Assessing Out-of-the-Box Bioinformatics Capabilities of Large Language Models

Varsha Rajesh, Geoffrey Siwo[1]

[1]Department of Learning Sciences, University of Michigan Medical School, Ann Arbor, MI, USA

Abstract

With immense improvements in artificial intelligence in proficiency and accessibility throughout the past few years, large language models (LLMs) have become a significant aspect of various fields, including programing, security, education, finance, and healthcare. Bioinformatics - the analysis of biological data - typically requires skills in the biological and computational sciences. The ability of LLMs to perform bioinformatics tasks posed as questions in natural language has not been extensively assessed. This study aims to determine the performances of OpenAI GPT 3.5, Llama 3 (70b), and GPT 4.0 in performing various levels of bioinformatic tasks. Utilizing Rosalind, an educational platform for bioinformatics, we compared the performance of the LLMs vs. humans on over 100 bioinformatics questions undertaken by 110 to 68,760 individuals. GPT-3.5 provided correct answers for 59/104 (~57.692%) questions, while Llama 3 (70b) and GPT 4.0 answered 49/104 (~47.115%) correctly. GPT3.5 was the best performing in most categories followed by Llama 3 and then GPT4. The best performing categories included DNA analysis and population dynamics, while the worst performing were sequence alignment/comparative genomics and motif finding/pattern matching. We found a linear correlation between the number of people who attempted a given question and the proportion who got it right (a potential measure of the difficulty of a question), especially in questions with fewer than 5000 human responses. The LLMs provided correct answers to several questions that require biological knowledge, statistical analysis and the use of computer code. These results highlight the potential and limitations for generalist LLMs to perform bioinformatics tasks and introduce its current capabilities in comparison to humans. This study will be a part of the larger effort to make bioinformatics more accessible, as with the help of artificial intelligence, expertise is not required for the calculations.

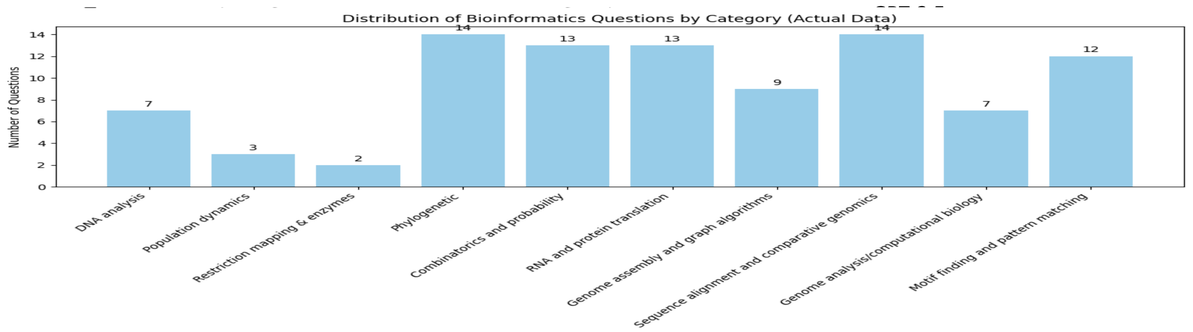
Introduction

Bioinformatics is a field that combines biology, computer science, and statistics, and it requires expertise in the field in order to perform various tasks used in the research process. If artificial intelligence (AI) was able to complete these tasks, the research process would be more efficient, accurate, and scalable. Artificial intelligence, although increasingly accurate, often lacks specific details and provides incomplete responses. This research aims to assess GPT-3.5, Llama 3 (70b), and GPT-4.0's performance in solving bioinformatics problems and compare it to human performance. A recent study published in *Scientific Reports* introduces an "optical model" to evaluate how effectively ChatGPT conveys genomics knowledge. Through three case studies—covering CRISPR-Cas systems, SARS-CoV-2 mutations, and cancer genomics—the study found that while ChatGPT was able to offer accurate high-level summaries, it often lacked the depth and precision required for expert understanding. Another study evaluates large language models on biomedical QA datasets, finding that although they can correctly answer many biomedical questions, they are prone to hallucinations, outdated references, and inconsistent performance on domain-specific queries. These previous applications of AI in bioinformatics highlight the importance of developing models that can accurately perform biostatistics tasks. This study contributes to the effort of facilitating biostatistical research with artificial intelligence, which can lead to more accurate data analysis and increase the discovery rate of biological observations.

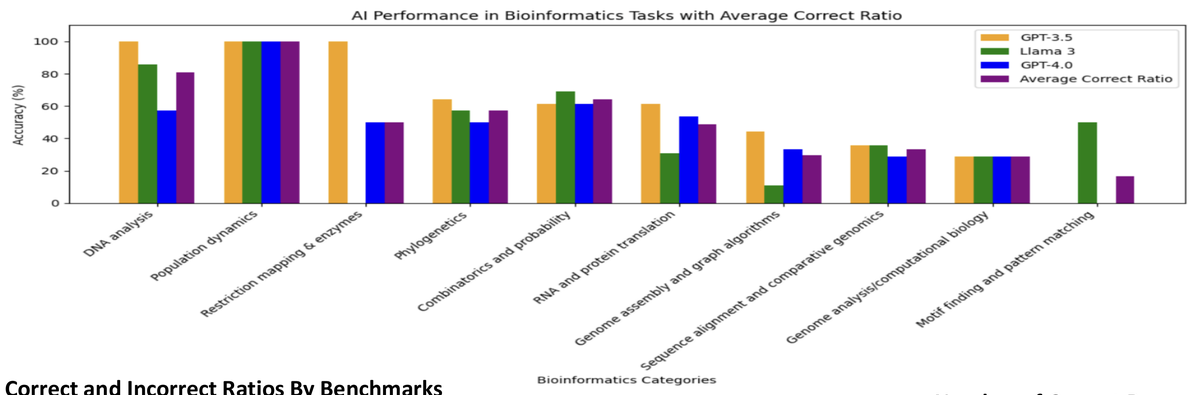
Results

This study evaluates the performance of GPT-3.5, Llama 3 (70b), and GPT-4.0 on 104 bioinformatics questions from Rosalind. GPT-3.5 answered 59/104 correctly (~57.7%), while Llama 3 and GPT-4.0 each answered 49/104 correctly. Human performance varied widely, with correctness rates ranging from ~20% to over 90%. A weak correlation was observed between human attempts and correctness, especially for questions with fewer than 5000 responses.GPT-3.5 had more incorrect than correct responses except for benchmarks >30% and >80%, while Llama 3 performed better except for >70 % and >90%. GPT-4.0 had more correct answers for >30% to >80%, with mixed results elsewhere. The ANOVA test yielded an F-statistic of 1.69 with a corresponding p-value of 0.186, indicating no statistically significant difference in performance among the models ($\alpha = 0.05$). The OLS regression analysis revealed negative coefficients for the models, with GPT-3.5 at -0.0435, Llama 3 (70b) at -0.0078, and GPT-4.0 at -0.0375, indicating inverse relationships between the predictors and model correctness. The 95% confidence intervals for correct responses were similar: GPT-3.5 (0.60, 0.67), Llama 3 (0.58, 0.66), and GPT-4.0 (0.59, 0.67). For incorrect responses, they were (0.54, 0.64), (0.57, 0.65), and (0.56, 0.64) respectively. 31 questions were answered correctly by all models, and 30 were answered incorrectly by all models. GPT3.5 and Llama 3 had the most overlap and Llama and GPT4.0 had the least. Top-performing categories included DNA analysis (GPT-3.5: 100%, Llama 3: 85.7%, GPT-4.0: 57.1%), Population dynamics (100% for all models), and Restriction mapping (GPT-3.5: 100%). Lower-performing areas included Genome assembly (GPT-3.5: 44.4%), Sequence alignment (GPT-3.5 & Llama 3: 35.7%), and Motif finding (GPT-3.5 & GPT-4.0: 0%). Linear regression showed slight negative correlations between AI and human correctness. While LLMs show promise in bioinformatics, they need refinement for specialized applications.

Distribution of Questions By Category

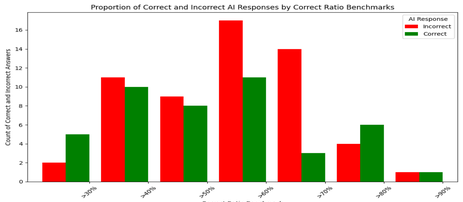


Performance of Models and Humans By Category

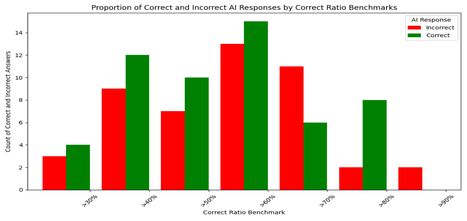


Correct and Incorrect Ratios By Benchmarks

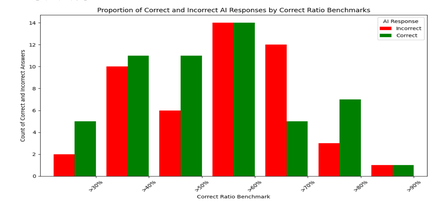
GPT3.5



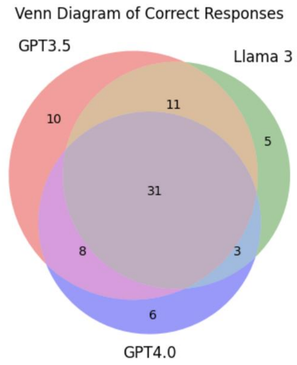
Llama 3(70b)



GPT4.0



Number of Correct Responses by Model



Research Methods

The study utilized data from over 100 bioinformatics questions on Rosalind, which had been attempted by 110 to 68,760 individuals per question. The performances of GPT-3.5, Llama 3 (70b), and GPT-4.0 were compared to human performance by evaluating the correctness of their answers against solutions provided by Rosalind. Performance was measured as the proportion of correct answers given by the models and humans. Difficulty of each question was estimated by the proportion of human users who answered it correctly. Additionally, the error was recorded. The tools used were OpenAI GPT-3.5, Llama 3 (70b), and GPT-4.0 to answer the bioinformatics questions. Python was used to analyze correlations between human performance and model performance. Google sheets/documents were used to keep track of observations. We assessed the performance of AI models compared to humans at different benchmarks (30%,40%,50%,60, 70%, 80%, 90%) defined by the proportion of humans who provided a correct answer to the questions. For example, at the 30% benchmark, we considered the performance of AI models only on the set of questions for which >30% of humans provided the correct answer. At each benchmark, the proportion of correct vs. incorrect answers for each AI model was obtained. To analyze the relationship between AI response accuracy and correctness, a binary variable was created for 'AI Response'. OLS regression was applied with the transformed response as the predictor and 'Correct Ratio' as the outcome variable, focusing on the coefficient to assess its impact on model performance. To evaluate whether model type had a statistically significant effect on accuracy, we performed a one-way Analysis of Variance (ANOVA) comparing the proportions of correct responses across three large language models: GPT-3.5, Llama 3 (70B), and GPT-4.0. Each model's responses were binarized as correct (1) or incorrect (0), and we found the F-statistic and p-value.

Discussion and Conclusion

This study explores the potential of the LLMs GPT-3.5, Llama 3 (70b), and GPT-4.0 in bioinformatics tasks. While these models correctly answered over half of the Rosalind platform's bioinformatics questions, they still fall short of human performance, particularly in biological sequence alignment and RNA translation. Weak correlations between human response and accuracy suggest that difficulty does not directly predict performance. LLMs excel in DNA analysis and population dynamics but struggle with genome assembly, phylogenetics, and sequence alignment. When the AI gets a question wrong, humans still answer correctly ~62-64% of the time on average. This suggests that the AI struggles with problems that majority of humans can solve, indicating potential gaps in reasoning or algorithmic application. The negative coefficients for AI response are the change in correct ratio when the AI response is correct in comparison to incorrect, underscoring that AI models tend to perform better on questions that are more difficult for humans. The model types do not significantly affect accuracy, suggesting similar capabilities overall. Despite limitations, LLMs demonstrate proficiency in biological knowledge, statistical analysis, and coding, showing promise for simplifying bioinformatics tasks. However, the tendency to give incomplete or incorrect answers points to gaps in training data. Future work should refine AI models for specialized tasks and explore AI integration into more complex bioinformatics tasks.

Works Cited

Author links open overlay panelIsrat Jahan a c, a, c, b, d, GPT-3.5, H., & AbstractRecently. (2024, February 20). *A comprehensive evaluation of large language models on benchmark biomedical text processing tasks*. Computers in Biology and Medicine. https://www.sciencedirect.com/science/article/pii/S0010482524002737

Shue, E., Liu, L., Li, B., Feng, Z., Li, X., & Hu, G. (2023, March 8). *Empowering beginners in bioinformatics with chatgpt*. bioRxiv : the preprint server for biology. https://pmc.ncbi.nlm.nih.gov/articles/PMC10028953/