BioinformaticsBench: A collaboratively built large language model benchmark for Bioinformatics reasoning

Abstract

Most of the existing Large Language Model (LLM) benchmarks on bioinformatics problem reasoning focus on problems grounded to niche research domains where datasets contain a small number of samples and, therefore are not truly representative of the broad domain of bioinformatics. To systematically examine the reasoning capabilities required for solving complex bioinformatics problems, we introduce an expansive benchmark suite BioinformaticsBench for LLMs. BioinformaticsBench contains a carefully curated dataset featuring a range of collegiate-level scientific problems from several bioinformatics domains, such as genetics, genomics, single celled analysis, proteomics, and metagenomics. Based on the dataset, we conduct an in-depth benchmarking study of representative open-source and proprietary LLMs with various prompting strategies. The results reveal that current LLMs are able to deliver a satisfactory performance, with an overall best score of 74%. Furthermore, through a detailed user study, we categorize the errors made by LLMs into ten problem-solving abilities. Our analysis indicates that while different models have different domains of expertise, GPT-4o is the best performing model overall. We envision that BioinformaticsBench will catalyze further developments in the reasoning abilities of LLMs, thereby ultimately contributing to scientific research and discovery.

1. Introduction

Bioinformatics is an interdisciplinary field at the nexus of biology, computer science, and statistics. In the past decade, there have been rapid advancements in high-throughput leading to a wealth of genomic, transcriptomic, and proteomic datasets, which has fueled a need for innovative computational approaches to extract insights from complex biological datasets. While dataset interpretation is central to understanding intricate biological phenomena1, the extraction of meaningful, novel insights presents a challenging endeavor due to their inherent complexity and volume of these datasets.

Recent advancements in artificial intelligence(Brown et al., 2020) and particularly in the development of large language models (LLMs), offer a promising direction in the pursuit of efficient and robust multimodal data interpretation in bioinformatics(Brandes et al., 2023; Livesey & Marsh, 2023; Al Ahdal et al., 2023). LLMs, such as GPT-4(OpenAI, 2023) (the fourth iteration of the Generative Pre-trained Transformer model (GPT-4), trained by OpenAI), have demonstrated remarkable capabilities in a wide array of natural language processing tasks(Yang et al., 2022). These models are built upon state-of-the-art deep learning architectures, trained on massive amounts of text data to simulate human conversations, and promise to generate coherent and contextually relevant responses across a range of scientific domains(Touvron et al., 2023). LLMs, with their ability to learn patterns, semantic relationships, and hidden structures in unstructured text data, offer a new perspective to assist bioinformatics research. They hold promise for tasks such as gene expression analysis, variant interpretation, protein folding prediction, and drug discovery(Elsborg & Salvatore, 2023).

The creation of benchmark sets is an active area of research in computer science, with several large scale benchmark sets created in the domains of Science(Wang et al., 2023), Mathematics(Lu et al., 2023), and Law(Guha et al., 2024). While several papers have been published on the use of LLMs in bioinformatics, there is very little work done on benchmark set creation. There are only 2 benchmarks that have been proposed till date, Bioinfo-Bench(Chen & Deng, 2023) and BioCoder(Tang et al., 2023), both of which have several drawbacks. Bioinfo-Bench only contains 200 questions and lacks a coverage of multiple sub-fields within the broad domain of bioinformatics. While BioCoder is more thorough, incorporating 1,026 Python functions and 1,243 Java methods extracted from GitHub, along with 253 examples from the Rosalind Project, all pertaining to bioinformatics, it is only limited to coding based problem solving. These drawbacks result in an incomplete assessment of the analytical and problem-solving skills required to tackle complex scientific problems. Therefore, it is essential to develop a new benchmark to evaluate the research progress of LLMs in solving bioinformatics problems.

2. Contribution

In this paper, we present BioinformaticsBench, which is to date the largest benchmark set containing 602 human annotated questions across 9 different bioinformatics domains.

Dataset selection criteria
Our dataset was selected to capture challenging problems within a diverse set of bioinformatics sub domains. These domains included bioinformatic algorithms, biostatistics, functional genomics, genetic linkage and equilibrium, mendelian genetics, molecular biology, phylogenetics, proteomics and sequence alignment. The questions were selected to ensure a balanced coverage across each of the domains. A breakdown of the category and question statistics is located in Table 1. To collect relevant questions, 10 textbooks were selected based on the textbooks reputability and wide use within the bioinformatic community. In addition to textbooks, 4 publicly accessible problem sets were collected from university websites, such as Massachusetts Institute of Technology (MIT) and University of Massachusetts, Boston (UMass Boston). A team of both graduate and undergraduate students were then tasked with consolidating questions from the chosen textbooks and problem sets to generate around 602 questions looking specifically for numeric, multiple choice, single word, and true or false questions. Numeric questions were required to have units attached. While parsing through questions from the given sources, if the question demanded a multi-word answer, the problem was reformatted to return a multiple choice, single answer or true/false solution, whenever feasible.

Data Preprocessing
Each of the four question types adhered to specific criteria as follows, aimed at mitigating potential ambiguities in the model's output:
1. Numeric: Each numeric question was modified to include the units in which the answer would be returned in (e.g. "Express your answer in KJ"), as well as a specified number of decimal places (e.g. "Round to one decimal place"). A standardized format for scientific notation was implemented when it was necessary for the question's specified output (e.g. "Write your answer in scientific notation. Format example: 1.00*(10^-1)").
2. Single Word: Since models often default to outputting a sentence or paragraphs to support the answer choice, questions requiring a single word answer were appended with the following statement: "Answer in one word".
3. Multiple Choice: Questions with the multiple choice format followed this structure: "Choose one of the following: (a) Choice 1 (b) Choice 2 (c) Choice 3". Each question included a set of answer selections, typically ranging from three to five choices.

Table 1: Summary of the BioinformaticsBench dataset. We report the total number of problems and percentage coverage for 9 key bioinformatics domains.

Title	Acronym	# Problems	Coverage
Bioinformatic Algorithms	algos	100	16.6%
Biostatistics	biostats	116	19.3%
Functional Genomics	funcgen	33	5.5%
Genetic Linkage & Equilibrium	genlink	33	5.5%
Mendelian Genetics	mendgen	115	19.1%
Molecular Biology	molbio	67	11.1%
Phylogenetics	phygen	27	4.5%
Proteomics	proteo	48	8.0%
Sequence Alignment	seqalign	63	10.5%

4. True/False: These questions were often presented as statements. For example, "True or False, the Burrows Wheeler Transform is a lossless compression algorithm". To mitigate an explanation, each statement incorporated the directive "Report your answer as True or False" for succinctness.
Additionally, each question was tagged with the corresponding citation, including page number or website address for a validity check. A team of 2 PhD students with high domain expertise, manually verified the correctness of the questions and their solutions, ensuring that the dataset was high accuracy.

3. Experiments

This section presents the experiments to assess the capabilities of LLMs in scientific problem-solving.

3.1. Experimental Setup

We evaluate our dataset on six unimodal LLMs,which include four proprietary models: GPT-3.5-Turbo (gpt-3.5-turbo-0125)(OpenAI, 2022), GPT-4-Turbo(gpt-4-turbo)(Achiam et al., 2023), GPT-4o (gpt-4o-2024-05-13)(openAI, 2024), along with four open-source models:LLaMA-2-7B (llama-2-7b-chat),(Touvron et al., 2023), LLaMA-3-8B(Meta-Llama-3-8B-Instruct)

We consider two different learning and prompting approaches described below:

1. **Zero-shot and few-shot learning.** In the zero-shot learning setting, models are not provided with any prior examples, which evaluates their inherent problem-solving capabilities with background knowledge and reasoning abilities. In the few shot setting, a few examples are given to the model before the test example. This aims to asses their capability to learn new information from the demonstrations and incorporate it into their problem solving process
2. **Prompting based approaches.** For our experiments, all settings begin with a system prompt that describes the type

and categories of questions. Additionally, we utilize a CoT prompting strategy in the zero setting. This was achieved by adding ”A: Let’s think step by step” at the end of the question prompt, and asking for the final answer.

Implementation details: We set temperature to default for all models to reduce the randomness of the predictions. After making an API call (GPT) or running a pretrained model (LLaMA3), our evaluation pipeline was used to parse the response received, and evaluate each questions’ correctness.

3.2. Results and Analysis

We report the model performance in terms of accuracy score for each textbook and an average score across all problems. The results of the LLMs in various settings are summarized in table 2. We have the following observations.

Observation 1: BioinformaticsBench is complex enough to differentiate among LLMs. Our results show that open-source models such as LLaMA were consistently outperformed by their proprietary counterparts across all settings within the textbook dataset. Notably, GPT-4o and GPT-4-Turbo lead in performance by a significant margin. For example, GPT-4o outperforms LLaMA-3 by 44% in the zero-shot setting. Additionally, within both LLaMA and GPT series, we observe a clear correlation between increased model capacity (i.e., larger parameter sizes) and improved performance. Intrestingly, while LLaMA performed poorly overall, it had an exceptionally high performance in the functional genomics domain, with a 21% performance gain compared to GPT-3.5 in the zero shot setting. Therefore, the complexity of BioinformaticsBench is able to differentiate the performance among different LLMs in different domains.

Observation 2. BioinformaticsBench highlights varied efficacy of prompting strategies across LLMs. Our findings suggest that the effectiveness of employing prompting strategies varies significantly among different LLMs. As shown in table 2, GPT-4-Turbo outperforms GPT-4o in becoming the top performing shows a marked improvement in the CoT setting over the zero-shot setting, with an average performance increasing from 67.6% to 72.8%. Interestingly, the performance overall increased in all categories except genetic linkage, for which the performance went down by around 9%. Meanwhile, despite the advanced capabilities of GPT-4o demonstrated by its zero-shot learning performance, it falls short compared to GPT-4o in chain of thought. This suggests a potential reduction in its program understanding capabilities. We also observed GPT-3.5-Turbo to exhibit a similar trend, with it’s performance falling by 5% in the CoT setting. Our findings illustrate that BioinformaticsBench can reveal the nuanced differences in the ability of LLMs to utilize prompting strategies effectively.

4. Evaluation Metrics

The accuracy was computed by comparing the response generated by the LLMs with the correct response. First, the text enclosed within the curly braces was extracted. For numerical quantities with values less than 1, a relative tolerance of 10% was allowed, while for quantities greater than 1, a relative tolerance of 5% was allowed. For string based responses, the strings were first converted to upper case and then matched. After extracting commas and text bracket, only exact string matches were considered. The accuracy was computed by dividing the number of correct responses by the total number of questions, per category and per dataset. Additionally, computational time was measured for each evaluation run using time.time().

5. Error analysis of prompting strategies

Considering the substantial advancements of current LLMs, an in-depth analysis of the particular skills that are either enhanced or limited under certain settings becomes imperative. Previous works have relied on human labor to annotate error reasons into different categories, which is both expensive and time-consuming (Zhong et al., 2023). In this section, we present an evaluation protocol that automates the classification of error reasons into deficient skills. This time-efficient approach enables large-scale analyses in future research. In order to quantify the impact of each setting on scientific problem-solving, we first define an essential skill set that is required by solving scientific problems. Then, an LLM verifier is employed to automatically classify each incorrectly solved problem based on the absence of a specific skill from the essential skill set. This approach generates error profiles, showcasing a direct comparison of different strategies. Firstly, we analyze the incorrect solutions made by GPT-3.5-Turbo for problems that provide detailed solutions. We chose GPT-3.5-Turbo since it had the highest error rate among the GPT models. Two PhD students who were highly familiar with the problems in our datasets, annotated the source of the error for each problem, indicating the specific line where the model makes a mistake and why. From 100 such error annotations, we distill these errors into ten essential skills that GPT-3.5-Turbo might lack:

1. Chain of thought: This ability involves decomposing the problem into smaller, manageable parts, understanding the relationships between these parts and maintaining a logical consistency between them.

2. Assumption identification: This skill involves the ability to recognize relevant and necessary assumptions in the problem.

3. Correct response, missing output: This describes the inability of the model to output the answer in the desired boxed format, despite correctly solving the question.

4. Incorrect response, missing output: This describes the

Table 2: Experimental results in terms of accuracy (%) on the textbook dataset. The best performing score is highlighted in **bold** and second-best is underlined. The average score is weighted by the number of problems in each textbook.

Model	Bioinformatics									
	ALGOS	BIOSTATS	FUNCGEN	GENLINK	MENDGEN	MOLBIO	PHYGEN	PROTEO	SEQALIGN	AVERAGE
Zero-Shot Learning										
LLaMA-2-7B-chat	26.00	8.62	42.24	6.06	17.39	26.86	33.33	20.83	14.28	19.60
LLaMA-3-8B-instruct	35.00	26.72	60.60	12.12	30.43	26.86	26.62	33.33	22.22	30.10
GPT-3.5-Turbo	35.00	47.41	39.39	12.12	42.61	31.34	44.44	41.67	33.33	38.20
GPT-4-Turbo	56.00	65.51	81.81	60.61	81.74	**65.67**	66.67	68.75	61.90	67.60
GPT-4o	**64.00**	**79.31**	**96.97**	**63.64**	**85.22**	61.19	**85.19**	**70.83**	**65.08**	**74.10**
Zero-Shot Learning + CoT Prompting										
GPT-3.5-Turbo	31.00	38.79	36.36	24.24	40.87	34.32	29.63	20.83	25.40	33.20
GPT-4-Turbo	**68.00**	**68.97**	**84.85**	**51.52**	**87.83**	**70.15**	**81.48**	**68.75**	**66.67**	**72.80**
GPT-4o	57.00	74.13	87.88	63.64	85.21	58.20	88.89	68.75	66.67	71.30

inability of the model to output the answer in the desired boxed format, while also incorrectly solving the question.
5. Problem deduction skills: This pertains to the ability to infer and deduce potential solutions or underlying principles from the given information in a problem.
6. Unit Conversion: This skill involves the ability to convert the answer to the unit specified in the question.
7. Domain Knowledge: This skill involves a comprehensive understanding of key scientific principles, terminology, and methodologies across bioinformatics subdomains.
8. Hallucination: This occurs when the model generates a response that is either factually incorrect, nonsensical, or disconnected from the input prompt.
9. Logical reasoning: This is the ability to make a reasoned argument and to identify fallacies or inconsistencies in an argument or set of data.
10. Arithmetic: This involves the ability to accurately carry out mathematical operations and computations.

After identifying this essential skill set, we assess the performance of the LLMs under different settings to discern the specific problem-solving skills they lack.

Can LLMs replace human domain experts as error evaluators?
Given the high cost of human annotations required to attribute the cause of incorrect solutions to specific skill deficiencies, we propose a novel self-critique protocol: we design a specific prompt that outlines these abilities, and employ a set of 3 LLMs to serve as classifiers and determine whether a specific error results from the lack of a particular problem-solving skill. Finally, we employ a consensus vote of the classification results to determine the class. This leads to the lack of human intervention while still maximizing the classification accuracy. To be specific, we utilize GPT-3.5-Turbo, GPT-4 and GPT-4 Turbo models as verifiers to determine the reason behind each error and pinpoint the missing skill. However, this approach resulted in only a 37.5% agreement among GPT-4o and GPT-4-Turbo models, and a 36% agreement among GPT-4o and GPT-3.5 models. This proves that while certain LLMs can be used to solve bioinformatics problems, error analysis still remains a challenge. (Figure 1).

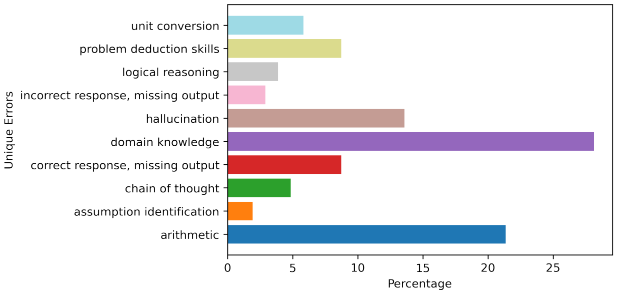

Figure 1: Error analysis via human annotation for GPT3.5. The total number of errors were distilled down into 10 categories for a set of 100 questions, and the percentages are plotted above.

6. Conclusion

This paper presents BioinformaticsBench, the largest to date benchmark containing more than 600 questions across 9 different bioinformatics sub domains. Our comprehensive evaluation includes a diverse arrays of Large Language Models (LLMs), spanning both open-source and proprietary models, and employing a variety of prompting strategies. We envision that the BioinformaticsBench dataset and evaluation protocol presented in this paper could lay a foundation for future research and enable advancements in understanding and enhancing problem-solving capabilities of LLMs.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Al Ahdal, A., Rakhra, M., Rajendran, R. R., Arslan, F., Khder, M. A., Patel, B., Rajagopal, B. R., Jain, R., et al. Monitoring cardiovascular problems in heart patients using machine learning. *Journal of healthcare engineering*, 2023, 2023.

Brandes, N., Goldman, G., Wang, C. H., Ye, C. J., and Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55 (9):1512–1522, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, Q. and Deng, C. Bioinfo-bench: A simple benchmark framework for llm bioinformatics skills evaluation. *bioRxiv*, pp. 2023–10, 2023.

Elsborg, J. and Salvatore, M. Using llm models and explainable ml to analyse biomarkers at single cell level for improved understanding of diseases. *bioRxiv*, pp. 2023–08, 2023.

Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Livesey, B. J. and Marsh, J. A. Advancing variant effect prediction using protein language models. *Nature Genetics*, 55(9):1426–1427, 2023.

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

openAI. `https://openai.com/index/hello-gpt-4o/`, 2024. [Accessed 24-05-2024].

OpenAI, R. Gpt-4 technical report. *ArXiv*, 2303, 2023.

OpenAI, T. Chatgpt: Optimizing language models for dialogue. openai, 2022.

Tang, X., Qian, B., Gao, R., Chen, J., Chen, X., and Gerstein, M. Biocoder: A benchmark for bioinformatics code generation with contextual pragmatic knowledge. *arXiv preprint arXiv:2308.16458*, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, X., Hu, Z., Lu, P., Zhu, Y., Zhang, J., Subramaniam, S., Loomba, A. R., Zhang, S., Sun, Y., and Wang, W. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

APPENDIX

Model Pricing Comparison

Table 3: Computational and pricing requirements for the different models. The model run time is measured in seconds for the zero shot setting and the pricing for input and output tokens are presented ($/1Mtokens).

Model	Run Time (sec)	Input	Output
GPT-3.5-Turbo	1784	0.50	1.50
GPT-4-Turbo	6012	10.00	30.00
GPT-4o	3377	5.00	15.00
LLaMA2-7B	3346	FREE	FREE
LLaMA3-8B	4048	FREE	FREE

For GPT models, pricing varies depending on usage and integration. On the other hand, LLaMA models, which are freely accessible, need to be install and deployed on a local machine in order to run.