM.

[1]https://github.com/GSK-AI/rambla

## Requirements for LLM Reliability

**Robustness to non-semantic variations:** LLMs should be robust to prompt variations that do not alter prompt meaning, and they should not display biases during few-shot prompting.

**High recall:** When operating on documents, LLMs should recall all relevant information, relying on either parametric knowledge or context exclusively, when instructed to do so.

**Hallucinations:** If they have insufficient knowledge or context information to answer a question, LLMs should refuse to answer.

## Results by Evaluation Task

**Robustness to non-semantic variations:** *QA paraphrase task; Few-shot prompt bias; Robustness to spelling mistakes*

- Larger models show superior performance to smaller models.
- Smaller models have larger bias to the label of the examples provided in the prompt.

**High recall:** *Recall from context vs knowledge; Recall from context with distraction*

- Models reliably use contextual knowledge when instructed to.
- Performance of smaller models is impacted more by distracting context.

**Hallucinations:** *Freeform QA baseline tasks; Conclusion generation; Question formation; "I don't know" task*

- Across all freeform tasks, larger models showed superior performance.
- Larger models successfully refrained from answering in every instance when insufficient context was provided, but smaller models occasionally provided hallucinated answers.

## Results Summary Table

| Task | Metric | GPT-4 | GPT-3.5 | Llama | Mistral |
|---|---|---|---|---|---|
| QA Baseline | F1↑ | 0.836 | **0.848** | 0.753 | 0.781 |
| QA Paraphrase | F1↑ | 0.819 | **0.836** | 0.728 | 0.780 |
| Few-shot Prompt Bias[1] | Bias↓ | **0.035** | 0.074 | 0.336 | 0.193 |
| Robustness to Spelling Mistakes[2] | F1↑ | 0.831 | **0.848** | 0.753 | 0.781 |
| Recall from Context vs Knowledge | F1↑ | **0.924** | 0.91 | 0.828 | 0.894 |
| Recall from Context with Distraction | F1↑ | **0.789** | 0.775 | 0.599 | 0.484 |
| Freeform QA Baseline | Acc↑ | **0.952** | 0.929 | 0.897 | 0.942 |
| Freeform QA Baseline (Bioasq) | Acc↑ | **0.948** | 0.939 | 0.921 | 0.943 |
| Conclusion Generation | F1↑ | **0.814** | 0.813 | 0.752 | 0.779 |
| Question Formation (Bioasq) | Acc↑ | **0.776** | 0.733 | 0.516 | 0.71 |
| "I don't know" Task[3] | Acc↑ | **1.0** | **1.0** | 0.62 | 0.872 |

[1] Bias shift to "yes" is defined as the proportion of excess "yes" answers in a balanced version of PMQA-L (PubMedQA Labelled), averaged across a sample of 4-shot example combinations.

[2] This set of results corresponds to 3 mutations.

[3] In this case the metric is the portion of times the model responds with "Unknown".

## Conclusions

- Worse performance in tasks requiring free text responses highlights the importance of benchmarking on real-world use-cases.
- LLMs themselves can act as evaluators in assessing semantic similarity with a ground truth.
- Our results suggest that in low-risk scenarios and with appropriate human oversight, LLMs can be a valuable resource in biomedical applications.
- Limitations: we did not conduct an exhaustive prompt search; we do not know the LLMs' training corpus; and our tasks require instruction-following, which may penalise smaller models.
- LLMs' high recall and robustness to prompt phrasing may allow them to support scientists in reviewing the biomedical literature. However, they are not ready for delegation in high-risk scenarios, such as applications impacting patients, because their outputs are difficult to verify even for biomedical domain experts.

William Bolton's work was done during an internship at GSK.ai, his PhD is supported by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. P/S023283/1).

GSK Imperial College London

Ahead Together