# Balancing Speed and Precision: A Comparative Analysis of Lightweight LLMs on SAT Reading and Writing Tasks

Zixuan Shang[1]*, Jesus Valdiviezo[2,3,4]

[1] Tesla STEM High School, Redmond, WA
[2] Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[3] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02215, USA
[4] Sección Química, Departamento de Ciencias, Pontificia Universidad Católica del Perú, San Miguel, Lima 15088, Peru
*Corresponding author: zixuanshang7819@gmail.com

## Abstract

This study evaluated six cost-efficient large language models (LLMs)—ChatGPT 4.1 mini, Gemini 2.0 Flash, Qwen3 235B-A22B, Llama 3.3 70B Instruct, Claude 3.5 Haiku, and DeepSeek V3—on SAT Reading and Writing multiple-choice questions. Using a structured pipeline with LangChain, we assessed 90 questions across difficulty levels (easy, medium, hard) and skill subdivisions (Craft and Structure, Expression of Ideas, Information and Ideas, Standard English Conventions). The LLMs were not tested on Command of Evidence questions that included a graphical representation of data. Key findings reveal **ChatGPT 4.1 mini and DeepSeek V3 as top performers (91.1% accuracy)**, closely followed by Gemini 2.0 Flash (88.9%), with Qwen3 235B-A22B lagging significantly (32.2%). Accuracy declined with question difficulty (e.g., ChatGPT 4.1 mini dropped from 96.7% on easy to 83.3% on hard questions), and all models struggled most with **Standard English Conventions** (16.7–72.2% accuracy), particularly grammar tasks like boundaries (44.4% average accuracy). While **Gemini 2.0 Flash delivered optimal speed-accuracy balance (88.9% accuracy in 94.51 seconds)**, DeepSeek V3 matched ChatGPT's precision (91.1%) at half the latency (221.84s vs. 184.5s). Models demonstrated moderate-to-high consistency (variability = 1.00–1.39). These results suggest smaller LLMs are viable for automated SAT-style assessments in comprehension (Information and Ideas: 96.3% accuracy) and analysis (Craft and Structure: 100% for top models) but require urgent improvements in grammatical precision and complex reasoning. Educators and developers should leverage ChatGPT 4.1 mini or DeepSeek V3 for high-accuracy feedback while reserving Gemini 2.0 Flash for rapid-response applications, with caution for grammar-focused tasks.

## Backgrounds

Previous research has explored the ability of LLMs to perform various standardized assessments, including the Graduate Record Examinations, the Law School Admission Test, and AP exams [2]. Studies have shown that state-of-the-art models such as ChatGPT-4 achieve human-like scores on some standardized tests, particularly those emphasizing logical reasoning and factual recall. However, systematic evaluations on SAT questions have not used the latest models [6].

Existing work tends to either evaluate LLMs on broad language comprehension tasks or focus on mathematical reasoning, leaving a gap in understanding how well these models handle identifying rhetorical purpose, understanding textual evidence, and applying grammar rules within context.

This research aims to fill this gap by systematically evaluating six selected LLMs—Llama 3.3 70B Instruct, Qwen3 235B-A22B, ChatGPT 4.1 mini, and Gemini 2.0 Flash—on SAT Reading and Writing multiple-choice questions. These models were chosen for their accessibility, cost-efficiency, and competitive performance. To ensure a rigorous comparison, each model was presented with an identical set of SAT questions sourced from the College Board, covering multiple skill areas and difficulty levels. The study excludes questions that require graphical interpretation, focusing solely on text-based comprehension and reasoning.
Each model was given a standardized prompt to minimize variability in response behavior. For example, prompts with extra examples were tested but did not improve accuracy, and some prompt variations led to inconsistencies in responses.
The models' accuracy is measured by comparing their chosen answers to the official correct answers. Evaluations were conducted over 3 runs, and the mode of correct answers was selected.
This study provides valuable insight into the current capabilities and limitations of LLMs in standardized test evaluation by identifying which model achieves the highest accuracy.
The central research question guiding this study is: Out of the six selected large language models that are comparatively cheap and fast, which one answers Reading and Writing SAT multiple-choice questions with the highest accuracy?
The findings will contribute to the growing body of research on LLM evaluation, informing educators, researchers, and developers about the reliability of these models for academic assessments.

## Methods

To evaluate the performance of various large language models (LLMs) on SAT Reading and Writing questions, we implemented a structured pipeline using LangChain, a framework for building LLM applications. The methodology consisted of model setup, data preparation, prompt engineering, and systematic evaluation to ensure fair and consistent comparisons.

### Model Selection and Setup

We selected six cost-efficient LLMs with fewer than 80 billion parameters to balance performance and computational expense (this evaluation was completed on May 17, 2025, when Claude 4.0 did not exist and Gemini 2.5-Flash was not available): ChatGPT 4.1 mini – Developed by OpenAI, Gemini 2.0 Flash – Developed by Google, Llama 3.3 70B Instruct – Developed by Meta, Qwen3 235B-A22B – Developed by Alibaba, DeepSeek V3 – Developed by DeepSeek, and Claude 3.5 Haiku – Developed by Anthropic.
Each model was initialized using their respective API providers (OpenAI, Google, Fireworks AI, and Anthropic via VertexAI). API keys were configured to authenticate access, and necessary libraries were installed to facilitate model interaction (*Installation, API Keys, Libraries*).

### Data Preparation

We sourced SAT Reading and Writing questions from the College Board [1], formatted as a pipe-separated CSV file stored on GitHub [5]. The dataset included: question text, multiple-choice options (A-D), correct answer, difficulty level (easy, medium, hard), and question type (e.g., Standard English Conventions, Information and Ideas). The data was loaded into a pandas DataFrame for structured processing (*Import Questions*). Each question was converted into a standardized query combining the question text and answer choices (*Constructing Query*).

**Prompt Engineering**
To ensure consistent responses, we designed a strictly formatted prompt that: instructed the model to analyze the question and select the best answer, restricted output to only the letter (A, B, C, or D) of the correct choice, and prevented explanations or additional text, reducing variability.

The final prompt used for evaluation (Figure 1) is denoted below.

```
"""
### Instruction:
You are a reasoning assistant.

1. Analyze the given context thoroughly.
2. Identify the best option from the provided choices.

### Response Format:
- Provide **only** the letter corresponding to the correct answer (e.g., A, B, C, or D).
- Strictly avoid additional text, explanations, or context in your response.

EXAMPLE INPUT
 Dolores Huerta's advocacy on behalf of farmworkers was rooted in her experience as a schoolteacher in Stockton,
 California, in the early 1950s. Hoping to help her students and their families outside the _____ Huerta left teaching to
 start the Stockton chapter of the Community Service Organization, a group focused on the needs of local farmworkers.

 Which choice completes the text so that it conforms to the conventions of Standard English?
 A. classroom.
 B. classroom;
 C. classroom,
 D. classroom

 EXAMPLE OUTPUT
 C

{input}
"""
```

Figure 1: Finalized prompt for LLM-inference model

**Evaluation Pipeline**
Using LangChain's RunnablePassthrough, we constructed a processing chain that passed the formatted question to the LLM, generated a response under controlled settings (temperature = 0 for deterministic outputs) and parsed the output to extract the predicted answer. Each question was evaluated three times per model to assess consistency. If a model failed to respond, the system retried up to three times before logging an error (*SAT Question Evaluation System*).

**Performance Metrics**
We measured accuracy (percentage of correct answers per difficulty level and question type), variability (consistency across multiple runs where lower values indicate more stable responses, and speed (time taken to process all questions). Results were aggregated into pivot tables, comparing performance across models (*SAT Question Evaluation System*).
The diagram of the workflow is included in the appendix.
We employed ChatGPT and DeepSeek during the writing part of the document [3, 4].

The code used to evaluate the SAT questions with the small LLM models can be found on GitHub:
**https://colab.research.google.com/drive/1eYo1zXpqSJoZG7ucztsDhSXQONNUXa9o?usp=sharing**.
The text in parentheses and italics after each step of this section corresponds to the section titles on GitHub.

## Results:
Average Variability = the average number of different answers the model gave for a specific set of questions (this number would range from 1 to 3 as each model was asked the same question three times; the closer the number is to 1, the more consistent the model was in its 3 answers)

Table 1: Performance Metrics for the LLMs Studied Separated by Difficulty.

| | Easy | | Medium | | Hard | | Total | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | Variability | Accuracy | Variability | Accuracy | Variability | Accuracy | Variability |
| Claude 3.5 Haiku | 100.0% | 1.03 | 70.0% | 1.23 | 63.3% | 1.20 | 77.8% | 1.16 |
| ChatGPT 4.1 mini | 96.7% | 1.03 | 93.3% | 1.07 | 83.3% | 1.10 | 91.1% | 1.07 |
| Gemini 2.0 flash | 96.7% | 1.00 | 86.7% | 1.03 | 83.3% | 1.00 | 88.9% | 1.01 |
| Llama 3.3 70B Instruct | 93.3% | 1.00 | 90.0% | 1.00 | 80.0% | 1.00 | 87.8% | 1.00 |
| Qwen3 235B-A22B | 30.0% | 1.00 | 33.3% | 1.00 | 33.3% | 1.00 | 32.2% | 1.00 |
| DeepSeek V3 | 100.0% | 1.00 | 93.3% | 1.13 | 80.0% | 1.17 | 91.1% | 1.10 |

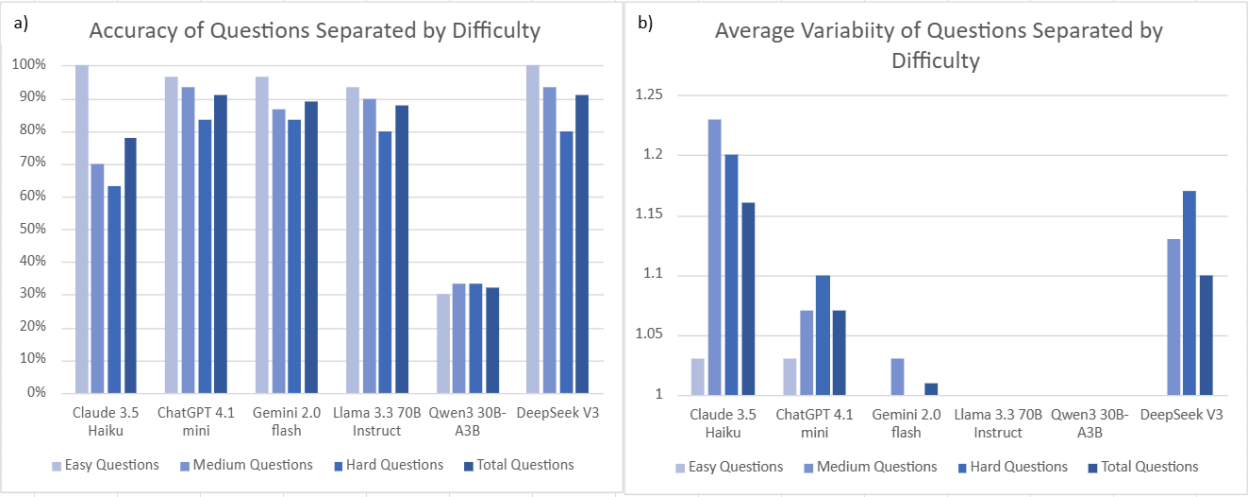


Figure 2: Performance Metrics for the LLMs Studied Separated by Difficulty. a) Accuracy. b) Average Variability.

**Skill subdivisions:**
Standard English Conventions: ["boundaries", "form, structure, and sense"]
Information and Ideas: ["central ideas and details", "inferences", "command of evidence"]
Craft and Structure: ["words in context", "text structure and purpose", "cross-text connections"]
Expression of Ideas: ["rhetorical synthesis", "transitions"]

Table 2: Performance Metrics for the Top 3 Performing LLMs Studied Separated by Type of Question/Skill.

| | Craft and Structure | | Expression of Ideas | | Information and Ideas | | Standard English Conventions | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Variability | Accuracy | Variability | Accuracy | Variability | Accuracy | Variability | Accuracy | Variability |
| ChatGPT 4.1 mini | 100.0% | 1.04 | 83.3% | 1.06 | 96.3% | 1.00 | 77.8% | 1.22 | 91.1% | 1.07 |
| DeepSeek V3 | 100.0% | 1.00 | 88.9% | 1.11 | 96.3% | 1.00 | 72.2% | 1.39 | 91.1% | 1.10 |
| Gemini 2.0 flash | 100.0% | 1.00 | 88.9% | 1.00 | 96.3% | 1.00 | 61.1% | 1.06 | 88.9% | 1.01 |

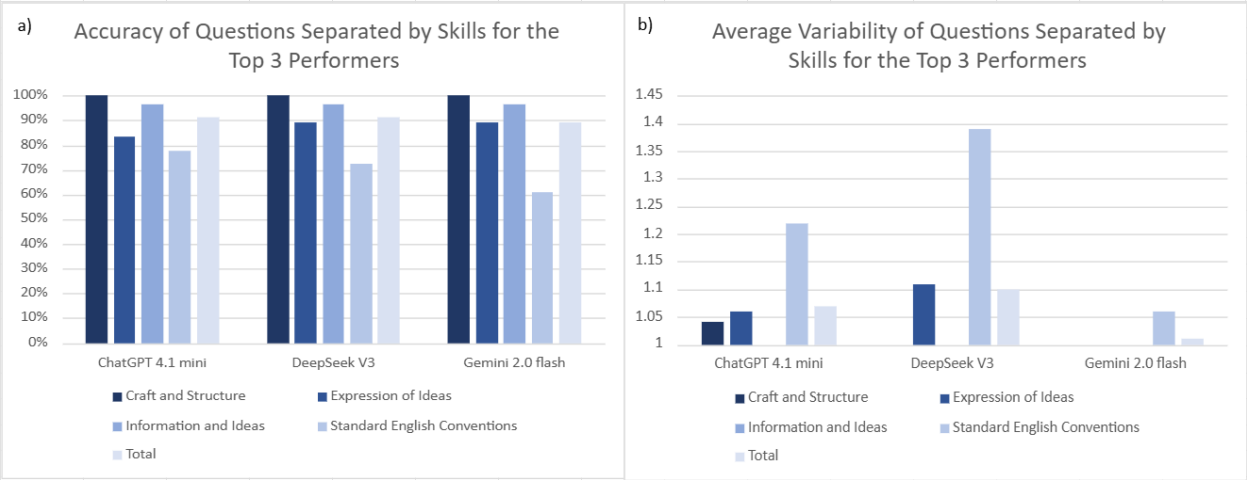


Figure 3: Performance Metrics for the Top 3 Performing LLMs Studied Separated by Type of Question/Skill. a) Accuracy. b) Average Variability.

Table 3: Evaluation Time for the LLMs studied.

| Model | Evaluation Time for all 90 Questions (Seconds) |
|---|---|
| Claude 3.5 Haiku | 319.50 |
| ChatGPT 4.1 mini | 184.50 |
| Gemini 2.0 flash | 94.51 |
| Llama 3.3 70B Instruct | 843.92 |
| Qwen3 235B-A22B | 2750.02 |
| DeepSeek V3 | 221.84 |

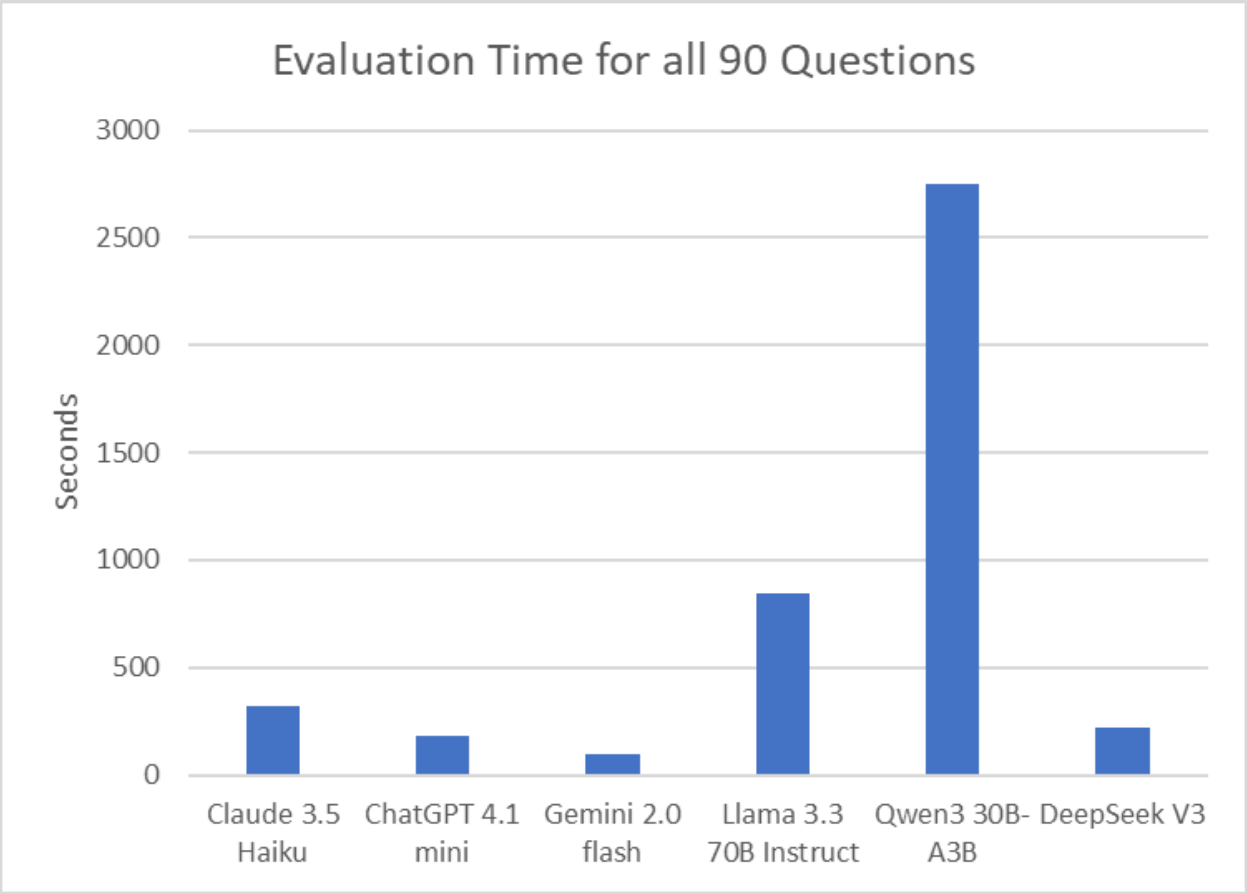


Figure 4: Evaluation Time for the LLMs studied.

Table 4: 10 Lowest Accuracy Questions Across All Models (All questions could be found in this GitHub link: https://github.com/1082098-LWSD/SAT-question-evaluation.git):

| Question ID | Difficulty | Skill | Correct Answers | Total Answers | Accuracy (%) |
|---|---|---|---|---|---|
| eb95235b | hard | Boundaries (Standard English Conventions) | 0 | 6 | 0.0 |
| a2816c7f | hard | Form, structure, and sense (Standard English Conventions) | 0 | 6 | 0.0 |
| 702eb7e3 | hard | Command of evidence | 0 | 6 | 0.0 |

| | | | | | |
|---|---|---|---|---|---|
| | | (Information and Ideas) | | | |
| 74ce2f05 | medium | Boundaries (Standard English Conventions) | 1 | 6 | 16.7 |
| 89fbc3eb | medium | Boundaries (Standard English Conventions) | 1 | 6 | 16.7 |
| e3edc138 | hard | Transitions (Expression of Ideas) | 1 | 6 | 16.7 |
| adf210e7 | hard | Boundaries (Standard English Conventions) | 2 | 6 | 33.3 |
| a7c85001 | easy | Boundaries (Standard English Conventions) | 2 | 6 | 33.3 |
| b46e0c8a | medium | Rhetorical synthesis (Expression of Ideas) | 3 | 6 | 50.0 |
| 1ee7b429 | medium | Form, structure, and sense (Standard English Conventions) | 3 | 6 | 50.0 |

The latest evaluation of six language models on SAT Reading and Writing questions highlights key trends in accuracy, speed, and consistency. ChatGPT 4.1 mini and DeepSeek V3 tied as top performers (91.1% accuracy), followed by Gemini 2.0 Flash (88.9%) and Llama 3.3 70B Instruct (87.8%). Claude 3.5 Haiku (77.8%) showed moderate performance, while Qwen3 235B-A22B (32.2%) remained severely deficient.

**Performance by Difficulty Level**

All models exhibited a clear trend: accuracy decreased as question difficulty increased, though top models demonstrated resilience. ChatGPT 4.1 mini and Gemini 2.0 flash maintained 83.3%

accuracy on hard questions, while DeepSeek V3 achieved 80.0%. Qwen3 235B-A22B failed across all levels (30.0–33.3%), indicating systemic flaws in foundational skills. This suggests that smaller LLMs handle straightforward questions well but struggle with more complex reasoning tasks.

**Performance by Skill Subdivision**
Most models excelled in *Information and Ideas* (96.3% accuracy for top models) and *Craft and Structure* (100% for ChatGPT, Gemini, DeepSeek). *Standard English Conventions* remained challenging, particularly boundaries (44.4% accuracy for Claude 3.5 Haiku) and form, structure, and sense (55.6% for Claude). Even top models like DeepSeek V3 scored only 72.2% in this subdivision. This suggests that while LLMs excel at understanding content, they still struggle with fine-grained grammatical rules.

**Consistency (Variability)**
Gemini 2.0 Flash (avg_variability = 1.01) and Llama 3.3 70B Instruct (1.00) showed stable outputs, even on incorrect answers. Claude 3.5 Haiku (1.16), ChatGPT 4.1 mini, and DeepSeek V3 (1.10) exhibited slight fluctuations, particularly in grammar tasks (*boundaries* variability = 1.44 for Claude). Qwen3 235B-A22B's perfect consistency (1.00) paired with low accuracy (32.2%) suggests rigid, repeatable mistakes.

**Performance vs. Speed Trade-off**
Gemini 2.0 Flash completed evaluations in 94.51 seconds (fastest) with strong accuracy (88.9%), ideal for real-time applications. ChatGPT 4.1 mini (91.1%) and DeepSeek V3 (91.1%) traded slightly higher processing times (184.5s and 221.84s, respectively) for precision. Qwen3 235B-A22B was both the slowest (2,750.02s) and least accurate (32.2%), rendering it impractical. The other models (Claude 3.5 Haiku and Llama 3.3 70B Instruct) were slower than the top three (319.50s and 843.92s, respectively) and less accurate (77.8% and 87.8%, respectively), reinforcing that the top three models offer the best balance for practical use.

**Implications for Real-World Use**
Gemini 2.0 Flash is optimal for instant feedback for a large number of questions (e.g., classroom tools). ChatGPT 4.1 mini and DeepSeek V3 are preferable for scoring high-stakes assessments such as SAT prep or detailed scoring. All models underperformed on *Standard English Conventions*, necessitating human oversight for grammar-focused tasks.

**Prompt Adjustments**
Initial tests with example-based prompts led to errors, as models sometimes repeated the example answer instead of responding to the actual question. Detailed, multi-step prompts also caused issues, as models occasionally ignored the question due to memory constraints. The final standardized prompt minimized these problems, ensuring fair comparisons.

**Lowest Accuracy Questions**
The 10 least accurate questions (0–50% accuracy) predominantly tested *Standard English Conventions.* There were zero correct answers for *boundaries* questions, which involved identifying correct punctuation/clauses (e.g., *Question eb95235b*). Models also performed poorly on *form, structure, and sense* questions that involved selecting grammatically correct verbs

(e.g., *Question a2816c7f*). *Transition* questions (e.g., *Question e3edc138*) challenged even top models (16.7% accuracy).

## Conclusions

This updated analysis confirms ChatGPT 4.1 mini and DeepSeek V3 (91.1% accuracy) as premier models for SAT-style questions, with Gemini 2.0 Flash (88.9%) leading at speed. Key findings include:

*Standard English Conventions* remains the weakest area (44.4 -- 72.2% accuracy), highlighting LLMs' unresolved challenges with syntactic rules and grammar. Accuracy consistently lowers on hard questions, which signals ongoing struggles with high-difficulty tasks. While Gemini 2.0 Flash and Llama 3.3 70B Instruct delivered stable outputs, Qwen3 235B-A22B's rigid errors (1.00 variability) necessitate algorithmic revisions.

Assuming unlimited time and money, this research could be furthered by evaluating LLMs' ability to score other types of questions and exams. LLMs could be leveraged to evaluate writing performance for GRE, AP exams, and other standardized tests that include a writing portion. The student's response could be scanned and fed to the LLMs as an image where the LLMs would parse out the student's response based solely on the image. The evaluation could also be broadened to include questions involving interpreting graphical data, which was omitted from this evaluation due to the chosen LLMs' limited ability to correctly obtain information from an image. LLMs could also be evaluated on their ability to score oral presentations with and without slides, where potential bias from a human scorer could be eliminated. LLMs would need to be able to recognize body movement as well as eye contact in real time to effectively score the student on their presentation skills as well as the content of the presentation. Scoring mathematical reasoning is another area where LLMs could be evaluated. LLMs could be trained to identify steps where the student made a mistake and offer a detailed and personalized explanation unique to each student, which is often not available from human teachers or tutors due to time constraints.

For educators, these results reinforce using ChatGPT 4.1 mini or DeepSeek V3 for high-accuracy needs but caution against relying on LLMs for grammar instruction. Future work should prioritize improving grammatical precision (e.g., *boundaries* questions) and complex reasoning, using the lowest-accuracy examples as benchmarks. Overall, while smaller LLMs are not yet perfect for high-stakes SAT grading, their strong performance on easier and medium-level questions makes them viable for tutoring, practice tests, and preliminary feedback, provided their limitations in grammar and advanced reasoning are acknowledged.

## References

[1] College Board. "SAT Suite Question Bank | ..." *Satsuitequestionbank.collegeboard.org*, satsuitequestionbank.collegeboard.org/.

[2] J.D. Capelouto. "Here's How GPT-4 Scored on the GRE, LSAT, AP English, and Other Exams." *Semafor.com*, 15 Mar. 2023, www.semafor.com/article/03/15/2023/how-gpt-4-performed-in-academic-exams?utm_source=chatgpt.com. Accessed 9 Feb. 2025.

[3] Liang, Wenfeng. "DeepSeek." Deepseek.com, Dec. 2023, deepseek.com.

[4] OpenAI. "ChatGPT." ChatGPT, OpenAI, 30 Nov. 2022, chatgpt.com/.

[5] Shang, Zixuan. "GitHub - 1082098-LWSD/SAT-Question-Evaluation." GitHub, 2024, github.com/1082098-LWSD/SAT-question-evaluation.git.

[6] Suresh, Vikram, and Saannidhya Rawat. “GPT Takes the SAT: Tracing Changes InTest Difficulty and Students’ Math Performance.” *Arxiv.org*, 2023, arxiv.org/html/2409.10750v1. Accessed 3 May 2025.

## Appendix:

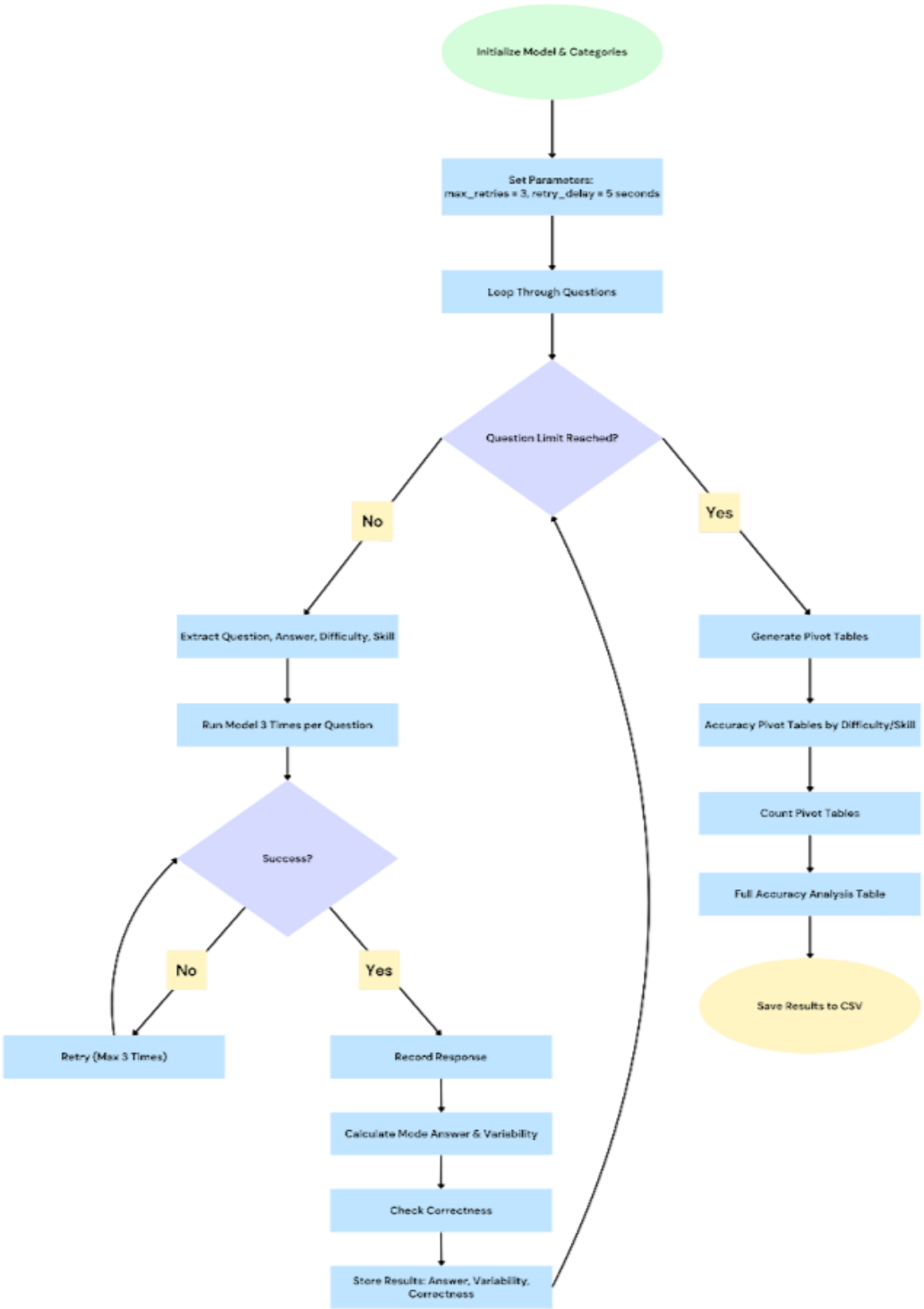


Figure 5: Flowchart of the Evaluation Pipeline