

WIP: Comparison of Large Language Models for Applied Mathematics Questions in Engineering Courses

Carlos Merlos
Department of Computer Science
Clarkson University,
Potsdam, NY 13699
merloscl@clarkson.edu

Faraz Hussain, Swati Kar, Lavanya Shri,
Olaoluwayimika Olugbenle, Mahesh Banavar
Department of ECE
Clarkson University, Potsdam NY 13699
{fhussain, mbanavar}@clarkson.edu

Abd AlRahman AlMomani
Mathematics Department
Embry Riddle Aeronautical University,
Prescott, AZ
almomana@erau.edu

Abstract—In this Innovative Practice WIP paper, we present a comparison study of large language models (LLMs) to see which would respond best to questions posed in classes. While the ultimate goal of the project is to develop classroom chatbots using LLMs, the first step, presented here, is to evaluate different models for multiple topic areas and select a few of them for further development.

In this preliminary work, we cover two subject areas: signal processing and differential equations. We take slightly different approaches to each area so we can get a better understanding of the range of capabilities of the LLMs. For signal processing, we use custom open-source LLMs and the free version of OpenAI’s ChatGPT 3.5. All models we use here are free. However, significant coding and processing power is required to implement the models in this method. On the other hand, for differential equations, we compare the paid version of ChatGPT 4.0 and compare that with a custom GPT, also from the paid version of ChatGPT. In this case, there is no requirement for coding skills. However, a monthly fee is required.

We evaluated the LLMs by testing them on question batteries of different difficulty levels. We found that all the models function well when the questions are simple and drawn directly from the source material. However, as the difficulty level of the questions increases, the “better” models in terms of training and parameters perform better, making a case for better training, but both ways of training come with costs. The results from this evaluation will be presented at the conference.

Future work involves further development of these models into more interactive chatbots and their deployment in classrooms for preliminary evaluation.

Keywords—undergraduate, large language models, chatbots, performance comparison, signals and systems, differential equations

I. INTRODUCTION

LLMs, including GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers), and LLaMa (Large Language Model Meta AI) are advanced AI models that can be trained to generate human-like text [1]. They offer instant feedback and interact with users in a conversational manner, which can be

leveraged for interactive learning and increasing student access and opportunities.

```
> Question:
(t) time-invariant? Is the system described by  $y(t) = \sqrt{t}(t)^2$  time-invariant?

> Answer:
Based on the context provided, the system described by  $y(t) = \sqrt{t}(t)^2$ 
(t) is not time-invariant.

A system is considered time-invariant if its behavior does not change with respect to time. In other words, if the input and output of the system are shifted in time, the output will also be shifted by the same amount.

In this case, the system described by  $y(t) = \sqrt{t}(t)^2$  is not time-invariant because if we shift the input  $x(t)$  by some constant time  $t_0$ , the output  $y(t)$  will also be shifted by the same amount, resulting in a different output.

Therefore, the answer to the question "Is the system described by  $y(t) = \sqrt{t}(t)^2$  time-invariant?" is no.
```

Figure 1. Example output from a custom-trained LLaMa-2 large language model. The correct answer is “No”, and we can see the model included a justification for the response.

The ultimate goal of this project is to develop a “classroom assistant” chatbot. This chatbot will be able to provide students with answers to simple questions, so that (a) students interact with a chatbot that they find less intimidating than interacting with course staff; (b) students have access to this resource 24x7; and (c) instructor and Teaching Assistant time when interacting with students is spent answering more complex questions.

The first step towards building this chatbot, covered in this Work-in-Progress paper, is to evaluate the suitability of different LLMs to be effective chatbots. We consider (1) OpenAI’s pre-trained LLM, ChatGPT 3.5 with no additional training provided [16]; (2) a non-customized version of OpenAI’s ChatGPT 4 [14]; (3) a customized version of ChatGPT 4 [13]; and (4) a custom-trained open source LLM, namely Meta’s LLaMa-2 [15] (see Figure 1).

To test the suitability of these different models, we use both quantitative and qualitative methods. For the quantitative evaluation, we ask these different LLMs (1) short questions in both multiple-choice and fill-in-the-blank formats; and (2) open-ended problems and questions that require some analysis. We evaluate the outputs obtained and consider the

accuracy of the output and the cost in terms of fees to access the models and computational resources/effort needed. Care was taken to train with equivalent data as far as possible. From this step, a smaller number of LLMs will be shortlisted for further evaluation, as described below.

To assess suitability as a “classroom assistant” chatbot, we will have students interact with the shortlisted LLMs in the Fall semester of 2024. We will conduct this exercise in two locations, both small private engineering universities in rural settings. One university will use material in signal processing and the other will use material in applied differential equations. We will use surveys to get feedback on student experiences with this tool. Student experience in both locations will also be compared.

The rest of this paper is organized as follows. An overview of large language models is provided in Section II. The use of LLMs for course topics and our training and evaluation methods are presented in Section III. Preliminary results from the benchmarking (cost, time, and accuracy) are shown here. Concluding remarks are in Section IV and the results from student interactions will be presented at the conference.

II. LARGE LANGUAGE MODELS

A. Large Language Models

LLMs are sophisticated artificial intelligence systems designed to understand and generate human-like text. These models are at the forefront of natural language processing (NLP) research and have achieved remarkable results in various language-related tasks such as text generation, translation, summarization, and question-answering. In large part, this is due to the fact that LLMs are trained on massive datasets of text, code, and other mediums, often exceeding terabytes in size [1].

At the core of LLMs lies a transformer architecture, which facilitates the processing of input sequences and the generation of output sequences [1]. The model consists of multiple layers of self-attention mechanisms and feed-forward neural networks, enabling it to capture contextual information and generate text word-by-word. During the pre-training phase, the model learns to predict the next word in a sequence given the preceding context, which helps it develop a deep understanding of language semantics and syntax. The network uncovers the statistical relationships between words in the massive datasets used for training. This allows the prediction of the next words in sequences with remarkable accuracy, making it ideal for translating languages while preserving meaning and nuance, or answering questions comprehensively and in informative ways.

Current work in the field includes few-shot learning techniques, which enable LLMs to adapt to new tasks with minimal additional training data, through a process known as fine-tuning [2] and the validation and verification of LLM outputs [3].

B. Advantages and Disadvantages

LLMs offer several advantages, revolutionizing natural language processing (NLP) tasks and enabling novel applications across diverse domains. These models excel in tasks such as text generation, translation, summarization, and sentiment analysis, thanks to their ability to capture complex language patterns and semantics. LLMs are able to be adjusted for specific tasks with minimal additional training data. Resources such as OpenAI’s ChatGPT [14], Google’s Gemini [17], and Meta’s Llama models [4] make LLMs widely accessible.

Despite their many advantages, LLMs also come with several limitations and challenges. One notable concern is the potential for bias and ethical issues in generated text. LLMs trained on large corpora of internet data may inadvertently perpetuate stereotypes or generate offensive content [21]. Addressing bias in LLMs requires careful curation of training data and the development of mitigation strategies to ensure fair and unbiased text generation [5, 6]. LLMs also can exhibit a phenomenon known as hallucination, where they generate output that appears coherent at first glance but lacks factual grounding [21]. This mainly stems from limitations in the training data or the inherent challenge of translating statistical patterns into truth. Mitigating this type of risk requires users to be critical of LLM outputs and cross references with verifiable sources for confirmation. Another challenge associated with LLMs is their computational resource requirements. Training and fine-tuning large-scale models like GPT-4 demand substantial computational power and energy consumption and have large latency [7].

C. Pre-Training LLMs

Pre-training an LLM involves a colossal dataset of text and code, often exceeding terabytes in size. This data encompasses books, articles, web text, code repositories, etc. The model learns to predict the next word in a sequence given the preceding context, using techniques like self-attention mechanisms and transformer architectures, using massive amounts of unlabeled text data [1]. Pre-training enables the LLM to develop a general understanding of language semantics and syntax, which can then be fine-tuned for specific tasks [8].

An important aspect of pre-training is unsupervised learning, which empowers an LLM with the ability to uncover patterns and relationships on its own and to learn the fundamentals of language structure and statistical relationships between words. An obvious advantage is that less data preparation is required. Massive unlabeled datasets that are readily available allow for broader language understanding. This method requires large computational resources and careful design of pre-training objectives. Additionally, learning can be less targeted but requires fine-tuning for specific tasks.

D. Fine-Tuning LLM Models

Fine-tuning is accomplished by supervised learning which requires labeled examples, where each piece of text has a

corresponding label indicating its function or meaning [9]. Fine-tuning typically requires less data and computation than pre-training, provides precise control over model behavior, and provides high performance on well-defined tasks.

One method for fine-tuning is Retrieval-augmented generation (RAG), which combines elements of traditional language model-based generation with retrieval-based methods [10]. In RAG, a retriever component is used to retrieve relevant context or information from a large knowledge base when presented with a query, which is then incorporated into the generation process of the language model [11, 12].

III. LLMs FOR COURSE MATERIAL

In this paper, we present our preliminary work in comparing the performance of multiple generative AI LLMs for two content areas: signal processing and differential equations.

A. Signal Processing

For this area, we used four LLMs, two commercially available LLMs, OpenAI’s ChatGPT 3.5 and Google’s Gemini, and two instances of Meta’s LLaMa 2 with 13 billion (13B) parameters that were fine-tuned on local machines. We queried ChatGPT and Gemini via their respective vendor web user interface. To differentiate between the two LLaMa models, we name them LLAMA-Notes and LLAMA-Textbooks. Both utilized semantic search with RAG [10]. LLAMA-Notes utilized class notes, quizzes, exams, lecture transcripts, and a Signals and Systems textbook, all formatted in PDF. LLAMA-Textbooks used over a dozen open-sourced books, both from signals and systems and prerequisite courses, all in PDF format.

Questions were sampled from former quizzes and exams. The questions were split into two groups based on difficulty levels. Questions in the first group were simpler and tested recall. Questions in the second group were more open-ended and required some analysis. Questions with mathematical symbols were converted to an inline text notation, pseudo LaTeX e.g. $e^{j\omega} + e^{-j\omega}$, so they could be queried by the four LLMs that we used. Some questions referenced supplementary images. Only Gemini was able to accept images along with text prompts. Gemini was also able to refer to online references and usually referenced relevant websites.

Overall, ChatGPT 3.5 and Gemini answered most questions correctly, but the discussion associated with each response varied in length and accuracy. There were instances, similar to real-life student responses, where the answer was correct, but the explanation was not. Gemini by default seemed to be less formal and more conversational, although these settings can be modified. LLaMa-Notes was the worst performer. We can explain these results due to the fact that LLaMa-Textbooks was augmented with multiple textbooks from this course as well as prerequisite courses. Naturally, there is some information that is lost when filtering academic settings, i.e. some chapters are briefly covered, or not at all. Selected questions and the performance comparison of the

four models are presented in Table I. We also saw some interesting results from LLaMa-Notes, which sometimes

Table I. Question topics, and comparison of responses from four LLMs. Each LLM responded with an answer and a discussion. Fully correct responses are marked with a green box, fully incorrect answers with a red box, and partially correct responses (either the answer or the discussion is correct, and the other is not).

Question Topic	ChatGPT 3.5	Gemini	LLaMa-Textbooks	LLaMa-Notes
Linear Algebra				
Calculus				
Impulse Function				
Trig properties				
Optimum values				
Periodicity				
Fourier computation				

referenced class notes in interesting ways. For example, in response to one question, the LLM responded with the following: “According to the context, Mahesh K. Banavar states that the Fourier series is not continuous in frequency for a square wave. However, for a triangle wave, the Fourier series is continuous in frequency, as shown by Mahesh K. Banavar’s example using the first five terms of the Fourier representation.” While this discussion is incorrect, it is interesting to see an effort for conversation and an effort to summarize the notes.

1. Limitations: We would like to note here that the quality of the training and supplementary data has a role in the quality of the outputs. Our implementation ignored handwritten notes overlaid in the course material, as well as images presented in lecture slides. Having example questions and text-based answers would aid in semantic searches. Where available, we used lecture transcripts but additional data was lost, such as supplementary whiteboard notes, lecturer gestures, and tonal emphasis, as well as questions raised by students. Additionally, some audio was lost when the presenter or students were away from the source microphones.

Running LLMs on local machines often requires significant computational resources. Cloud-based LLMs, such as Gemini and ChatGPT, alleviate the need for local hardware.

They do however introduce some latency as queries must pass through a network. Local instances can have lower latency, but can rarely compete with the computing power and speed of commercial cloud-based LLMs. Specific to our work, we note that ChatGPT and Gemini benefit from the vast computational power of data centers, enabling them to handle larger models and more complex tasks efficiently. They also have the added benefit of learning from millions of prompts generated by hundreds of thousands of users.

Conversely, one benefit of executing a local LLM is improved privacy. When using web-based LLMs, prompts entered are sent to remote servers. Local processing offers greater control over user data and avoids potential privacy risks associated with online implementations.

B. Differential Equations

This research examines the effectiveness of Large Language Models (LLMs), specifically two flavors of ChatGPT-4, in enhancing the learning experience for students in differential equations. Our research is centered on two versions of ChatGPT: the standard model and a fine-tuned variant specifically tailored to handle differential equations, referred to as “DiffEq” (see Figure 2).

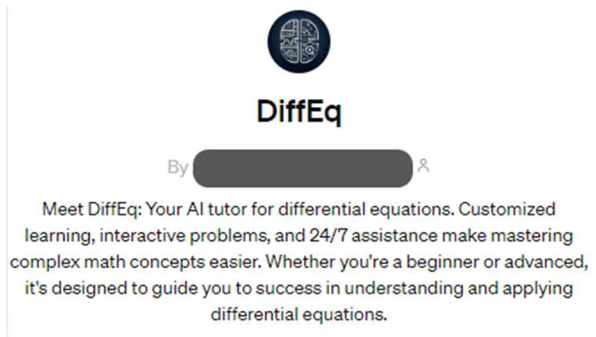


Figure 2. Our customized differential equation GPT (“DiffEq”). ChatGPT-4 was fine-tuned with training data specific to differential equations to develop this custom LLM. The model was created by co-author AlMomani.

1. **Fine-Tuning Process:** The fine-tuning of the “DiffEq” model was meticulously designed to align closely with the specific needs of a differential equations course. The process involved curating a diverse dataset comprising online resources, textbooks, and lecture notes pertinent to differential equations [18-20]. This targeted training approach aimed to enhance the model's proficiency in not only recognizing specific types of equations but also in applying systematic methods to solve them, thereby ensuring it could function as an effective class chatbot.

2. **Evaluation Methodology:** Our evaluation framework was structured in two distinct phases to assess both the general and fine-tuned models:

- a) Phase 1: Multiple-choice questions were used to gauge the models for immediate comprehension and their ability to identify key concepts in differential equations

accurately. A set of 26 questions was selected from sources external to the training data to avoid bias. This phase served as a preliminary measure of each model's basic understanding and recall abilities. The general ChatGPT-4 model achieved an accuracy of 85%, while the DiffEq model demonstrated superior performance with an accuracy of 96%.

- b) Phase 2: The second phase involved more complex, open-ended problems requiring detailed, step-by-step solutions. This phase was crucial in assessing the models' ability to engage in deeper reasoning and explanatory processes. The problems selected tested various capabilities, including handling initial value problems (IVPs) and algebraic simplifications within differential equations. For example, in solving the initial value problem, $y'' - 4y' + 4y = 0, y(0) = 12, y'(0) = -3$, both models performed well under varying initial conditions, although “DiffEq” tended to provide more concise and targeted explanations.

A particularly illuminating comparison arose with the problem, $2x(dy/dx) - y = 3$, which tested algebraic manipulation skills. Here, DiffEq outperformed the general ChatGPT-4 model, which struggled with some of the more intricate algebraic simplifications, underscoring the effectiveness of the fine-tuning process.

These results highlight the significant benefits of fine-tuning LLMs like ChatGPT for specific academic disciplines. The enhanced performance of the DiffEq model in both the multiple-choice and problem-solving tests illustrates its potential as a valuable tool for educational support in differential equations courses. The fine-tuning process, by focusing on relevant content and problem-solving methodologies, equips the model with a deeper understanding and an improved ability to navigate this task.

These results not only reaffirm the capabilities of LLMs in educational settings but also open avenues for further exploration into subject-specific adaptations of AI technologies, aiming to create more effective, personalized learning aids for students across various academic disciplines.

IV. CONCLUSIONS AND FUTURE WORK

In this work, we compared six different LLMs in two topic areas. For signal processing, we used four models: two were commercially available (ChatGPT and Gemini) and two were custom-trained versions of Meta's LLaMa-2 13B models. LLMs were trained on different datasets: solely textbooks vs primarily class notes. Overall, ChatGPT and Gemini performed well but justifications varied. Gemini offered more conversational explanations. LLaMa-Notes performed the worst, likely due to less comprehensive information, and sometimes referenced notes in unexpected ways, showing potential for discussion-based learning.

In the fall semester, we will shortlist and further develop LLMs, and have students interact with them. Student feedback will be presented at the conference and will guide our work in developing the course assistant chatbots.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [2] I. Schlag, S. Sukhbaatar, A. Celikyilmaz, W.-t. Yih, J. Weston, J. Schmidhuber, and X. Li, “Large language model programs,” *arXiv preprint arXiv:2305.05364*, 2023.
- [3] T. Gao, H. Yen, J. Yu, and D. Chen, “Enabling large language models to generate text with citations,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.
- [4] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du et al., “Lamda: Language models for dialog applications,” *arXiv preprint arXiv:2201.08239*, 2022.
- [5] L. Rinaldi, E. S. Ruzzetti, D. Venditti, D. Onorati, and F. M. Zanzotto, “A trip towards fairness: Bias and debiasing in large language models,” *arXiv preprint arXiv:2305.13862*, 2023.
- [6] B. H. Zhang, B. Lemoine, and M. Mitchell, “Mitigating unwanted biases with adversarial learning,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.
- [7] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.
- [8] J. D. M.-W. C. Kenton and L. K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT, 2019*, pp. 4171–4186.
- [9] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 328–339.
- [10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [11] M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith, “Knowledge enhanced contextual word representations,” in *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.
- [12] K. Vladimir, O. Barlas, M. Sewon, L. Patrick, L. Wu, E. Sergey, C. Danqi, and Y. Wen-Tau, “Dense passage retrieval for open-domain question answering,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 6769–6781.
- [13] F. Kieser, P. Wulff, J. Kuhn, and S. Kückemann, “Educational data augmentation in physics education research using chatgpt,” *Physical Review Physics Education Research*, vol. 19, no. 2, p. 020150, 2023.
- [14] J. Kocón, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniec, M. Gruza, A. Janz, K. Kanclerz et al., “Chatgpt: Jack of all trades, master of none,” *Information Fusion*, vol. 99, p. 101861, 2023.
- [15] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [16] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, “A brief overview of chatgpt: The history, status quo and potential future development,” *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023.
- [17] H. R. Saeidnia, “Welcome to the gemini era: Google deepmind and the information industry,” *Library Hi Tech News*, no. ahead-of-print, 2023.
- [18] P. Dawkins. (2023). Paul's Online Notes: Differential Equations. [Online]. Available: <https://tutorial.math.lamar.edu/classes/de/de.aspx>
- [19] J. Chasnov. (2016). Introduction to Differential Equations: Lecture notes for MATH 2351/2352. [PDF document]. Available: <https://www.ms.uky.edu/~ejwh226/Spring2018/Chasnov.pdf>
- [20] C. Tracy. (2017). Lectures on Differential Equations. [PDF document]. Available online at: <https://www.math.ucdavis.edu/~tracy/courses/math22B/2BBook.pdf>.
- [21] S. Alshahrani, E. Wali, J. Matthews, “Learning From Arabic Corpora But Not Always From Arabic Speakers: A Case Study of the Arabic Wikipedia Editions,” *WANLP*, 2022.