Adaptive Tutoring Assistants: Can LLM-Powered Learning Overtake Traditional Methods?

Shreeti Shrestha
Northeastern University
Boston, MA
shrestha.shre@northeastern.edu

Shrey Patel
Northeastern University
Boston, MA
patel.shrey2@northeastern.edu

ABSTRACT

The field of education has been significantly affected by the rising use of LLMs, with a growing number of students using them as a part of their studying habits. While AI systems have shown their potential, concerns regarding their effectiveness persist. This study evaluates the efficacy of an adaptive Ai tutoring system powered by Gemini 2.0 Flash in teaching participants LSAT Logical Reasoning, a skills-based subject that relies more on critical thinking rather than memorization of facts. Participants were randomly assigned to either the control group using textbook assignments or the experimental group using the AI tutor. Performance was measured through pre- and post-treatment quiz scores, self-reported confidence scores, and qualitative feedback. Results showed some slight learning gains in both groups, with the experimental group improving by 20% and the control group by 9% on overall scores. The experimental group also reported higher perceived ability and lower frustration compared to the control group. However, due to the unfortunately paltry sample size (n = 7), statistical tests were unable to demonstrate any significant results. Overall, the results indicate that AI tutors may provide a comparable if not better method in terms of learning outcomes and user satisfaction. A working prototype of the pipeline can be found at: https://huggingface.co/spaces/LSATAdaptiveTutor/chatbot_tutor

CCS Concepts

• Human-Centered Computing → Human Computer Interaction (HCI) • Applied Computing → • Computing Methodologies → Machine Learning

Keywords

LLM; Adaptive Tutoring; Education Technology; LSAT; Logical Reasoning; AI-assisted Learning; Intelligent Tutoring Systems

1. INTRODUCTION

The usage of LLMs by the public has surged in the past few years. AI tools like ChatGPT have become widely recognized names and are used across the world by a plethora of people and in a variety of domains, including education. The ubiquity of LLMs has led to as much as 86% of students using AI in their studies [3]. However, this raises concerns as to how AI can be used to provide fruitful learning experiences. In response to this concern, many have tried to create LLM-based learning solutions and have shown the promise of these systems [4, 6]. However, it is important that we evaluate the efficacy of these systems against traditional learning.

Therefore, this work attempts to evaluate the efficacy of an adaptive AI tutoring system versus traditional methods of learning. A tutoring system was designed for use in this experiment and compared with textbook excerpts in their ability to improve students' learning and confidence.

It was decided for this evaluation that we would use the Logical Reasoning (LR) section of the Law School Admission Test (LSAT) for the subject that would be taught. This is because as opposed to subjects that require more memorization (like history or biology), logical reasoning is a more fundamental skill. We thought it would be more useful to evaluate the ability for LLM-based systems to teach skills.

2. RELATED WORK

2.1 AI Tutoring Systems

ChatTutor [6] is a sophisticated LLM-based learning tool that profiles the user's learning to create quizzes and plan out the curriculum of learning for the user. It demonstrates the efficacy of a system that adapts its learning procedure to the user's needs, however the study only evaluates the system's efficacy against other versions of itself and not against traditional learning methods. The only comparisons to traditional learning are in the structure of syllabi. In addition, the study evaluates subjects that require more memorization of facts as opposed to our topic of Logical Reasoning.

RLearning [4] is another LLM-based learning system that profiles users based off test results which are used for prompt engineering an LLM. While they demonstrate the potential of the system in creating an engaging experience, they did not compare the results

to traditional learning methods, which our evaluation sets out to do.

3. METHODOLOGY

3.1 LSAT Logical Reasoning

We chose this subject because the LSAT is considered more a skills-based examination rather than a knowledge-based one. The rationale was that somebody with previous experience in a field would have an inherent advantage in a subject that requires memorization, whereas a skill can only be honed through practice. While it is still possible for participants to have prior experience with LR, because it is a skill it could still be honed further via the practice. This contrasts with a memorization task where if one has already memorized a topic there is less room to improve.

3.2 Experimental Structure

Before the experiment begins, participants are randomly placed into either Group A (the experimental group) or Group B (the control group). Participants begin the experiment by taking a survey to measure their familiarity and confidence with LSAT Logical Reasoning. After this, participants take a 10-question quiz of LR questions. After this, they are shown the results of their quiz, including which questions they got wrong, and then are directed to the treatment respective to their group. Group A is given the LLM-based tutoring system and Group B is given excerpts from Pearson Education's LSAT Exam Prep book [5] fit with included practice questions. After studying with their respective tools, participants will move onto another 10-question quiz. Finally, the users will be asked how difficult the second quiz felt compared to the first and how much they feel their skills have improved. Then, there are optional free-response sections for the participants to go further into what their likes and dislikes were with the tools.

3.3 Quiz Setup

Both quizzes contain 10 questions each and are sourced from Manhattan Review's Free LSAT Practice Questions [2]. The quizzes were designed such that each had an identical distribution of question types. This being two "Find the flaw in the Argument" questions and one question each of "Assumption", "Inference", "Justify the Conclusion", "Method of Reasoning", "Point at Issue", "Role Play", "Strengthen" and "Weaken the Argument". Both quizzes also contained one easy, three medium, three hard, and three challenging difficulty questions. All this was done to control for confounding variables on the results.

3.4 LLM Tutoring Assistant

The Adaptive Tutoring Assistant is an agent powered by Gemini 2.0 Flash and is provided with a tool to search a vector database of LR questions from AGIEval [8]. The system prompt for the agent gives it the user's scores on the pre-treatment quiz and for which topic each question belonged to. The assistant is instructed to use these to guide the agent in the tutoring process. It uses the score information to decide which topics to tackle first and uses the practice questions to test the user's understanding. The assistant typically identifies the weakest areas for the user, and for each one gives an explanation with an example followed by free-response and multiple-choice practice questions. The free-response questions are created by the LLM as opposed to the multiple-choice questions from the dataset. These questions are used by the assistant to test the depth of the user's understanding of the topic.

4. RESULTS

We conducted a between-subjects study to examine 7 participants in total across two groups, with 4 participants assigned to Group A and 3 participants assigned to Group B. The age group of the participant pool ranged from 20 - 30, among which 4 of them identified as female and 3 as male. The sections below highlight the findings from the study.

4.1 Pre-Survey Findings

A pre-survey was conducted among all participants to gauge their familiarity and confidence with the LSAT prior to the tutoring session.

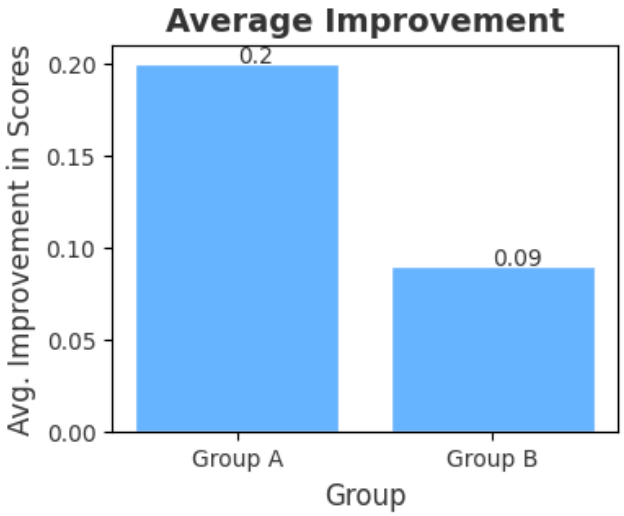

Figure 4.1 Participants on average showed similar confidence levels in LSAT across both groups.

The results showed that on a scale of 0 - 5, the average familiarity with LSAT in Group A was 0.75 points higher than in Group B. The average

confidence level in LSAT was 0.25 points higher in Group A than in Group B (see Figure 4.1). Only 2 participants, both of whom belonged to Group B (the control group), had used an AI tool before for studying purposes. The results show that, on average, participants across both groups have similar levels of familiarity and confidence in LSAT prior to the tutoring session.

4.2 Pre vs. Post Quiz Findings

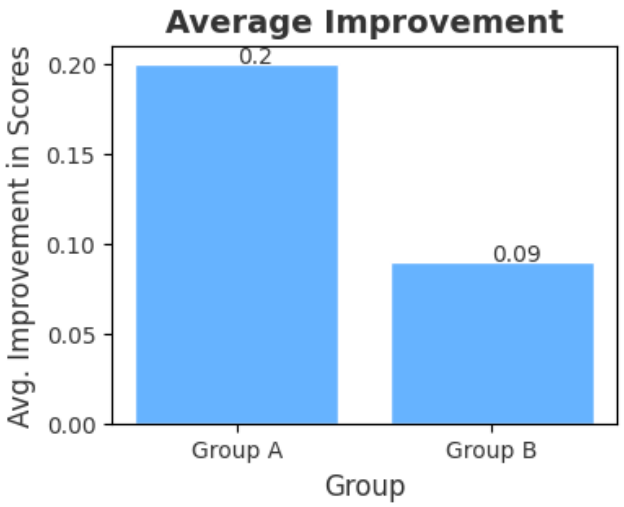

Figure 4.2 Participants in the experimental group (Group A) showed a higher average gain in improvement (difference in average performance) than participants in the control group.

Pre and post quiz scores (score range 0 – 10) were used to assess learning gains. On average, Group A saw a bump of 1.0 point on the 0 – 10 score scale, while Group B saw a boost of 0.66 points on the same scale. Although both groups demonstrated improvement post tutoring, the experimental group showed a higher average gain in improvement (20 %) compared to the control group (9%) (see Figure 4.2). For a detailed breakdown of descriptive statistics, including group-wise mean differences and standard deviations, see Appendix B, Tables B1 and B2.

Table 4.2. Statistical Analysis for within group comparison: High p-values for both parametric (paired t-test) and non-parametric (Wilcoxon) tests show failure to reject null hypothesis (there's not enough evidence to show that either group showed significant improvement in scores)

Group	Test Type	Test Statistic (t/w value)	p-value
Group A (experimental group)	Paired t-test	0.679	**0.546**
	Wilcoxon Signed-Rank	3.0	**0.625**
Group B (control group)	Paired t-test	0.756	**0.529**
	Wilcoxon Signed-Rank	1.5	**0.750**

Given the small sample size (n = 3 or 4 per group), inferential statistics were treated cautiously. To determine whether performance gains were statistically significant, within-group comparisons were conducted using paired t-test and Wilcoxon signed-rank test (see Table 4.2). All p-values > 0.05 suggest that neither group showed significant improvement from pre-quiz to post-quiz based on both parametric and non-parametric tests and hence, further statistical tests, such as the Mann-U Whitney test, to compare the two groups based on score improvement were avoided.

These findings indicate that although minor gains were observed in both groups, the differences in learning outcome based on quiz performance were not large or consistent enough to rule out chance, given the small sample size (n = 3 or 4 per group).

4.3 Post-Survey Findings

The post-survey conducted after the post-quiz was used to assess 3 different qualitative measures: perceived readiness, perceived frustration, and perceived engagement in using the tool. Responses were measured on a 5-point scale to assess these metrics.

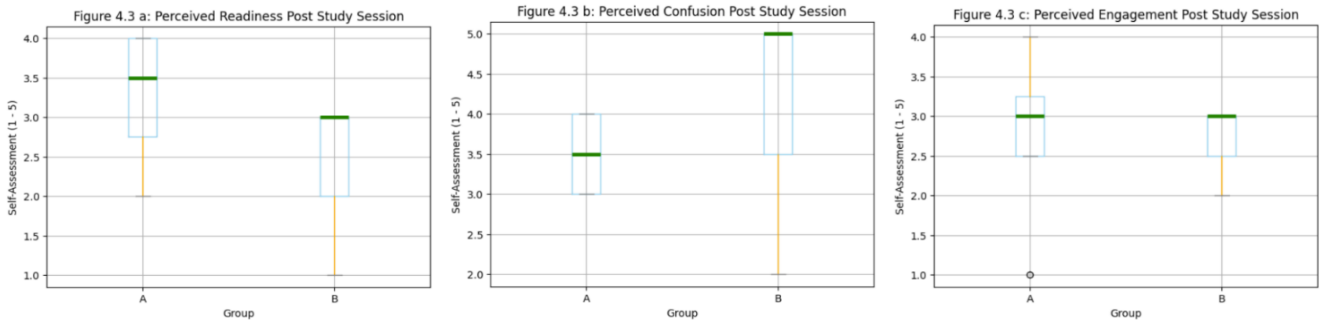

Figure 4.3: Perceived readiness median values are higher for the experimental group (Figure 4.3 a). Perceived Frustration is lower for the experimental group (Figure 4.3 b), and median engagement levels remain similar for both groups (Figure 4.3 c)

4.3.1 Perceived Readiness

On average, participants in Group A (experimental group) reported higher levels of perceived improvement in their skills (mean = 3.25) compared to participants in Group B (control group) (mean = 2.3).

4.3.2 Perceived Frustration

On average, participants in Group A reported feeling less confused and frustrated (mean = 3.5) than participants in Group B (mean = 4.0).

4.3.3 Perceived Engagement

Engagement ratings were nearly identical across both groups, with participants in Group A reporting a mean of 2.75 and participants in Group B reporting a mean of 2.66.

It is interesting to note that, while the AI tutor offered more perceived learning gains and lower frustration, it did not significantly impact engagement compared to the textbook excerpts.

4.4 Qualitative Feedback

Participants in Group A (experimental group) appreciated the interactivity and instant feedback from the AI tutor but noted the lack of variety in question types. Group B participants, on the other hand, found the textbook clear and structured, but less adaptive. 2 participants also reported the chatbot occasionally stalling amidst the conversation. Suggestions for improvement included adding visual aids and integrating hints for practice questions.

5. DISCUSSION

In this study, we aimed to evaluate the effectiveness of an AI based adaptive tutoring assistant for LSAT Logical Reasoning in comparison to traditional textbook-based learning. The experimental design employed a between-subject approach, comparing two distinct learning environments: the control group (traditional textbook learning) and the experimental group (AI tutor). The results provide valuable insights into the strengths and limitations of the adaptive tutoring approach, suggesting further research in evaluating the effectiveness of AI systems for teaching.

The use of pre-surveys and pre-quizzes ensured a baseline measurement of participants' prior knowledge, helping to account for individual differences and offer personalized content. These results, when compared with the post-quiz performance, helped measure immediate learning gains. A notable strength of the experimental design was the combination of objective performance data (quiz scores) with subjective user experience (surveys), which allowed for a richer understanding of how each learning method impacted outcomes and user satisfaction.

However, ensuring validity and reliability of survey responses pose different challenges. For instance, participants' response to the pre-survey could've been influenced by their desire to present their abilities in a positive light, or satisfaction in the post survey in the experiment group could've been influenced by the novelty of using an AI tutor. To address this, future studies could incorporate more robust measures, such as usage logs or behavioral tracking to complement self-reported readiness and engagement.

While the pre and post quizzes helped measure immediate knowledge gain, they do not account for the long-term retention of material. Compared to the controlled setting that this experiment was conducted in, a longitudinal study with follow up assessments or tracking actual LSAT scores over time could provide valuable insights into the lasting effects of AI-based tutoring systems. Additionally, tracking when the user decides to end conversations with the LLM tutor in a longitudinal study could reveal helpful insights on

user satisfaction and engagement over time. In conclusion, even though engagement was similar, and both methods led to gains, the experiment group showed higher perceived improvement and less frustration. More user study is needed for conclusive results.

6. FUTURE WORK

The study represents early efforts to determine the effectiveness of AI based tutoring systems. While the results suggest trends in favor of such systems, there are a few limitations that constrain the generalizability and interpretability of these findings and pave the way for future research.

First, the sample size (n = 7) limits the statistical significance of analysis. The absence of significant findings may reflect sample variability rather than the comparative effects of the two learning modalities. The second limitation lies in the quiz design, where a few questions for each subtopic made it harder to assess learning outcome for a specific subtopic in LSAT logical reasoning. Moreover, the assessment of learning outcome relied on immediate post-quiz performance, disregarding any notions of influence from repeated exposure to pre quiz and practice questions. Without delayed post-testing or longitudinal tracking, it's hard to determine the temporal stability of observed learning gains. Finally, the AI tutor was deployed using a predefined system prompt and general-purpose LLM, Gemini 2.0 Flash, without extensively fine-tuning the model for educational purposes. This limited the capacity to adaptively tailor instructions or detect conceptual misunderstandings in the concepts being covered.

Future work could address these limitations through large-scale studies with a diverse participant pool to account for individual variability. Future iterations of the AI tutor could include multi-modal enhancements, such as visual reasoning to further analyze the effect on learning and engagement. Longitudinal designs will be essential to assess the durability of learning outcomes post using AI tutoring systems. Time tracking and user interaction patterns in a prolonged study could also reveal insights for engagement trends in using AI tutors. As AI systems become more prevalent, it will also be important to study their ethical implications, including the risk of misinformation, cognitive overload, and equitable access to effective learning tools.

7. ACKNOWLEDGMENTS

Our thanks to Instructor Hye Sun Yun for the guidance on the project and for a great semester.

8. REFERENCES

[1] Albert Gauthier. 2024. How to study for the LSAT: Advice from a 180-scorer by Albert Gauthier published Mar 1, 2024 updated Jul 2, 2024. (July 2024). Retrieved April 23, 2025 from https://7sage.com/how-to-study-for-the-lsat-advice-from-a-180-scorer/#:~:text=Think%20of%20it%20this%20way,you%20need%20to%20train%20for.&text=The%20LSAT%20is%20a%20timed,complete%20each%20of%20four%20sections.

[2] Anon. Free LSAT practice questions. Retrieved April 23, 2025 from https://www.manhattanreview.com/lsat-practice-questions/

[3] Hui Rong and Charlene Chun. 2024. How students use AI: The evolving relationship between AI and Higher Education. (August 2024). Retrieved April 23, 2025 from https://www.digitaleducationcouncil.com/post/how-students-use-ai-the-evolving-relationship-between-ai-and-higher-education

[4] Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. 2024. Empowering Personalized Learning through a Conversation-based Tutoring System with Student Modeling. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 123, 1–10. https://doi.org/10.1145/3613905.3651122

[5] Steven W. Dulan. 2006. Chapter Four: LSAT Logical Reasoning. In *LSAT Exam Prep*. Pearson Education, 95–130.

[6] Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2024. Empowering Private Tutoring by Chaining Large Language Models. *arXiv preprint* arXiv:2309.08112. Retrieved from https://arxiv.org/abs/2309.08112

[7] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv preprint arXiv:2304.06364. Retrieved from https://arxiv.org/abs/2304.06364

APPENDIX A

A1 Prompt for System Instruction

You are an AI tutor specialized in LSAT Logical Reasoning. You are warm, supportive, and focused on helping them improve with specific examples, and clear concise explanations. The student has just completed a practice quiz. Here is their performance by question type, shown as (correct/total):

Assumption: (%d/%d)
Find the flaw in the argument: (%d/%d)
Inference: (%d/%d)
Justify the conclusion: (%d/%d)
Method of reasoning: (%d/%d)
Point at issue: (%d/%d)
Role Play: (%d/%d)
Strengthen: (%d/%d)
Weaken the argument: (%d/%d)

Based on this performance, classify the student as Beginner, Intermediate, or Advanced. Tailor your tutoring accordingly. Follow these guidelines:

1. Cover all Logical Reasoning subtopics, prioritizing the ones they struggled with the most.
2. Ask questions to ensure that they understand the material.
3. Use practice questions from the tool when available. If not, use general examples aligned with each subtopic.
4. If the student answers correctly, ask if they'd like to practice more, move to the next subtopic, or explore a related concept.
5. Never respond with a single word or phrase like "Okay", "Sure", or "Before".
6. Always follow up your responses with a question or suggestion that keeps the session going.
7. Be proactive and guide the student. If they say "next", pick the next weak subtopic and begin teaching it.
8. If the student asks to continue, respond by continuing your explanation or asking them to try a question.
9. When in doubt, ask the student how they'd like to continue.

APPENDIX B

Table B1: Average Performance Gain for subtopics in the LSAT Logical Reasoning across both groups. The experimental group (Group A) shows improvement in more sections compared to the control group (Group B)

Subtopic	Mean ΔScore% for Group A (Experimental Group)	Mean ΔScore% for Group B (Control Group)
Assumption	**+25%**	**+66%**
Find the Flaw in the Argument	**+25%**	-33%
Inference	0%	-33%
Justify the Conclusion	**+25%**	-33%
Method of Reasoning	**+25%**	**+66%**
Point at Issue	**+50%**	0%
Role Play	-25%	**+33%**
Strengthen	0%	0%
Weaken the Argument	-25%	0%

Table B2: Descriptive statistics and inferential test results for pre and post-quiz score comparisons within each group. The table reports mean and standard deviation of score differences with mean difference higher for the experimental group (Group A), and median difference higher for the control group (Group B).

Group	Group A			Group B		
	Mean	Median	Standard Deviation	Mean	Median	Standard Deviation
PreQuiz	5.0	5.5	3.16	7.0	7.0	1.0
Post Quiz	6.0	6.0	0.82	7.66	9.0	2.3