Evaluation of Large Language Models for an AI Chat Assistant Focused on Pumas and Pharmacometrics

Agastya Deepak Vinchhi ♠ Juan José González Oneto ♠ Vijay Ivaturi ♠

♠PumasAI

1 Objective

We analyzed and ranked a set of Large Language Models (LLMs) to create an AI assistant aimed at helping PumasAI's product users perform their work more efficiently. Focusing on chat completion models, we investigated the effectiveness of various LLMs in facilitating pharmacometric-related inquiries. This approach creates a standardized framework to benchmark and assess LLMs to find top-performing models. We included open-source and self-hosted solutions in our assessment, which allows the assistant to handle sensitive information and ensure data privacy.

2 Methodology

2.1 Exploration of LLMs

We chose a total of 26 LLMs based on a range of factors that consider the LLMs' abilities, performance, feasibility, and robustness. These factors are a guide for us to understand the capabilities and comprehensiveness of choosing LLMs, to explore as potential candidates for an AI assistant. Factors include: Model Size (Parameters), Implementation (access method through APIs or self-hosted LLMs), Licensing, Purpose, and Pre- training Data.

2.2 Establishment of a Prompt Repository

A collection of 145 prompts was assembled, serving as a standardized test for comparison. To collect these prompts, we conducted and recorded six interviews with scientists at PumasAI, asking them about their challenges using Julia and PumasAI tools and libraries. Each recorded interview transcript was manually parsed and passed through OpenAI's GPT-4-turbo to generate over 600 potential prompts that could be posed to an LLM. We narrowed down the total prompts to 145 based on diversity, specificity, and relevance to our research goal.

2.3 Benchmarking

We hosted the LLMs on a cloud computing platform to generate outputs based on our prompt repository with the use of Ollama - a library used to run LLMs locally. The following methods were used to evaluate the LLM outputs: Human Vetting, RAGAS, and LLM Rubric.

2.3.1 Human Vetting

Experts from PumasAI assessed LLM outputs to decide on whether they were suitable or not. Moreover, the standard of suitability by the human experts is based on criteria such as accuracy (output quality and correctness), relevance (question interpretation and understanding), and completeness (output covers the question) of the outputs. LLMs that were eliminated during the human vetting step were not considered for the rest of the evaluation. After human vetting, 7 models were eliminated, which resulted in our final selection of 19 models that would be evaluated.

2.3.2 Automated Model Evaluation using RAGAS (Retrieval Augmented Generation Automated Scoring):

RAGAS provides strong tools to automate the evaluation of Retrieval Augmented Generation (RAG) pipelines. RAGAS produces a quantitative score, providing us with a measurable metric of the chosen LLM's performance. We chose faithfulness and answer relevance from the various metric options that RAGAS provides.

- **Faithfulness:** This measures how closely the claims made in the LLM output align with the retrieved input context. A high faithfulness score implies that the LLM accurately reflects information from the retrieved context, which is critical in RAG applications.
- **Answer Relevance:** This measures how directly the LLM's output addresses the user's original question. It calculates a score by using the cosine similarity value between the vector embeddings of "n" artificially generated questions (generated by a separate LLM) and the original user question, based on the LLM output.

2.3.3 LLM Rubric - Grading Outputs with Assistance from Another LLM

We prompted an external LLM (OpenAI's GPT-4-turbo) to grade outputs (scale of 0-100) by providing a fixed rubric that explores factors that an output must have. These factors in our rubric have different weights, where some are deemed more important and hold higher scores than others. Factors include: Relevance (15%), Comprehensiveness (25%), Clarity and Coherence (20%), Depth and Detail (30%), and Citations (10%).

3 Results

Normalized results from the metrics described before are shown in the following table. Results are presented in descending order based on the total score, which can be found in the last column. The total score was computed as the average of all metric scores.

Model	Rubric	Faithfulness	Relevancy	Total
GPT 4o	0.91	0.91	0.87	0.90
Dolphin Mistral	0.83	0.91	0.94	0.90
Mistral	0.86	0.93	0.84	0.88
Claude 3 Haiku	0.89	0.9	0.83	0.87
Claude 3.5 Sonnet	0.92	0.88	0.77	0.86
Claude 3 Opus	0.86	0.86	0.73	0.82
Qwen 2	0.9	0.56	0.84	0.77
Starling-lm	0.88	0.46	0.94	0.76
GPT 4o Mini	0.76	0.78	0.73	0.76
Gemini 1.5 Flash	0.7	0.75	0.74	0.73
Llama 2	0.86	0.33	0.95	0.71
Zephyr 7	0.87	0.35	0.91	0.71
Phi 3	0.89	0.43	0.79	0.7
Vicuna 7b	0.79	0.33	0.93	0.68
Stable Code	0.73	0.36	0.92	0.67
Stablelm2	0.74	0.32	0.93	0.66
Gemma 2b	0.57	0.34	0.75	0.55
Gemini 1.5 Pro	0.58	0.55	0.51	0.55

For further analysis, we clustered the LLM outputs using the k-medoids algorithm. By first computing the vector embeddings of all LLM outputs, we calculated the pairwise distance matrix D using the cosine distance of vector embeddings. We determined that the ideal number of clusters would be $k = 2$. The plots below explore the relationship between the different metrics and are colored according to the cluster assignment.

This distance matrix D can be represented by

$$D_{ij} = \frac{1}{145} \sum_{k=1}^{145} \text{cosine\_distance}(v_{ki}, v_{kj})$$
$$i, j = 1, 2, \ldots, 19$$

where each entry is the average cosine distance value between the i-th and j-th vector embedding $v \; \forall \; k = 1, 2, \ldots, 145$, representing each prompt.

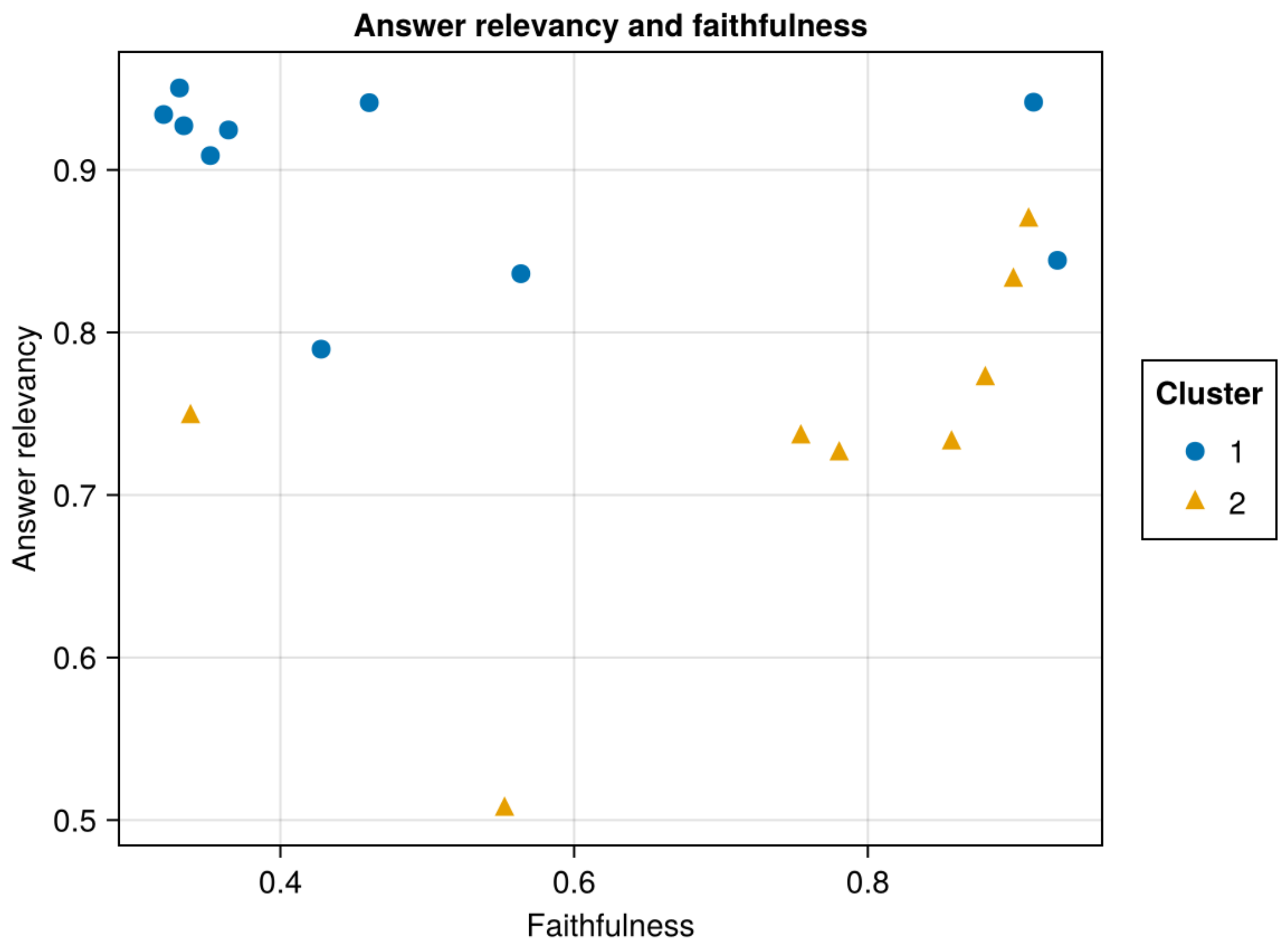

Figure 1: Answer relevancy and faithfulness, colored by cluster

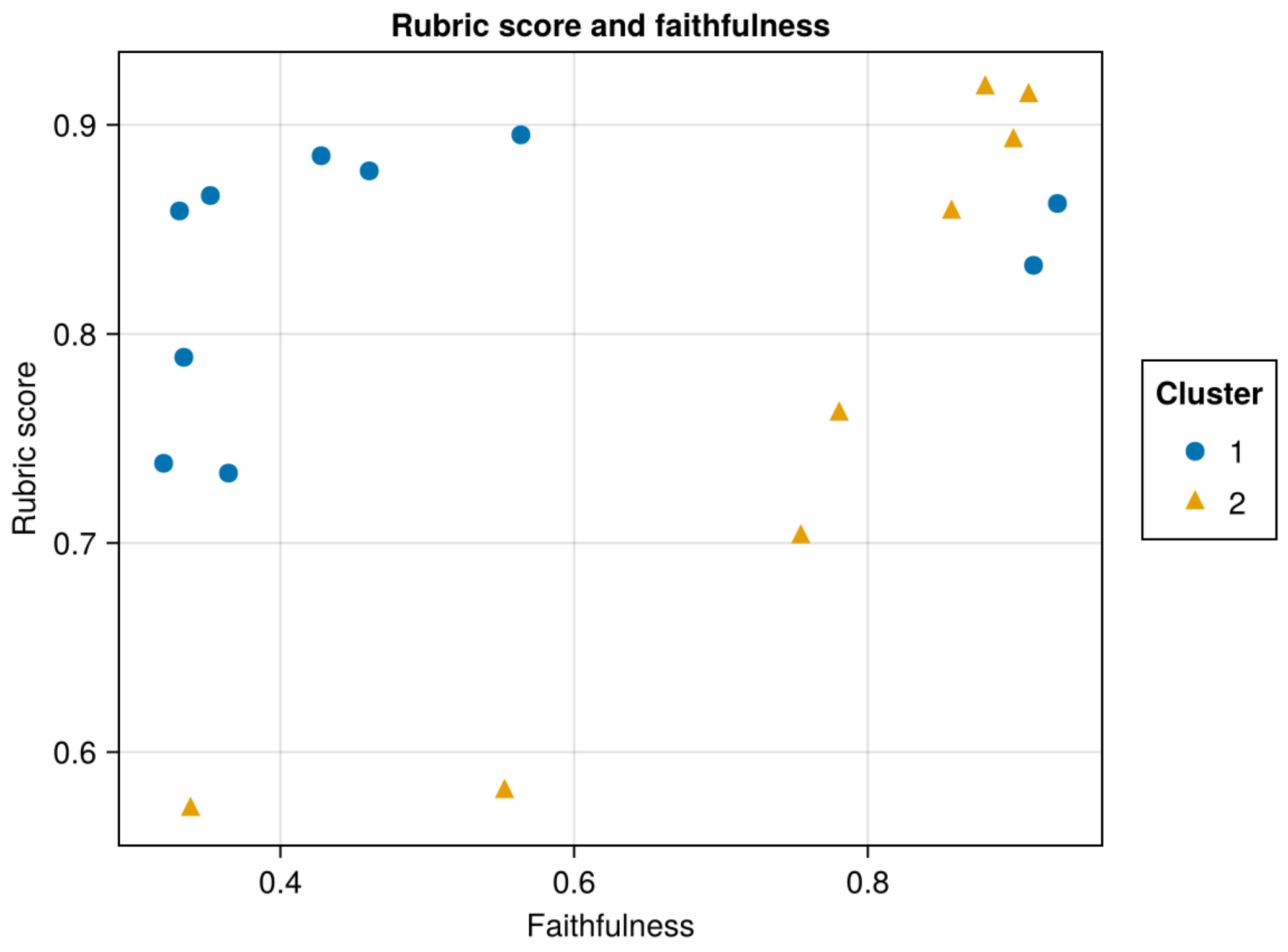

Figure 2: Rubric score and faithfulness, colored by cluster

4 Discussion and Conclusion

- Our evaluation reveals that the top 5 models for an AI assistant for Pumas and Pharmacometrics are GPT 4o, Dolphin Mistral, Mistral, Claude 3 Haiku, and Claude 3.5 Sonnet.
- To no surprise, we see GPT 4o, Mistral, and Claude models performed exceptionally well as they are popularly favored LLM choices. Notably, an AI assistant that implements Mistral models could handle sensitive information since it is open source and self-hosted, which partially mitigates privacy concerns.
- Our clustering could almost perfectly distinguish proprietary, API-based LLMs from self-hosted ones (with the only exception being Gemma 2b, which is related to the Gemini models).
- This research creates an evaluation method that allows the continuous exploration of an assessment of LLMs; this is important as the LLM landscape is constantly evolving.