Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #49192

Leveraging Large Language Models for Early Study Optimization in Educational Research

Mikayla Friday, University of Connecticut
Mr. Michael Thomas Vaccaro Jr, University of Connecticut

Michael Vaccaro is a fourth-year Ph.D. student in the School of Civil and Environmental Engineering at the University of Connecticut. He received his Bachelor of Science in Civil Engineering from the University of Connecticut in 2021. In addition to his work in structural engineering, Michael's interests in teaching and learning have inspired him to pursue interdisciplinary research spanning the fields of engineering, artificial intelligence, and neuroscience. His recent work in these areas has been supported by his major advisor's NSF MCA project and a transdisciplinary NSF Research Traineeship (TRANSCEND). Michael's engineering education research explores artificial intelligence's potential in K-12 science education, particularly in developing personalized learning environments.

Prof. Arash Esmaili Zaghi P.E., University of Connecticut

Arash E. Zaghi is a Professor in the Department of Civil and Environmental Engineering at the University of Connecticut. He received his PhD in 2009 from the University of Nevada, Reno, and continued there as a Research Scientist. His latest

Leveraging Large Language Models for Early Study Optimization in Educational Research

Mikayla Friday[1], Michael Vaccaro[1], and Arash Zaghi[1]
[1]University of Connecticut, School of Civil and Environmental Engineering

Abstract

Participant-based research remains essential in education-based experimental designs. However, there are many barriers to optimizing these studies prior to experimental deployment. While it would be ideal to include human participants in every stage of research design including early development, this is often not possible. In this study, we propose that Large Language Models (LLMs) can serve as preliminary participants during the early phases of experimental design. The robust role-playing capabilities of LLMs allow researchers to conduct trials with a simulated population representative of the target audience. We note that this approach is intended to optimize study designs prior to student involvement, rather than to replace human participants.

By leveraging LLMs in this manner, researchers can gauge potential pitfalls in and refine various aspects of their studies prior to deployment, such as question phrasing, question ordering, and response types. This is especially valuable in personalized learning environments powered by LLMs as the performance of these models is highly dependent on the prompts used. Using LLMs to simulate the experiment in early stages allows researchers to conduct as many trials as needed to tweak prompting, hyperparameters, and study design wherever necessary. We applied these principles in a small-scale text personalization study, where LLMs were used to adapt academic texts to users' learning preferences.

Using LLMs in place of human participants in the preparatory stage allowed us to account for potential flaws in our earliest experimental designs. Ultimately, we propose that these tools can be used to improve the design of participant-based studies. While not a replacement for humans, LLMs can serve as a valuable tool in the development and optimization of human subject studies. Future work will explore the scalability of this approach among different types of educational research.

Introduction

Large language models (LLMs) have emerged as powerful tools in education, offering the potential to transform classroom dynamics through automation, personalization, and enhanced student engagement [1]. Educators have already begun utilizing LLMs to generate lesson plans, streamline grading, and provide personalized feedback to students [2]. Additionally, LLMs have been implemented as Intelligent Tutoring Systems, assisting students in gaining a deeper understanding of challenging topics by offering tailored explanations and interactive learning experiences [3]. One particularly promising but underexplored application of LLMs in education is their potential for personalized learning (PL), specifically in the realm of text adaptation.

Unlike traditional PL approaches, which categorize students into predefined groups and assign them static learning materials, generative artificial intelligence (AI) enables a more dynamic and flexible personalization process. With LLMs, text can be adjusted in real time to align with an individual student's comprehension level, learning pace, and interests [4]. A recent systematic review by Khan et al. [5] discussed the uses of generative AI in engineering education, highlighting the need for further research. Our work helps to address some of the shortcomings of current approaches by developing a method to systematically evaluate the performance of LLM-based PL platforms prior to their deployment with human subjects.

Another recent literature survey by Razafinirina et al. [6] demonstrated that LLMs have the potential to advance PL within the classroom, but face challenges to implementation. One of the key challenges in the effective use of LLMs is designing optimal prompts. The relevance of LLM-generated content is highly dependent on the structure and refinement of these prompts, requiring iterative adjustments to ensure the generated responses can meet the user's needs [7]. The process of designing effective inputs, including iterative adjustment or refinement, is termed prompt engineering [8]. Some relevant prompt engineering techniques include providing examples in the prompts and breaking large tasks into smaller components [9]. It can be challenging to know which method will be the most effective in each case without the ability to test and compare the results produced by different prompting strategies.

To navigate the challenges associated with effective prompt engineering, we propose a novel method that leverages the role-playing capabilities of LLMs [10] to simulate human participants. This approach allows researchers to preemptively identify pitfalls in prompt design and pilot the data collection process before involving human participants. This may reduce the risks associated with deploying an ineffective intervention; for example, an LLM-based personalization platform that cannot effectively adapt to student preferences due to poor prompting. Simulation frameworks have been explored in various disciplines, such as survey refinement [11] and economic research [12]. However, such a framework has yet to be systematically applied to education research and, more specifically, the use of LLMs for PL.

This paper seeks to bridge this gap in research by outlining how LLM simulations can be used to refine prompts and inform the design of PL systems. The remainder of this paper is structured as follows: first, we provide an overview of our methods for developing the simulation framework, followed by an evaluation of its efficiency across model versions, a discussion of our findings, and our concluding thoughts. By introducing an LLM simulation framework for educational research, this study aims to advance the field of AI-driven PL, providing a structured methodology for refining and evaluating LLM-based interventions before real-world implementation.

Methods: Designing a Simulation Framework

Overview of the Simulation Framework

In this paper, we develop and test a method to optimize the design of LLM-powered text-personalization systems. The performance of these systems is determined by several factors, including the model version, model hyperparameters, and the prompting strategies. To iteratively test each of these factors, we present a framework where we utilize multiple LLM agents to simulate the text-personalization platform. Our text-personalization platform works as follows. First, students are administered a brief reading-based survey to gather information about their individual learning preferences. This short survey is composed of a few text excerpts that are designed to appeal to different types of learners. We then use students' feedback on these excerpts to modify new texts based on their identified learning preferences [13], [14].

We simulate the text-personalization platform described above using a series of LLM agents, each defined by its own system and user messages. The relationships between these agents are demonstrated in Figure 1. The first agent, which we refer to as the Profile Generator, is used to generate hypothetical student profiles. Generating a diverse set of profiles in this stage is crucial because it allows the text personalization technique to be evaluated for several types of learners. These generated profiles are then used within the system message for another LLM agent, known as the Student Bot. This agent plays the role of a student using the text-personalization platform. When presented with a series of educational texts, the Student Bot is prompted to indicate its preferences based on its given profile. Finally, two agents are used to assess the Student Bot's choices and to generate new content that is personalized for the user. These models are known as the Profiler and the Rewrite Bot, respectively.

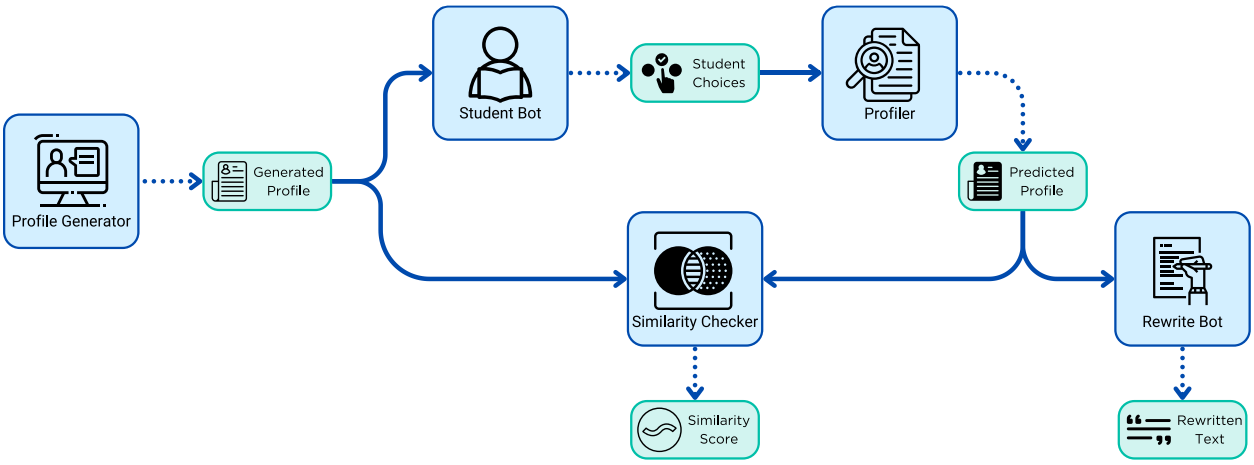

Figure 1. LLM agent roles, outputs, and interactions. Blue boxes represent individual LLM agents and green boxes represent the output. Dotted arrows connect the agents to their respective output, while solid arrows represent model inputs.

Intermediate Evaluation

To test the ability of the reading survey to effectively discern a student's learning preferences, we established an agent called the Similarity Checker. This agent is prompted with both the generated and predicted profiles, as shown in Figure 1. This agent provides a score between 0 and 100 that quantifies the similarity between the two profiles. Using an LLM-based Similarity

Checker enabled rapid evaluations of the platform's performance, allowing us to quickly update and refine prompting strategies, hyperparameters, and the study design.

While past researchers have used LLMs as evaluators [15, 16], we recognize that it is crucial that these models be used with human oversight. As such, we have conducted a reliability analysis between the scores generated by the Similarity Checker and by each of the first two authors, MF and MV. Specifically, one hundred profile pairs that were scored by the Similarity Checker were also scored independently by MF and MV. We found the Intraclass Correlation Coefficient (ICC) between the Similarity Checker, MF, and MV. The ICC was calculated following the methods outlined by Koo and Li [17] using SPSS Statistics version 29 based on a single rater, absolute agreement, two-way mixed effects model. Results found an ICC of 0.710 with a 95% confidence interval of (0.620, 0.785). This is typically viewed as moderate to good agreement in social science research [17].

In addition to the similarity scores, which quantified the similarity between the generated and predicted profiles, we evaluated the performance of the Rewrite Bot by recording how many times the Student Bot selected the personalized text over a generic text. Having multiple methods of performance evaluation embedded in the simulation allowed for the iteration of individual components. For example, high similarity scores and a low selection rate of the personalized texts would indicate that the Profiler is working well while the prompts or hyperparameters of the Rewrite Bot need modification. Having multiple methods of evaluation and the ability to repeat the entire simulation hundreds of times allowed us to optimize the PL platform.

Initial Paragraph Presentation to the Student Bot

We considered several options while deciding how to present the initial paragraphs to discern user preferences. The goal was to optimize the presentation of the initial paragraphs such that the most information possible could be gleaned about the users' learning preferences while minimizing the cognitive load placed on the user. We considered three main methods: ranking, rating, and choosing. For the ranking method, we displayed two sets of three paragraphs where the Student Bot was instructed to rank the paragraphs in each set from their favorite to least favorite. Within each set, the three paragraphs described the same topic but were written in different styles. For the rating method, we displayed six individual paragraphs and prompted the Student Bot with a 4-point Likert scale for them to report how much they liked or disliked the text presentation. Finally, for the choosing method, we provided three pairs of paragraphs and had the Student Bot select their favorite from each pair. We ran the simulation one hundred times with each of these frameworks and found no significant difference in profile similarity scores produced by the Similarity Checker agent between the different paradigms. The mean profile similarity scores of the ranking, rating, and pairwise choosing methods were 96.04, 96.22, and 94.91, respectively. A one-way ANOVA for differences between the methods yielded $p >> .05$, which is not statistically significant. Simulating the experiment allowed us to see that the initial presentation of the paragraphs did not affect the outcome and allowed us to focus on the method that was most applicable to our experiment.

In addition to evaluating the performance of the Profiler and the Rewrite Bot, the ability to simulate the experiment also enabled us to refine the system and user messages of the Profile Generator. Initially, we tried giving limited guidance to the Profile Generator, allowing it the freedom to generate student profiles randomly. However, we found that this was not an effective method of prompting because the profile content was inconsistent between generations. Constraining the content produced by the Profile Generator ensured that profiles would focus on the same educational aspects of the users between separate generations. Constraining the Profiler in a similar way further enabled meaningful comparisons between the generated and predicted profiles. For this reason, we used the Felder-Silverman Learning Style Model (FSLSM) [18] as a basis for both the generated and predicted profiles. The FSLSM has precedence within the fields of educational technology and PL [19, 20], making it a pragmatic choice for our PL platform. After testing our experiment with the updated prompts, we saw a marked improvement in the consistency of the profiles and the results of the text-personalization experiment. The FSLSM has four dimensions, and each dimension has two extremes. Because of this, a pairwise presentation of texts (i.e., pairwise choices) rather than ranking or rating items was used. Since the FSLSM has four dimensions, we included four pairwise choices rather than three.

Establishing the Need for the Profiler: Best Method for Generating Personalized Content

We were interested to know if having the Profiler in between the output of the Student Bot and the input of the Rewrite Bot benefited the outcome of the text personalization or if it was an unnecessary step. We therefore tested two different methods for generating personalized texts: explicit and implicit. The explicit framework was described previously and is demonstrated in Figure 1. We bypass the Profiler in the implicit framework, incorporating the Student Bot's choices directly into the prompt of the Rewrite Bot. The implicit framework is presented in Figure 2. We tested each framework one hundred times in our simulated environment. An independent samples *t*-test between the mean number of times the personalized text was chosen for each group shows a statistically significant improvement when the explicit framework is used ($p < .001$***). This finding aligns with OpenAI's suggested strategy of chunking tasks and breaking prompts into steps [9]. Therefore, in our final experimental design, we prompted the Rewrite Bot with the output from the Profiler. In the next Section, we evaluate the performance of the explicit simulation framework with different model versions and hyperparameters.

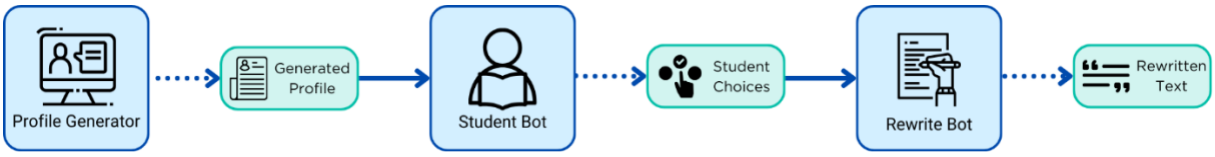

Figure 2. Simulation framework without the Profiler (implicit framework). Dashed arrows connect the agents to their respective output, while solid arrows represent model inputs.

Evaluating the Simulation Framework with Various Models and Hyperparameters

OpenAI is consistently developing new GPT models [21]. To test the stability of the framework across releases, we ran the simulations using six different GPT versions, including *gpt-3.5-turbo* and several models in the GPT-4 family. Tables 1 and 2, below, summarize the performance of the different models in terms of the similarity scores and the number of times the personalized text was selected. In general, we found that the performance decreased when using newer GPT-4 models. In an attempt to account for the decrease in performance with allegedly more powerful models, we adjusted the hyperparameter that controls the randomness of the output, i.e., the temperature. Higher temperatures lead to greater randomness in the output, decreasing the coherence of the model's response [22]. We found that the performance of newer models, namely *gpt-4o*, were optimized for temperatures near 0.7 while older releases like *gpt-4-1106-preview* worked well with temperatures near the default of 1.0. We believe that newer models may need lower temperatures to ensure that the responses are consistent and remain relevant to the task provided in the prompt.

Table 1. Profile Similarity Scores Across GPT Models and Associated Accuracy Scores.

	Mean Profile Similarity Scores			
Model	**Actual**	**Opposite**	**Random**	**Accuracy**
GPT-3.5-Turbo	70.05	63.95	68.05	53%
GPT-4	77.68	49.8	64.32	77%
GPT-4-1106-Preview	79	34.35	62.5	88%
GPT-4-Turbo	66.05	48.85	58.6	65%
GPT-4o	72.9	40.45	61.4	84%
GPT-4o-2024-08-06	68.8	37.35	55.1	78%

Table 2. Number of Times Rewritten Paragraph was Selected by Student Bot.

	Times Rewrite was Selected			
Model	**Both**	**One**	**Neither**	**Total**
GPT-3.5-Turbo	66	22	12	77%
GPT-4	91	7	2	94.50%
GPT-4-1106-Preview	86	4	10	88%
GPT-4-Turbo	85	5	10	87.50%
GPT-4o	70	14	16	77%
GPT-4o-2024-08-06	73	14	13	80%

To further ensure the validity of the Similarity Checker, "Opposite" and "Random" profiles are also generated. Recall that the Profiler takes the student choices as input. Opposite profiles are generated by reversing the Student Bot's choices, while random profiles are generated using a random set of choices. In Table 1, the mean similarity scores over one hundred trials are shown comparing the provided profile to each of the Actual, Opposite, and Random profiles generated by the Profile Generator. It is important to note that certain model versions possess distinct strengths and weaknesses, even within the same family of LLMs [21]. Thus, the ability to test each model in a simulated environment can help researchers select the most appropriate model for the task at hand. Finally, we note that even though our experiment only assessed OpenAI

GPTs, this method would also be beneficial in testing different LLMs, e.g., Claude or Gemini, and assessing the skills of these different models.

Discussion

The ability to test many variations of our experiment provided the opportunity to modify and refine our framework in ways that were nearly impossible prior to the advent of LLMs. While we applied this principle in the case of a PL platform, we believe that the significance of this framework extends beyond the field of educational technology. The ability to repeatedly test your experiment, modify it incrementally, and test for different variables is a significant contribution to the scientific method. Importantly, this method provides researchers with an opportunity to create an effective and optimized experimental framework prior to deployment with human subjects.

Another benefit of repeated simulation is the ability to assess trends within the experiment. This allowed us to ensure that the platform would perform with the same level of accuracy for different student profiles. For example, we could assess if certain profiles were associated with lower rates of personalized paragraph selection, and retroactively tweak prompting to better accommodate users with those profiles.

Limitations

The basis of our experiment hinges upon the role-playing capabilities of LLMs and their ability to act as participants in early phases of experimental design. While the role-playing capabilities of LLMs are documented [10], we are aware that LLMs cannot account for human complexities. LLMs cannot emulate a single human; rather, they can only embody a combination of traits that a human may have. Additionally, LLMs do not experience situational shifts, such as mood, that humans experience. We believe that it is very likely that there will still be experimental elements that are not perfectly optimized when the study is deployed and, as such, we believe that human-based pilot tests will remain critical to experimental design. We emphasize that the simulation method discussed here should only be applied in preliminary stages of experimentation. For example, we used the results of this research to inform an exploratory study investigating the ability of an LLM to personalize science texts for middle school students [23].

Conclusion

In this study, we designed a simulation framework to aid in the development of a small-scale text personalization platform for middle school students. By implementing the simulation framework, we were able to repeatedly modify different aspects of the experiment, such as paragraph presentation and how personalized text was generated. This allowed us to assess the effectiveness of our changes in the early and intermediate stages of the design process. However, we recognize that LLMs cannot fully simulate the nuance of human behavior and, thus, do not serve as a replacement for human subjects. By using LLMs as simulated participants during

experimental development, researchers can test different ideas in an efficient and cost-effective manner. Future work will aim to extend this approach to larger-scale studies and more diverse educational contexts to assess the scalability of simulation-aided design. While our work has been limited to the field of educational technology, we anticipate that this approach can be adapted to human-based research studies in different fields. We hope that this paper serves as a practical guide for using LLMs for the simulation and early optimization of experimental designs across disciplines.

Acknowledgments

This material is based upon work supported by the National Science Foundation under MCA Grant No. 2120888. The first and second authors (MF and MV) were supported by an NSF Research Traineeship (TRANSCEND) under Grant No. 2152202 at the time this research was conducted. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] B. Dong, J. Bai, T. Xu and Y. Zhou, "Large Language Models in Education: A Systematic Review," in *2024 6th International Conference on Computer Science and Technologies in Education (CSTE)*, 2024, doi: 10.1109/CSTE62025.2024.00031.

[2] E. Bonner, R. Lege and E. Frazier, "Large Language Model-Based Artificial Intelligence in the Language Classroom: Practical Ideas for Teaching," *Teaching English with Technology,* vol. 23, no. 1, pp. 23-41, 2023, doi: 10.56297/BKAM1691/WIEO1749.

[3] J. Stamper, R. Xiao and X. Hou, "Enhancing LLM-Based Feedback: Insights from Intelligent Tutoring Systems and the Learning Sciences," in *International Conference on Artificial Intelligence in Education*, 2024, doi: 10.1007/978-3-031-64315-6_3.

[4] D. Zapata-Rivera, I. Torre, C. S. Lee, A. Sarasa-Cabezuelo, I. Ghergulescu and P. Libbrecht, "Editorial: Generative AI in education," *Frontiers in Artificial Intelligence,* vol. 7, 2024, doi: 10.3389/frai.2024.1532896.

[5] R. Khan, S. Bhaduri, T. Mackenzie, A. Paul, S. KJ, Sen and I, "Path to Personalization: A Systematic Review of GenAI in Engineering Education," in *AI4Edu Workshop at KDD '24*, Barcelona, Spain, 2024.

[6] M. A. Razafinirina, W. G. Dimbisoa and T. Mahatody, "Pedagogical Alignment of Large Language Models (LLM) for Personalized Learning: A Survey, Trends and Challenges," *Journal of Intelligent Learning Systems and Applications,* vol. 16, no. 4, 2024, doi: 10.4236/jilsa.2024.164023.

[7] L. J. Jacobsen and K. E. Weber, "The Promises and Pitfalls of Large Language Models as Feedback Providers: A Study of Prompt Engineering and the Quality of AI-Driven Feedback," *AI,* vol. 6, no. 2, p. 35, 2025, doi: 10.3390/ai6020035.

[8] G. Mizrahi, Unlocking the Secrets of Prompt Engineering: Master the art of creative language generation to accelerate your journey from novice to pro, 1st ed., Packt Publishing, 2024.

[9] OpenAI, "OpenAI API-Prompt Engineering," [Online]. Available: https://platform.openai.com/docs/guides/prompt-engineering. [Accessed 21 February 2025].

[10] Z. M. Wang, Z. Reng, H. Que, J. Liu, W. Zhou, Y. Wu, H. Guo, R. Gan, Z. Ni, J. Yang, M. Zhang, Z. Zhang, W. Ouyang, K. Xu, S. Huang, J. Fu and J. Peng, "RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models," *[arXiv Preprint],* 2024, doi: 10.48550/arXiv.2310.00746.

[11] D. Dillon, N. Tandon, Y. Gu and K. Gray, "Can AI language models replace human participants?," *Trends in Cognitive Science,* vol. 27, no. 7, pp. 597-600, 2023, doi: 10.1016/j.tics.2023.04.008.

[12] A. Filippas, J. H. Horton and B. S. Manning, "Large language models as simulated economic agents: What can we learn from Homo Silicus?," in *EC '24: Proceedings of the 25th ACM Conference on Economics and Computation*, 2024, doi: 10.1145/3670865.3673513.

[13] M. Vaccaro, M. Friday and A. Zaghi, "Transforming K-12 STEM Education with Personalized Learning through Large Language Models (Fundamental)," in *2025 ASEE Annual Conference & Exposition*, Montreal, Quebec, Canada, 2025.

[14] M. Vaccaro, M. Friday, Z. G. Akdemir-Beveridge and A. Zaghi, "Designing and Testing AI-Based Text Personalization Tools," in *2025 ASEE Annual Conference & Exposition*, Montreal, Quebec, Canada, 2025.

[15] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J. Li, M. Jiang and R. A. Metoyer, "Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks," in *IUI '25: Proceedings of the 30th International Conference on Intelligent User Interfaces*, 2025, doi: 10.1145/3708359.3712091.

[16] L. Ouyang, J. Wu, X. Jiang, D. Almeida, W. C. L, P. Mishkin, C. Zhang, S. Agarwal, K. Salma, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike and R. Lowe, "Training language models to follow instructions with human feedback," in *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.

[17] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine,* vol. 15, no. 2, pp. 155-163, 2016, doi: 10.1016/j.jcm.2016.02.012.

[18] R. M. Felder and L. K. Silverman, "Learning and teaching styles in engineering education," *Engineering education,* vol. 78, no. 7, pp. 674-681, 1988.

[19] C. A. Carver, R. Howard and W. Lane, "Addressing different learning styles through course hypermedia," *IEEE Transactions on Education,* vol. 42, no. 1, pp. 33-38, 1999, doi: 10.1109/13.746332.

[20] J. Kuljis and F. Liu, "A Comparison of Learning Style Theories on the Suitability for elearning," *Web Technologies, Applications, and Services,* pp. 191-197, 2005.

[21] OpenAI, "Hello GPT-4o," 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/. [Accessed 21 February 2025].

[22] M. Peeperkorn, T. Kouwenhoven, D. Brown and Jordanous, "Is temperature the creativity parameter of large language models?," *[arXiv Preprint],* 2024, doi: 10.48550/arXiv.2405.00492.

[23] M. Vaccaro, M. Friday and A. Zaghi, "Evaluating the capability of large language models to personalize science texts for diverse middle-school-age learners," *[arXiv PrePrint],* 2024, doi: 10.48550/arXiv.2408.05204.