Journal of Research High School

Vol. 2026 (2) 109 – 117
ISSN 2688-3651 [online]

Can LLMs Pass the SAT? A Study on US and World History Exams

Izhan Ali[1] *

[1]Niskayuna High School, Niskayuna, NY, USA
*Corresponding Author: izhan.ali08@gmail.com

Advisor: Rostyslav Korolov, korolr2@rpi.edu

Received April 20, 2025; Revised September 27, 2025; Accepted November 14, 2025

Abstract

Large Language Models (LLMs) have found usage in various language-based tasks. One of the most powerful applications of LLMs is the sophisticated question-answering (Q&A) chatbot. These chatbots have the ability to answer questions without being trained on data pertaining to the topic. This is called zero-shot learning (ZSL) that uses already learned information to answer new questions. The goal of this study is to analyze the effectiveness of a well-known LLM for history exam preparation under two subject categories: US and World History under a ZSL approach. In this study, the LangChain platform, which is open source, was used for building Artificial Intelligence (AI) applications driven by an existing LLM. Google Gemini Pro model was used to create an application that generates text by utilizing some relevant input pertaining to SAT exams (History). A publicly available dataset from known sources for SAT level history course (SAT History) was used. Additionally, the quality of the text generated by fine-tuning model parameters was analyzed. Experiments were conducted with various hyperparameters of the model to assess the impact on the correctness of the results along with a comprehensive error and harmful content analysis. Results indicate strong performance on US history questions and slightly lower effectiveness on World history. This indicates content familiarity bias. Findings from the study highlight potential benefits and limitations of LLMs as educational tools for standardized test preparation. This application and the results of this analysis can be used both by students and teachers to understand the effectiveness of LLMs in helping them better prepare for exams in general.

Keywords: Artificial intelligence, Large Language Models (LLMs), Zero-shot learning (ZSL), Education, Google Gemini, LangChain

1. Introduction

With the advent of ChatGPT (OpenAI, 2023) the use of LLMs surged in the field of education. For students, these models offer the advantage of having access to more study material in addition to what is provided at school. Multiple studies suggest that more worked-examples are helpful for learners in introductory college courses and having access to more relevant content can help lower cognitive overload (Jury et al., 2024). Students who aim to seek admission in competitive institutions for their undergraduate degree opt for AP exams (similar to the old SAT subject-tests), to improve their admission chances and enhance the ability to get scholarships. In general, SAT subject tests offered a college level curriculum to high school students until 2021. Many of the AP (Advanced Placement) courses have a similar curriculum and these courses offer a path to obtain college level knowledge in a shorter period of time (Conger et al., 2021). All these factors contribute to the popularity and significance of these exams. Despite their advantages, these exams and courses are significantly challenging given the advanced content covered. Recent advances in LLM research have seen a surge in chatbots powered by these models. ChatGPT and Google Gemini are two of the prominent ones. Many studies point towards the effectiveness and potential benefits of these tools for educational purposes i.e. their usage by students (Halaweh, 2023). Researchers have examined the usage of these models to pass

Vol. 2026 (2) 109 – 117
ISSN 2688-3651 [online]

actual exams (Chang et al., 2024, Kortemeyer, 2023). If students are using the content generated by the LLMs for exam preparation, then there should be no misinformation in the generated material. LLMs are known to generate (hallucinate) incorrect and fabricated information (Perković et al., 2024, Chang et al., 2024). According to literature, hallucination in an LLM is a situation when the model produces text that is fictional, misleading or entirely fabricated. This is a bigger concern for fields like history where factually correct information is extremely critical. Using fabricated content for exam preparation can be detrimental to student success in general. Additionally, many LLMs are trained to counter hate speech and avoid sensitive topics that are usually part of the history curriculum (Albladi et al., 2025).

This study investigates how well these LLMs can assist students in preparing for exam when used as zero-shot learners (Bhattacharjee et al., 2024, Xian, 2017). The focus of this study is the SAT World History exam using a dataset of SAT (Scholastic Assessment Test) style questions. The dataset used for this analysis is a publicly available dataset (see Appendix for source). It consists of a collection of questions and answers for the SAT subject test in World History and US History. While the SAT subject tests were discontinued in 2021, the content coverage matches with the AP level curriculum; therefore, this can give a good estimate of how effective LLMs are in helping students prepare for history AP and SAT style standardized exams in general. The final goal is to assess the effectiveness of a well-known LLM (Google Gemini (Team G et al., 2023)) for this problem by creating a LangChain (framework for enabling LLM integration with other sources) powered application. Google Gemini is known to have features that are useful in education (Imran & Almusharraf, 2024) which is why this model was selected for this study. An in-depth analysis of LLM generated results is performed using various hyperparameters of the model via a carefully created experimentation process. This was done to uncover the model's effectiveness under different configurations for result improvements.

The key contributions of this research are:

1. A question-answering model is developed by leveraging the LangChain platform to integrate Google's Gemini Pro model, optimizing it for SAT-level history question generation via a ZSL approach.
7. A publicly available data set is used to analyze the quality of the responses generated, ensuring the effectiveness and correctness of the model for exam preparation. This is done for two sub-categories of history i.e. US versus World history.
8. To improve the quality of the responses generated an experimental framework for hyperparameter tuning is developed.
9. A framework for in-depth error analysis is created that takes a holistic approach towards analyzing wrong answers. This includes accuracy, precision, recall, F-1 score and semantic analysis of errors using transformers.
10. The effectiveness of LLMs is analyzed for sensitive topics and hate speech.
11. The resulting application is a useful tool for teachers and students. Students can use this as a tool for exam preparation. Teachers can use this as an effective pedagogical framework.

2. Background and Literature Review

Natural language processing (NLP) is a sub-field of AI that enables computers to interpret and generate human language. It is at the intersection of machine learning (ML) and linguistics and enables the creation of text and speech analysis tools such as chatbots, translation applications, search engines and text summarization. LLMs are NLP models that have been trained on a vast variety of textual data from various sources. LLMs have the capability to answer questions, generate and summarize human-like text and solve new problems. Of particular interest to us is their ability to answer questions without requiring any additional domain-specific knowledge. This is referred to as zero-shot learning (Zhang et al., 2023). Most people who use LLMs like ChatGPT do not update or train the model therefore it is relevant in the context of usage to understand the effectiveness of these models. This feature is useful in education applications especially for the usage of LLMs by students and teachers. Chatbots built on LLMs significantly improve the learning process (Neumann et al., 2014). They can act as virtual tutors helping students filter through excessive information, find course materials and act as aides in problem solving (Stamper et al., 2024). Modern day LLMs mimic human critical thinking and generate human-like text with inputs like basic prompts (Zhang et al., 2024). With little effort students have access to a wide range of educational content through these models.

Vol. 2026 (2) 109 – 117
ISSN 2688-3651 [online]

Some recent studies suggest there are challenges associated with LLMs like producing wrong information, lacking reasoning and logic leading to hallucinations (Searson, Langran & Trumble, 2024). There is research that verified inconsistencies in the output from ChatGPT especially if the output involved long text (Alkaissi & McFarlane, 2023). LLMs are largely trained on data available on the internet which may include misinformation. Some of the recent models have incorporated some filters to deal with hate speech and other sensitive content (Albladi et al., 2025). Students who use LLM based chatbots as study tools for subjects like history have two important challenges: 1) the information generated by the chatbot should be factually correct 2) any fact pertaining to historical events should not be filtered and/or altered by the speech monitoring algorithm. Therefore, it is important to understand the effectiveness of these LLMs when used as a preparation tool for advanced tests for example a history test.

This study investigated the effectiveness of a well-known LLM Google Gemini (Team G et al., 2023) by using a publicly available dataset (see Appendix for source) for SAT World and US history. The dataset has 1380 questions and answers targeting the multiple-choice section of the SAT exam. This dataset was provided as input to the Gemini model and the responses for each question were recorded. The effectiveness analysis was done by calculating the metrics Accuracy, Recall, Precision and F1-score. An in-depth semantic analysis was done on the error to find underlying patterns. Additionally, an improvement analysis was conducted by changing the hyperparameters of the model for randomly selected inputs and comparing the results.

3. Materials and Methods

3.1 Dataset and Preprocessing

The dataset used for this analysis is a text classification dataset that is publicly available on Kaggle (link in Appendix). Data analysis was performed by using python on Google Colab Pro. The dataset is in a csv format. Data was uploaded and converted to a pandas data frame (McKinney, 2011). A data frame is an efficient data structure used to store and manipulate data. This dataset has 1379 usable questions (one invalid observation was dropped). The column called subject has two categories – US history (1107 questions) and world history (272 questions). There are five columns for answer choices labeled with English alphabet from letter A to letter E. These are options for the multiple-choice question from the prompt column. The column called answer has the correct answer saved in it as the correct English alphabet letter. An additional column called "formatted_prompt" was created by combining the value from prompt and the five answer columns. Finally, a column called "predicted output" was added to the data frame where the output of the Gemini model was recorded for every prompt. Only the letter answer (A, B, C, D, E) was recorded as the predicted answer to compare with the actual answer. Therefore, the LLM response was saved in a variable and the letter related to the response was extracted and saved into the predicted answer column. Another predicted answer 'F' was created for any questions that were labelled as harmful content and not answered by the LLM. Additionally, the study utilized the first 801 observations only for analysis (529 for US history and 272 for World history). This was done because the LLM has restrictions on the number of questions it can answer in a given amount of time. The dataset had only 272 questions for world history therefore, any number greater than or equal to this is reasonable for comparison for US history.

3.2 LLM Selection

LLMs are deep learning models built on the idea of transformers introduced in (Vaswani et al., 2017). Different LLMs are trained to enhance different application capabilities (Razafinirina et al., 2024). Google Gemini is a powerful LLM that is found to be effective in education related tasks as demonstrated in (Imran & Almusharraf, 2024). This study showcases Gemini's ability to handle education related tasks like communication, adaptive and informative responses and efficient and systematic assessment of assignments and tasks. This study also found that Gemini has a higher capability to assist learners for advanced research and analysis tasks. Gemini is known to have better qualitative analysis capabilities with an improved ability to understand and reason (Team G et al., 2023). The goal of this research is to analyze an LLM's ability to assist learners in learning advanced topics in SAT History. Therefore, Google Gemini was the first choice. Additionally, using the best available model in this domain will set a baseline for subsequent

Vol. 2026 (2) 109 – 117
ISSN 2688-3651 [online]

research in this field. ZSL is a technique used in machine learning where a model can classify data from categories it has not seen during training. This technique gives LLMs the ability to utilize semantic relationships that it learnt during training and convert it into knowledge transfer when dealing with unknown classes/data. This is the technique for which the model is tested.

3.3 Experimental Set-up

All experiments were run on Google Colab Pro notebooks. The Pro version of google colab is a paid version that allows access to GPU and TPU.

Installation

The langchain-community package was installed. This package contains community-maintained integrations and tools that enable developers to use LangChain with various data sources, APIs (Application Programming Interface) and services. Next, langchain-google-genai was installed that provides integrations between LangChain and Google Generative AI models (like Gemini). This installation allowed the usage of Google's AI models within the LangChain framework enabling text generation.

Importing Necessary packages

After installation the packages used for data analysis and model building (GoogleGenerativeAIEmbeddings and ChatGoogleGenerativeAI) were imported. The former class provides access to embedding models that are necessary for any text related task. Embeddings are word representation in numeric (vector) form. The latter class allowed interaction with Google's Gemini model in a chat-like interface.

Account configuration

The genai.configure() function was used to set up authentication for accessing Google generative AI services within the LangChain package that is required to run this application. This is unique for each user.

Model usage

The list_models() function was used to generate the list of available models and the respective methods that each model supports. For example, gemini-pro has one of the available methods as 'generateContent' which was the required method for the problem at hand. To process the entire dataset all at once, a for loop was created to generate the relevant answer for each question in the dataset. The generated answer was saved in a new column called predicted_answer in the original dataframe. While running this loop an error was encountered that pertained to the rate limit of the model. To deal with this issue a time delay of a few seconds was incorporated in the loop (see code in Appendix). Many of the requests were flagged as harmful i.e. these requests fall under the category of hate speech, violence, harm, misinformation etc. For all such questions the answer was recorded as 'F' because this was not an option in the given set of answers.

4. Results

The goal of this study is to assess the effectiveness of a well-known LLM when preparing for the AP style SAT history exam. The following evaluation metrics were used to assess the ZSL capability of the Google Gemini model for US and World history:

4.1 Accuracy Measurement

Accuracy measures the correctness of the answers as compared to the correct answer. It is calculated as the ratio of the number of correct answers given by the LLM divided by the number of

Table 1: Based on the results of the Google Gemini model, Accuracy, Precision, Recall and F1-score are calculated for US and World History exam questions.

Metric	US History	World History
Accuracy	84.12	77.20
Precision	89.00	82.68
Recall	84.12	77.20
F1-score	86.49	79.20

Vol. 2026 (2) 109 – 117
ISSN 2688-3651 [online]

questions in all. A higher accuracy value for US history questions (83.93%) was observed compared to World history questions (72.05%) as shown in Table 1.

4.2 Precision, Recall and F1-score

Precision is a measure of the correctness of model answers for each category. Recall measures the model's ability to identify correct answers in each category. F1-score combines both these measures by calculating their harmonic mean. F1-score is beneficial in identifying both true positives and false positives. A higher F1-score (Table 1) was recorded for US history questions compared to World History (around 7% lower).

4.3 Baseline Comparison:

To contextualize the results of the LLM (Table 1), two baselines were compared that involve human performance benchmarks. The first is a random guess baseline. Under this model the assumption is that all questions have four answer choices, therefore the expected accuracy of random guessing will be 25%. The Gemini model performs much better than random guessing for both US and World history. The second baseline is student performance on AP History exams. According to the College Board AP score distribution data from 2025, percentage of students getting a 3 (passing grade) or higher was 73% for AP US History and 64% for AP World History. For a valid comparison with LLM performance, the scores were converted to percentage of questions answered correctly. A score of 3 corresponds to answering 50-60% questions correctly, a 4 to 70-80% and a 5 to 80-90%. The Gemini model's accuracies were comparable to high performing AP students for both exams (who scored a 4 or a 5, with 70-90% correct answers approximately).

4.4 Semantic Similarity Analysis

In history exams many responses are similar to the correct answer but are incorrect. Sometimes these are candidates for further investigation. At the core of any LLM, there exists a numeric representation of each sentence. This is called vector embedding. In this analysis the semantic similarity (cosine similarity) between predicted (wrong) answer and correct answer was calculated to understand why some of the responses were incorrect. A pre-trained sentence transformer from the Hugging Face Transformers library was used to generate sentence embeddings for each response. The model used was "all-MiniLM-L6-v2" which is a lightweight and fast sentence embedding model. These sentence embeddings are numeric representations of text (sentences) and can be used to determine sentence similarity (cosine similarity in this case). Cosine similarity is a metric that is used to measure similarity between two vectors, and its value is bounded between 1 and -1. 0 means no similarity, 1 means maximum similarity and -1 means maximum dissimilarity. We found a 0.592 average similarity for US history and 0.583 for world history embeddings. Both the values are above zero but not close to 1. In practical terms it means these items share some commonalities, but they are not strongly similar. These moderate similarities between the correct answer and predicted (wrong) answer could be a source of confusion for the LLM.

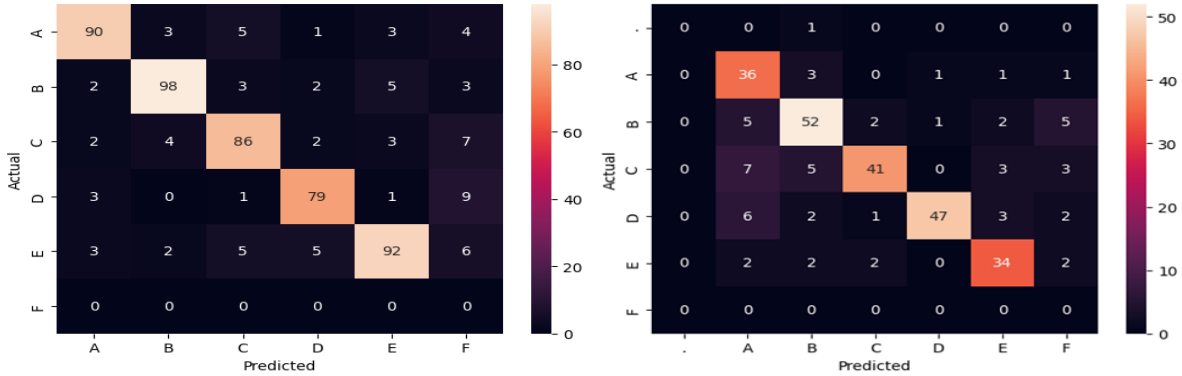

Figure 1: Left Confusion matrix for each answer option in US history. Right Confusion matrix for each answer option in World history.

Vol. 2026 (2) 109 – 117
ISSN 2688-3651 [online]

4.5 Harmful Content Analysis:

Many of the questions were flagged as harmful by LLM. These were marked separately as F. Figure 1 shows the confusion matrix with different categories of responses flagged as harmful by Google Gemini. Out of the wrong answers 34.5% answers were marked as harmful for US history and 20.9% for World history. The average cosine similarity between all answers marked as F (calculated pairwise) was 54.6% for US history questions and 58% for World history questions. This points to the fact that a varied range of topics were marked as harmful owing to their dissimilarity.

4.6 Error Analysis

To understand the wrong answers better a hyperparameter tuning scheme was devised to check for the right calibration of the model for difficult history exam questions. For all the incorrect answers experiments were conducted with the model's hyperparameters to tune it for better results. Here is a description of each hyperparameter that was used (Team G et al., 2023):

Temperature
This has a range between 0.0 to 2.0 and it controls the randomness of the output. Lower values align responses with more probable responses while higher value makes the model more creative and diverse. The default temperature value for Google Gemini is 1 therefore wrong answers were tested for two temperature values (0 and 1.5) to analyze if correct answers can be obtained by decreasing or increasing temperature. For 0 temperature value all questions within the experiment sample were answered correctly; for 1.5 value, out of 84 wrong answers (by the original model) 4 were answered correctly. For World History 2 out of 62 were answered correctly for temperature value 1.5 and only one was answered wrong when temperature value was 0. Our findings show that increasing temperature will not lead to more accurate answers for history questions.

Top-k sampling
This controls the number of probable tokens. The values range from 1 to 100. Lower values will lead to more deterministic responses and higher values make responses more diverse. The experiments tested values 16, 32 and 40. For both datasets all three resulted in the same answers. For US history, only 4 out of the 84 wrong answers were correct. For World history 3 out of 62 were correct. There was no notable trend here.

Candidate count
This parameter specifies how many candidate solutions the model should generate before choosing the final one. Higher values lead to selecting the best response. Through the experimental runs (April 2025) it was discovered that a limit of 1 has been added to this parameter, therefore we could not evaluate this further.

5. Discussion

In this research, the performance of a well-known LLM (Google Gemini) was investigated for SAT History exam preparation using a ZSL framework. This analysis involved the creation of a question-answering approach via a publicly available dataset that had questions from US and World history exams. Evaluation was performed across multiple metrics like accuracy, precision, recall, F1 score, semantic similarity, harmful content filtering and hyper-parameter tuning. The rationale for in-depth evaluation is to understand the effectiveness of using LLMs for fact-intensive subjects like history for exam preparation by students.

The results of this study indicate that Gemini performs better in answering questions from US history than World history evincing higher accuracy and F1-scores. One of the reasons behind this difference could be attributed to training data distribution. Google Gemini, like other LLMs, is trained on publicly available internet data (Google, 2025), therefore it's likely to encounter content that is more related to US history compared to region-specific topics appearing in World History. These results align with previous studies that identified topic-based variability in LLM

Vol. 2026 (2) 109 – 117
ISSN 2688-3651 [online]

performance (Kortemeyer, 2023). In the semantic analysis of wrong answers, it was found that the average similarity scores were 0.592 (US history) and 0.583 (World History). This indicates that when the model produced wrong results, its answers were semantically less related which points to the confusion regarding a wide range of facts in both US and World history topics. For high stakes environments, this semantic confusion can lead to bigger issues like introducing inaccuracies and misleading students. An important concern in this study was the issue of sensitive content handling by LLM. A significant number of questions were marked as harmful especially in US history. Responsible content moderation is needed for these tools, but this raises an important pedagogical issue for subjects like history. History inherently includes sensitive and politically charged topics. Filtering these topics by labeling them as 'harmful' may compromise the integrity of the material presented to students. The hyperparameter tuning analysis indicated there exists a trade-off between precision and creativity in terms of the LLM responses. Therefore, there is a need for domain-specific update and tailoring of LLMs when used in educational settings.

6. Conclusion

In this research, the performance of Google Gemini's LLM was examined on standardized history exam questions via ZSL. Results reveal a promising performance in accuracy, precision, recall and F1-score particularly on US history questions compared to World history. The in-depth error analysis revealed that accuracy was affected by issues such as overcautious content filtering and inconsistent reasoning for questions that required historical inference. Overall, this study emphasized the potential of LLMs as useful tools for education and standardized test preparation. Large scale usage will require careful calibration and training on diverse datasets to reduce inaccuracies. Adjustment to generation parameters, especially temperature, can have a notable impact on performance. There is a need for higher transparency in these models when used by educators for teaching. In general, both educators and students need more information on confidence level of answers and some rationale behind content generation.

This study was limited to the use of a single LLM, and it is acknowledged that results may vary with newer versions of the same LLM or other LLMs trained on different datasets. One future research direction will be to compare the performance of multiple models. The dataset used was SAT level, but it may not represent the entire history curriculum, and future research will include considering datasets from other sources as well. Further, Gemini's rate limit and some restrictions on parameters limited the scope of the experimentation. Future research will explore avenues of deeper analysis to assess the potential of the model. Another future research direction is to compare human performance with the LLM performance on standardized tests.

This work contributes to the fast-growing area of research called domain-specific evaluation of LLMs (Jeong, 2024). This is of relevance in fields like history that require subjective interpretation, contextual nuance and factual integrity for content expertise. The technique presented in this study involves a combination of ZSL with qualitative error analysis and hyperparameter tuning. This can serve as a replicable framework for assessing LLM performance in other subjects. LLMs are helping us getting closer to personalized and enhanced learning environment. However, there is a trade-off between novel learning techniques and academic rigor that needs to be preserved as the world moves forward with AI usage in education.

Appendix
Code: https://github.com/IzhanAli08/LLMResearch
Data: https://www.kaggle.com/datasets/trainingdatapro/sat-history-questions-and-answers

References

Albladi, A., et al. (2025). Hate speech detection using large language models: A comprehensive review. IEEE Access. doi:10.1109/ACCESS.2025.3532397

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2). doi:10.7759/cureus.35179

Vol. 2026 (2) 109 – 117
ISSN 2688-3651 [online]

Bhattacharjee, A., et al. (2024). Zero-shot LLM-guided counterfactual generation for text. *arXiv Preprint*. doi:10.48550/arXiv.2405.04793

Chang, Y., et al. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1–45. doi:10.1145/3641289

Conger, D., et al. (2021). The effect of Advanced Placement science on students' skills, confidence, and stress. *Journal of Human Resources*, 56(1), 93–124. doi:10.3368/jhr.56.1.0118-9298R3

Google. (2025). Gemini Code Assist overview. Google for Developers. Retrieved from https://developers.google.com/gemini-code-assist/docs/overview

Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology*, 15(2).

Imran, M., & Almusharraf, N. (2024). Google Gemini as a next generation AI educational tool: A review of emerging educational technology. *Smart Learning Environments*, 11(1), 22. doi:10.1186/s40561-024-00310-z

Jeong, C. (2024). Fine-tuning and utilization methods of domain-specific LLMs. *arXiv Preprint*. doi:10.13088/jiis.2024.30.1.093

Jury, B., et al. (2024, January). Evaluating LLM-generated worked examples in an introductory programming course. In Proceedings of the 26th Australasian Computing Education Conference (pp. 77–86). ACM. doi:10.1145/3636243.3636252

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, 19(1), 010132. doi:10.1103/PhysRevPhysEducRes.19.010132

McKinney, W. (2011). pandas: A foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14(9), 1–9.

Neumann, A. T., et al. (2024). An LLM-driven chatbot in higher education for databases and information systems. *IEEE Transactions on Education*. doi:10.1109/TE.2024.3467912

OpenAI. (2023). ChatGPT (Apr. 9 version) [Large language model]. Retrieved from https://chat.openai.com/chat

Perković, G., Drobnjak, A., & Botički, I. (2024, May). Hallucinations in LLMs: Understanding and addressing challenges. In 2024 47th MIPRO ICT and Electronics Convention (MIPRO) (pp. 2084–2088). IEEE. doi:10.1109/MIPRO60963.2024.10569238

Razafinirina, M. A., Dimbisoa, W. G., & Mahatody, T. (2024). Pedagogical alignment of large language models (LLM) for personalized learning: A survey, trends and challenges. *Journal of Intelligent Learning Systems and Applications*, 16(4), 448–480. doi:10.4236/jilsa.2024.164023

Searson, M., Langran, E., & Trumble, J. (Eds.). (2024). Exploring new horizons: Generative artificial intelligence and teacher education. Association for the Advancement of Computing in Education (AACE). Retrieved from https://www.learntechlib.org/p/223928/

Stamper, J., Xiao, R., & Hou, X. (2024, July). Enhancing LLM-based feedback: Insights from intelligent tutoring systems and the learning sciences. In International Conference on Artificial Intelligence in Education (pp. 32–43). Springer Nature. doi:10.1007/978-3-031-64315-6_3

Team, G., et al. (2023). Gemini: A family of highly capable multimodal models. *arXiv Preprint*. doi:10.48550/arXiv.2312.11805

Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Vol. 2026 (2) 109 – 117
ISSN 2688-3651 [online]

Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning – The good, the bad and the ugly. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4582–4591). IEEE. doi:10.1109/CVPR.2017.485

Zhang, Z., et al. (2023). Defending large language models against jailbreaking attacks through goal prioritization. *arXiv Preprint*. doi:10.48550/arXiv.2311.09096

Zhang, Z., et al. (2024). Simulating classroom education with LLM-empowered agents. *arXiv Preprint*. doi:10.48550/arXiv.2406.19226