ABSorBEnT: AI Benchmarking Suite for Biochemical Engineering Tasks

Tara Pande*
Department of Bioengineering, University of California—Berkeley
*tarapande@berkeley.edu

***Abstract*—Large Language Models (LLMs) have evolved from simple "chatbots", to AI assistants that are capable of understanding and communicating a variety of technical subjects. In particular, LLMs have the potential to aid researchers studying synthetic biology, molecular biology, organic chemistry, or biochemistry. However, it is still unclear what types of biochemistry tasks are best delegated to LLMs. I present ABSorBEnT: Ai Benchmarking Suite fOR Biochemical ENgineering Tasks. ABSorBEnT is a simple framework that aims to understand how LLMs reason accurately with synthetic biology concepts by testing prominent models ChatGPT-3.5 and Bard with questions from 3 different bioengineering domains and question difficulties. Amongst the many conclusions I've gathered, overall, the results show that Bard is 5% more accurate than ChatGPT. My findings also reveal that LLMs are best at performing "textbook look-up" questions, rather than generative application based questions. While these insights offer strong initial conclusions, a larger and more complete evaluation set would offer a compelling supplemental discussion. All of the evaluation data used in ABSorBEnT is publicly available at: github.com/TaraSPande/ABSorBEnT.**

I. Introduction

In the last few months, Large Language Models (LLMs) have shown promise in revolutionizing a variety of fields and careers. One such example has been in the life sciences industry. Not only have LLMs shown the remarkable ability to summarize and explain dense scientific texts or concepts, but many researchers also desire a better understanding of the depth of knowledge that LLMs possess across various fields in molecular biology, organic chemistry, computational biology, or bioengineering.

I aim to uncover the performances of various LLMs through my benchmarking suite called ABSorBEnT (Ai Benchmarking Suite fOR Biochemical ENgineering Tasks). In addition to summarizing scientific texts, a crucial application of LLMs in the life sciences would be to generate custom correct answers in a simpler manner than complex software and more powerful than a generic search engine. LLM's abilities in this department is what ABSorBEnT aims to determine.

LLMs are typically evaluated in a variety of ways, from something simple like mathematically comparing question-answer word embeddings using cosine similarity or using RAGAS scores like factual accuracy, answer relevancy, retrieval context precision, and context recall [1].

The goals with these approaches, while generally looking for correct answers, also prioritizes word semantics and clarity. However, it is already well understood that various LLMs like OpenAI's ChatGPT [2] can "organize or summarize text", as it is a main selling point of LLMs [3]. For these reasons, benchmarking semantic clarity is not a priority for me. In my initial literature review, it became apparent that creating an LLM benchmark for life science applications is not a well address topic. Therefore, my strategies to create my own benchmark were fairly open ended.

To be clear: my goal is to determine the depth of accurate knowledge that LLMs have. Therefore, asking open-ended questions (ex. "How would I go about…" or "What would we include in a…"), while great for evaluating semantic clarity, are problematic when evaluating brute factual accuracy because LLMs can give vague answers that are not technically wrong, but are certainly not the most helpful, specific, or knowledgeable. Because semantics and conversational qualities are less important for this application, I designed a data set with straightforward questions that require conclusive, factual answers.

II. Methods

Because I am creating a biochem-based benchmarking suite, I wanted to test LLMs' accuracy when forced to give specific answers and judge them on a binary [correct] or [incorrect] grading criteria. Because this binary grading framework may critique harshly, I gave the LLMs three attempts for each question to see if LLMs are capable of improving when given simple feedback on incorrect responses.

Following a similar methodology to Chen 2023 [4], I created a database of 120 questions that are subdivided into 3 different categories: Biology, Chemistry, and Coding. Coding questions have specific applications in Bioengineering.

TABLE I
A BREAKDOWN OF THE LLM EVALUATION DATA SET.

	Biology	Chemistry	Coding	Total
Easy	18	0	3	21
Medium	9	12	0	21
Hard	23	26	30	78
Total	50	37	33	120

The questions are also categorized by their difficulty: Easy, Medium, or Hard. Easy and Medium questions are designed to be questions that could be easily found in a textbook. Easy questions could be found in middle school textbooks, whereas Medium questions could be found in college textbooks. Hard

questions, on the other hand, require custom generation or applying concepts to specific prompts, rather than just stating back textbook content. This difficulty was assigned to match the stereotypical strengths and weaknesses of LLMs [3].

I generated these 120 questions by hand, given the specificity of my goals, taking inspiration from content covered in lecture materials from UC Berkeley Professor John C. Anderson's course: Genetic Design Automation [5]. I tested these 120 questions on two LLM models: ChatGPT-3.5-turbo [6] and Bard [7]. Then, I standardized their outputs using Python and RDKit [8] in order to calculate their respective accuracies.

III. Results

To state the most anticipated finding: Bard performs better than ChatGPT, but mildly so (by 5%). With three attempts, Bard has an accuracy of 47.50% overall for biology, chemistry, and coding questions. Meanwhile, ChatGPT only performs with 42.50% accuracy overall across the 3 domains.

Interestingly, across three attempts, Bard and ChatGPT seem to improve by a similar rate: 7.50% for Bard and 6.70% for ChatGPT. Qualitatively, Bard seems more "self-aware", in the sense that it was dramatically more likely to give a response explaining that it doesn't know how to solve the problem or asks for feedback more regularly after repeated failed attempts. Over 95% of "I don't know" responses were from Bard. On the other hand, ChatGPT was more likely to repeat back the same answer to the user, despite being told it's wrong. ChatGPT is a more "confident" engine.

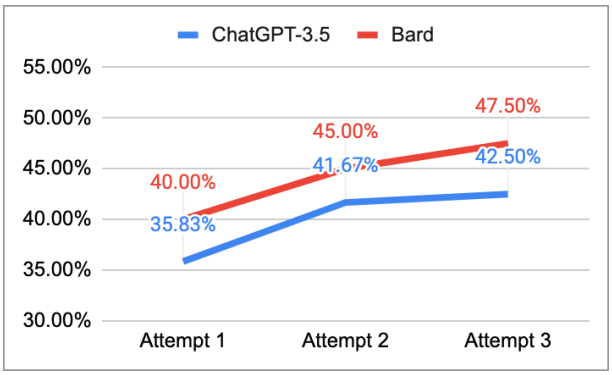

Fig. 1. This figure shows that while Bard mildly performs better than ChatGPT overall, both models do not dramatically improve in performance when given multiple attempts to find the correct answer.

Additionally, I wanted to compare the performance of ChatGPT and Bard across the different difficulties and domains of questions. As expected, ChatGPT and Bard performed worse as the difficulty of the questions increased, across all three domains. This result is not surprising, as the difficulties of these questions were assigned with this hypothesis in mind.

Another notable finding is that Bard outperforms ChatGPT on Medium Chemistry questions by 44.28%. One example of a Medium Chemistry question would be: to count all of the atoms in a SMILES, a notation for encoding molecular structures computationally. Additionally, Bard also performed better on Hard Coding questions compared to ChatGPT by 36.38%. The two models were generally comparable in the other categories.

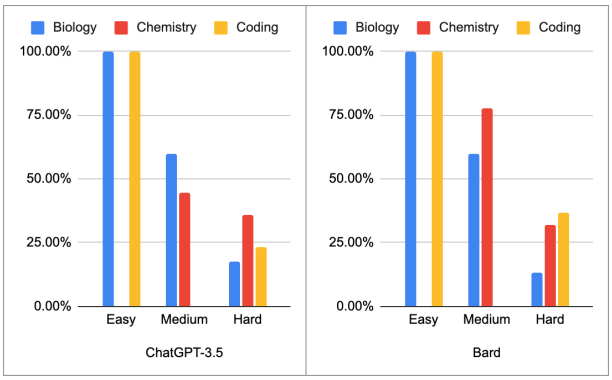

Fig. 2. This figure shows the accuracy of ChatGPT (left) and the accuracy of Bard (right). As expected both LLMs performed worse as the questions increased in difficulty, across all 3 domains.

Lastly, it is important to dissect the differences between Bard and ChatGPT on specific question topics. My 120 questions cover 16 different topics, with 5 to 16 questions per topic.

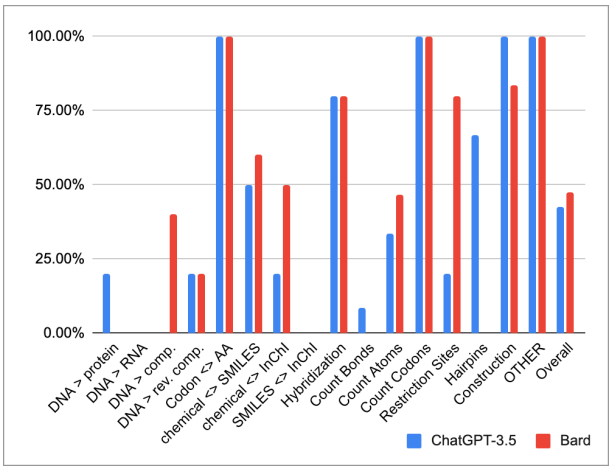

Fig. 3. There are 16 different topics of questions in my entire 120 question data set. Here is a more-detailed breakdown of the topics that *can* and *can't* be accomplished by LLMs.

One interesting finding is that ChatGPT outperforms Bard significantly on questions related to hairpins. Although, I only asked 3 hairpin related questions, so this result may be inaccurate with such a small sample.

Furthermore, Bard has a much stronger understanding of the cleavage location of restriction sites found in DNA. But once again, I only asked 5 restriction site questions, so more sampling is necessary to verify this inital finding.

Lastly, and most importantly, both ChatGPT and Bard are very bad at understanding samples of DNA, RNA, or amino acid chains longer than 10 characters. Oftentimes, these

LLMs would get close, but would delete bases or completely hallucinate a string of characters that didn't exist which both cause inaccurate answers.

In summary, while these results are an interesting initial conclusion, given more time, I would greatly want to increase the sample size of my data set in order to verify these results with higher confidence.

IV. Discussion

This research project was successful in drawing initial conclusions that benchmark the performances of two models: ChatGPT-3.5 and Bard. I created a data set of 120 questions that span 3 different domains: Biology, Chemistry, and Coding, as well as 3 difficulties: easy, medium, and hard.

Without a doubt, the greatest limiting factor throughout this project is time. In the future, with more time, it will be great to test more LLMs with ABSorBEnT, such as Llama-7b [9], Galactica-30B [10], or ChatGPT-4.0 [2]. Additionally, there are many open-source, finetuned LLM models that specialize in various bioengineering domains, such as BioGPT [11], GenePT [12], scGPT [13], GeneFormer [14], scBERT [15], and many more! It would be very interesting to pin these specialized finetuned models against more generalized models and compare their performances with this benchmark suite.

Furthermore, in the future, it would become very important to increase the size of my data set, as well as increasing the number of topics that my data set covers. I already have several ideas of types of questions to incorporate into the ABSorBEnT benchmark suite later on.

As a much later goal, I would love to automate my benchmark suite, particularly when it comes to data set generation. Potentially, it would also be great to create an open-source, easy-use GUI that can allow other scientist to contribute to the benchmark suite. This collaborative effort would facilitate the ability for researchers to learn how LLMs perform against biochemical engineering prompts.

Data Set Availability

My entire 120 question data set, answers from all 3 attempts for ChatGPT and Bard which includes data tables, graphs, qualitative notes, and python code are all available on Github: github.com/TaraSPande/ABSorBEnT.

Acknowledgments

I would like to acknowledge my professor John C. Anderson for his mentorship and assistance.

References

[1] "RAGAS Metrics", *RAGAS*, docs.ragas.io/en/latest/concepts/metrics/index.html

[2] "ChatGPT 3.5: How can I help you today?", *OpenAI*, chat.openai.com.

[3] Michael Schade "How ChatGPT and Our Language Models Are Developed", *OpenAI*, 2023, help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

[4] Chen and Deng, "BIOINFO-BENCH: A Simple Benchmark Framework for LLM Bioinformatics Skills Evaluation", *bioRxiv Preprint*, October 2023. doi.org/10.1101/2023.10.18.563023.

[5] "Chris Anderson, Associate Professor, Bioengineering", *Berkeley Bioengineering*, https://bioeng.berkeley.edu/faculty/chris_anderson.

[6] "Text generation models Documentation", *OpenAI API*, platform.openai.com/docs/guides/text-generation.

[7] "Bard: A conversational AI tool by Google", *Bard Experiment*, bard.google.com.

[8] "The RDKit 2023.09.3 documentation", www.rdkit.org/docs/source/rdkit.Chem.inchi.html

[9] "Introducing LLaMA: A foundational, 65-billion-parameter large language model", *Meta Research*, 24 February 2023, ai.meta.com/blog/large-language-model-llama-meta-ai/

[10] R. Taylor, M. Kardas, G. Cucurull, et. al., "Galactica: A Large Language Model for Science", *Meta AI*, 16 November 2022, arxiv.org/pdf/2211.09085

[11] R. Luo, L. Sun, Y. Xia, et. al., "BioGPT: generative pre-trained transformer for biomedical text generation and mining", *Briefing in Bioinformatics*, vol. 23, no. 6, November 2022, doi.org/10.1093/bib/bbac409

[12] Y. Chen, J. Zhou, "GENEPT: A SIMPLE BUT HARD-TO-BEAT FOUNDATION MODEL FOR GENES AND CELLS BUILT FROM CHATGPT", *Cold Spring Harbor Laboratory*, 19 October 2023, doi.org/10.1101/2023.10.16.562533

[13] H. Cui, C. Wang, H. Maan, et. al., "scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI", *Cold Spring Harbor Laboratory*, 2 July 2023, doi.org/10.1101/2023.04.30.538439

[14] Z. Cui, Y. Liao, T. Xu, et. al., "GeneFormer: Learned Gene Compression using Transformer-based Context Modeling", *Cornell University arXiv*, 31 January 2023, doi.org/10.48550/arXiv.2212.08379

[15] F. Yang, W. Wang, F. Wang, et. al., "scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data", *Nature Machine Intelligence*, vol. 4, pp. 852-866, 26 September 2022, doi.org/10.1038/s42256-022-00534-z