Building a More Accurate Artificial Intelligence Chatbot for Engineering Problem Solving in the Water Industry

Michael Demko, Rob Sinclair, and Kevin Hou

The integration of generative artificial intelligence (AI), exemplified by ChatGPT, in engineering is already impacting communications within the water sector—but how has it affected accuracy? ChatGPT, as a language model, demonstrates the capability to generate human-like responses and assist with engineering questions; however, its accuracy is contingent upon learned patterns from training data, lacking true understanding or real-time information.

In 2023 this team demonstrated that three leading chatbots proved to be extremely inaccurate (<10 percent) when prompted with 100 water industry engineering problems. The team decided to design its own chatbot to respond to the Principles and Practice of Engineering (PE) exam-style questions with greater accuracy. Two main challenges for a large language model (LLM) in engineering problem solving are the need for complex mathematical reasoning and the ability to work with text and image inputs. As a result, the team built the chatbot on GPT-4o, one of OpenAI's latest multimodal foundational models.

Experiments were performed on a practice exam consisting of 76 multiple-choice questions and results were analyzed, both quantitatively and qualitatively, to assess the chatbot's overall performance and its specific skill sets. The baseline experiment took each question and its figure simultaneously to generate an answer. The team then implemented other techniques to improve the chatbot's performance depending on its observed deficiencies. This chatbot has reached an initial mean accuracy of 72.4 percent (a passing score) across three trials on the full practice exam.

Advancements in the Field of Language Models

One of the latest advancements in the field of LLMs is the emergence of multimodal LLMs. While the traditional transformer architecture was designed to process text, multimodal LLMs take inputs in several forms, such as texts and images. This invention represents a huge step forward in deep learning as it blurs the boundary between vision models and language models, enabling neural networks to gain insights from higher dimensional data. Such a capability has a wide range of applications, from visual question answering to content creation, and it has been applied in many fields, such as healthcare and marketing, where processing complex data across multiple modalities is essential.

Engineering issues, however, bring their own challenges for these new models in two respects. First, engineering problem solving requires complex reasoning across modalities. For example, a typical problem in the water industry can involve a question and a drawing where the question refers to specific elements in the drawing. To solve such a problem, an engineer must draw insights across two modalities—text and image—and perform complex reasoning to arrive at the correct conclusion. To be a useful AI assistant in the water industry, a multimodal LLM must be able to do the same. Second, engineers in the water industry make safety-critical decisions all the time. An AI assistant's limitations must be thoroughly and objectively tested, otherwise its advice cannot be relied on. Since multimodal LLM is still a novel tool, its capabilities in engineering problem solving are largely untested.

In 2023 the team tested three popular chatbots with 100 engineering exam-style questions; that test did not provide the multiple-choice responses. The intent was to determine how accurate the chatbots were in real-world scenarios that are not multiple choice. Answers were considered correct if they were within 5 percent of the correct answer; GPT-4 performed the best at 11 percent.

Literature

Language models have been around for a long time. They are formally defined as probabilistic models that predict the likelihood of a sequence of words or tokens, and these models can be readily used for generative tasks through sampling. Rule-based models, such as n-grams (Shannon, 1948), which are sequences of items (usually words) from a given text or speech, had dominated the field for decades until models based on neural networks appeared. Variants of recurrent neural networks became the preferred architecture for language models since they were able to capture long-range dependencies to some extent (Hochreiter, 1997; Rumelhart, Hinton, and Williams, 1986). These models, however, relied on a set of hidden states that were iteratively updated over time. Inevitably, they struggled to handle very-long-range dependencies in natural language and did not support parallelized computations at training time.

Michael Demko, P.E., is senior project manager at Wade Trim in Palm Bay. Rob Sinclair is an advanced design technology practice lead at Wade Trim in Beaver, Penn. Kevin Hou is an AI engineer at Wade Trim in Pittsburgh.

In 2017, the transformer architecture emerged and quickly dominated the field of natural language processing (Vaswani, 2017). This architecture replaced sequential computations with the attention mechanism, making parallel computation possible at training time. Two years later Lu, Batra, Parikh, and Lee (2019) introduced ViLBERT, the world's first multimodal model that took visual and text input simultaneously to draw insights across both modalities.

Since then, many works have investigated the performance of multimodal models (Achiam et al., 2023). Beyond their ability to process text, researchers have also evaluated their capabilities to see (Yang et al., 2023), to reason (Ahn et al., 2024), and to retrieve relevant information with techniques such as retrieval augment generation, or RAG (Lewis et al., 2020). In the context of engineering problem solving, Pursnani et al. (2023) investigated the performance of GPT-4 on the U.S. Fundamentals of Engineering exam with textural input and the model ultimately scored 75.37 percent on four-option multiple-choice questions; however, the model was tested with unimodal input where figures were described with text. This work extends Pursnani et al. (2023) to evaluate the ability of state-of-the-art multimodal models to solve engineering problems from bimodal inputs consisting of image and text.

Methods

All experiments (Table 1) are performed

Continued on page 10

Continued from page 8

on GPT-4o (OpenAI, 2024). Except for experiment 5, the model's input exclusively consists of three components:

- The original question
- The figure that comes with the question (if any)
- A short prompt that instructs the model to solve the problem

Zero-shot prompting is used everywhere. The data set is a PE exam where each question is passed into the model in an individual call. In total, five experiments are performed, where experiment 1 serves as the baseline. These experiments are designed to test the impact of the RAG and a variety of prompting techniques on performance.

Specifically, the prompts used in the five experiments are shown in Table 2. In this table, text enclosed in double quotes is a string, while italic text in parenthesis is a variable. A Python script, a file that generally contains a short self-contained set of instructions, i.e., lines of code, that performs a specific task, automates all steps in the experiments, from importing data to parsing results.

Experiment 5 implements RAG with the PE Civil Handbook and the PE Environmental Handbook (NCEES, n.d.). This corpus is divided into chunks based on the following separators in order:

- Double new lines
- New line
- Spaces
- Characters

These splits are done recursively until each chunk falls below a chunk size of 300 characters; adjacent chunks overlap by 30 characters. A unimodal vector database with an embedding size of 1536 is constructed with text-embedding-3-small (OpenAI, n.d.), the smallest embedding model provided by OpenAI. At query time, a semantic search is performed on this embedding space with L2 loss, which is used to measure model performance by calculating the deviation of a model's predictions from the correct predictions to identify the five most relevant chunks to use as the context out of 4545 chunks stored in the vector database.

Table 1. Experiments and Prompting Techniques

Experiment	Prompting Technique
Experiment 1 (baseline)	Include reasoning steps
Experiment 2	Handle figure description in a separate LLM call
Experiment 3	Suppress reasoning steps
Experiment 4	Include reasoning steps and describe figures
Experiment 5	Retrieval augmented generation with PE Civil Handbook and PE Environmental Handbook

Table 2. Experiments and Prompts

Experiment	Prompt
Experiment 1 (baseline)	“Your job is to answer multiple-choice questions from the PE exam, civil - water resources, and environmental discipline. First reason through the question step by step, and then provide your choice as A, B, C, or D in the format 'Answer: [LETTER]' at the end. The following is the question: *(exam question)*.” + *(drawing/table) if applicable*
Experiment 2	“Your job is to answer multiple-choice questions from the PE exam, civil - water resources, and environmental discipline. First reason through the question step by step, and then provide your choice as A, B, C, or D in the format 'Answer: [LETTER]' at the end. The following is the question: *(exam question)*.” + *(textural description of drawing/table from the API call below) if applicable* Figure Description Prompt (separate application programming interface (API) call that precedes the standard call): “Please describe this figure in detail. Note that the description will be used in a downstream task to solve a related engineering problem.” + *(drawing/table)*
Experiment 3	“Your job is to answer multiple-choice questions from the PE exam, civil - water resources, and environmental discipline. Only provide your choice as A, B, C, or D in the format 'Answer: [LETTER]'. Do not include any explanation. The following is the question: *(exam question)*.” + *(drawing/table) if applicable*
Experiment 4	“Your job is to answer multiple choice questions from the PE exam, civil - water resources, and environmental discipline. First, reason through the question step by step. Second, if there is a figure, describe its key elements and their respective positions. Finally, provide your choice as A, B, C, or D in the format 'Answer: [LETTER]' at the end. The following is the question: *(exam question)*.” + *(drawing/table) if applicable*
Experiment 5	“Consider the following context if you find it relevant and useful: *(context retrieved through RAG)*. Your job is to answer multiple-choice questions from the PE exam, civil - water resources, and environmental discipline. First reason through the question step by step, and then provide your choice as A, B, C, or D in the format 'Answer: [LETTER]' at the end. The following is the question: *(exam question)*.” + *(drawing/table) if applicable*

There are 80 problems in the data set, out of which 76 are four-option multiple-choice questions. Accuracy is used as the quantitative metric for performance evaluation, so the four nonmultiple-choice questions are removed from the data set. For each question, the chatbot's solution is marked as correct if it provides a single choice in the specified format and the choice is correct. Since transformer-based language models are inherently stochastic (randomly determined) at inference time (Bender et al., 2021), each experiment is repeated three times. This approach allows observation of the variance in accuracy across three trials to ensure the data set's size is sufficiently large to reflect the chatbot's steady performance at the model's default temperature setting.

Experiment Results

Results from all experiments are shown in Table 3.

As a point of comparison to the 2023 results, the original questions, again without the multiple-choice options, were re-run with the understanding the chatbots are improving in accuracy and GPT-4o was now multimodal. With the improved chatbot, the accuracy improved to 48 percent correct within a 5 percent margin of error.

Conclusion

This GPT-4o-based chatbot demonstrates strong performance in solving engineering problems in the water industry. With proper

Continued on page 12

Continued from page 10

prompt engineering, it achieves a steady accuracy of 74.1 percent. A comparison between experiments 1 and 2 illustrates the model's capability to simultaneously draw insights across modalities and perform complex mathematical reasoning. In experiment 2, the textural description of a question's figure is generated in a separate model call so the model solves the problem itself from unimodal input. This setup is proved redundant since it underperforms the baseline setup in experiment 1. These results suggest that the model is entirely capable of integrating inputs across modalities and performing complex mathematical reasoning at inference time.

In terms of prompt engineering techniques in the context of multimodal inputs, the model benefits from instructions to describe visual input and reason through the question before attempting to answer it; the results from experiments 1, 3, and 4 support this observation. The model performs the worst under the setup of experiment 3, where the prompt suppresses the reasoning step and instructs the model to directly output the final answer. On the other hand, the prompt in experiment 4 instructs the model to describe key elements in the figure and reason through the question before providing the final answer. Under this setup, the model outperforms the baseline and achieves the highest accuracy of 74.1 percent, with the lowest standard deviation of 1.23 percent across three trials.

As the capabilities of the latest multimodal models in engineering problem solving are explored, it is crucial to identify its limitations. Experimental results indicate that the lack of low-level understanding of an image's content is the primary bottleneck. In a complex drawing, the model often struggles to associate specific annotations to the elements they refer to; it also has difficulties interpreting the spatial relationships among drawing elements. On the other hand, the lack of knowledge is not a limiting factor of the GPT-4o zero-shot performance since implementing RAG with PE Civil Handbook and PE Environmental Handbook in experiment 5 does not help the model outperform the baseline.

Acknowledgment

The authors would like to acknowledge Ripley Raubenolt for her contributions to the original 2023 project.

Works Consulted

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... and McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
2. Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., and Yin, W. (2024). Large language models for mathematical reasoning: progresses and challenges. arXiv preprint arXiv:2402.00157.
3. Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).
4. Hochreiter, S. (1997). Long Short-Term Memory. Neural Computation MIT-Press.
5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 33, 9459-9474.
6. Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32.
7. NCEES. Civil. (n.d.). Retrieved September 23, 2024. https://ncees.org/exams/pe-exam/civil/.
8. OpenAI. (n.d.). New Embedding Models and API Updates. Retrieved September 23, 2024. https://openai.com/index/new-embedding-models-and-api-updates/.
9. OpenAI. Hello GPT-4o. May 13, 2024. Retrieved Sept. 23, 2024. https://openai.com/index/hello-gpt-4o/.
10. Pursnani, V., Sermet, Y., Kurt, M., and Demir, I. (2023). Performance of ChatGPT on the U.S. fundamentals of engineering exam: comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. Computers and Education: Artificial Intelligence, 5, 100183.
11. Shannon, C. E. (1948). A mathematical theory of communication. The Bell System technical journal, 27(3), 379-423.
12. Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.
13. Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.
14. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C. C., Liu, Z., and Wang, L. (2023). The dawn of lmms: Preliminary explorations with gpt-4V(ision). arXiv preprint arXiv:2309.17421, 9(1), 1.

Table 3. Results and Summary

Results

Experiment 1 (baseline)	**Trial 1**	**Trial 2**	**Trial 3**	**Mean**	**Standard Deviation**
Score	55/76	59/76	51/76	n/a	n/a
Accuracy	72.4%	77.6%	67.1%	72.4%	4.29%
Experiment 2					
Score	56/76	52/76	54/76	n/a	n/a
Accuracy	73.7%	68.4%	71.1%	71.1%	2.16%
Experiment 3					
Score	43/76	40/76	44/76	n/a	n/a
Accuracy	56.6%	52.6%	57.9%	55.7%	2.26%
Experiment 4					
Score	57/76	57/76	55/76	n/a	n/a
Accuracy	75%	75%	72.4%	74.1%	1.23%
Experiment 5					
Score	57/76	51/76	50/76	n/a	n/a
Accuracy	75%	76.1%	65.8%	69.3%	4.07%

Summary

	Experiment 1 *(baseline)*	Experiment 2	Experiment 3	Experiment 4	Experiment 5
Mean Accuracy	72.4%	71.1%	55.7%	**74.1%**	69.3%
Standard Deviation	4.29%	2.16%	2.26%	**1.23%**	4.07%