Testing the Future: Open Source vs. Proprietary AI

Final Project Report

Authors:
Shizra Tariq
Kevin Dai
Vivan Nguyen
Niketha Sabesan

Date:
April 29, 2025

Executive Summary

Artificial Intelligence (AI) is advancing rapidly, with open-source models like DeepSeek-R1 competing with or even outperforming proprietary ones like ChatGPT-4 (DeepSeek-AI et al., 2024). While open-source AI offers cost-effectiveness and accessibility, its real-world performance compared to proprietary models remains unclear. Businesses, researchers, and policymakers need a comprehensive understanding of the strengths and weaknesses of both approaches to make informed decisions. This project aimed to systematically compare the latest open-source and proprietary AI models across various real-world tasks, such as conducting literature reviews for doctors, thus improving our understanding of AI and its role in jobs, businesses, and our society.

To address this issue, this project evaluated the performance of open-source AI models against proprietary models using real-world test cases sourced from University of Minnesota professors across a variety of subjects, such as math, English, history, and medicine. The tests covered natural language processing, image recognition, and other practical tasks. The results were benchmarked against existing studies, and a comprehensive report was prepared, summarizing the methods, test results, and comparative analysis. The results allowed stakeholders to determine the best types of models for their needs.

One of the first deliverables is the Benchmark Comparison Dataset Report which described how the AIs were tested in each subject category. It also discussed the methods used to rate how well the models performed. A Performance Analysis Report provided a detailed analysis of each model's real-world applicability, thus allowing potential stakeholders to understand how each model would best fit their needs. Finally, a Presentation and Comparative Report outlined the strengths and weaknesses of both AI types.

Key findings showed that closed-source models outperformed open-source models in accuracy, relevance, and speed, while open-source models offered greater affordability and openness but suffered from slower response times and server congestion.

Overall, the project plan had three main stages. The first stage mainly focused on the project management tasks such as completing the Statement of Work and Work Breakdown Structure. The second stage involved acquiring all of the questions that would be tested on the AIs as well as background research on benchmarks that could be used to rate the results. The second stage then moved on to testing the questions on every AI. The final stage finalized the testing of the AIs then moved on to analyzing the results and conducting a seminar for the Gopher AI Club. After presenting the results, the final stage shifted into the formal handoff of the project.

The project had several key risks, such as delays in test case submission and the rapid evolution of AI technology. While most of these risks didn't occur, the team did encounter the delay in test case submission risk. This delay impacted the critical path, but the team crashed tasks to recover time. In the end, the team successfully finished testing the AIs, summarizing the results, and presenting the seminar for the Gopher AI Club.

The Gopher AI Club seminar served as a key outreach deliverable, allowing us to present our findings, engage with students, and gather feedback on the implications of AI model performance in real-world academic settings. The project taught us the importance of adaptability, planning, and efficiency–from handling the unexpected risks from our testing models and delays in test-case collections. We managed to successfully adapt to the changes and stick with our contingency plans in order to have a successful presentation in the end. We were able to learn from this and deliver our completed and final product despite these constraints.

Contents

1 Project Overview/ Problem Opportunity Statement

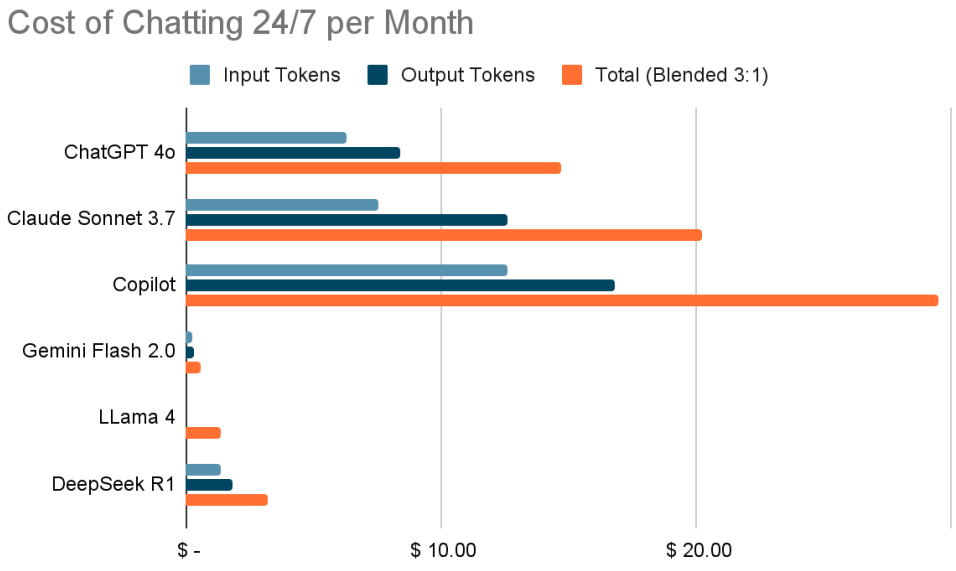
1.1 Overview Statement

The goal of the research project *Testing the Future: Open-Source vs. Proprietary AI* is to evaluate the effectiveness of emerging open-source systems against their proprietary counterparts, like GPT-4. We then wanted to give a presentation to showcase this research in order for other students to gain insights and to hopefully take this knowledge into the development of our future of AI. Given the speed at which open-source rivals like DeepSeek-V3 are developing, the activity aimed to assess how effectively they handle actual academic and professional problems. The original plan was to create a set of standardized questions, test models against benchmarks, and provide comparison findings. To provide a less complicated evaluation framework without deviating from the project's core objectives, the scope was trimmed to fit within practical limitations such as limited API access, token restrictions, and tardy data accumulation.

1.2 Problem Statement

Currently, when considering an AI model to use, most companies will resort to using some form of proprietary AI, with 79% of organizations using ChatGPT 3.5 in their cloud environment (Kaduri & Stansfield, 2024). Taking a conservative estimate of an enterprise subscription cost for a ChatGPT of $40 per user/month (Howarth, 2025) and a conservative fraction of a company's size, a company like Medtronic could spend up to $4.8 million (10,000 users × $40 × 12 months) per year on AI usage. If a company decides to not subscribe to an enterprise key and instead chooses to use an Application Programming Interface (API) key instead, the company would incur the costs seen in **Figure 1** per user.

Figure 1: Cost of Chatting 24/7 per Month



Calculated at 60 words per hour, 1.3 tokens per word, 24 hours every day, blended 3:1 input-output ratio.

When comparing these costs to the cost of open-sourced AIs, it is clear that in general, proprietary models cost much more than open-source models. Therefore, it is important to consider whether that money could have been saved for other uses. By systematically comparing the performances of different open-sourced AI models against proprietary models, stakeholders will be able to better understand the models that may suit them best.

In terms of how researchers and policymakers may benefit from this project, understanding the performances of each model will be important for how they do their work in the future. For researchers, utilizing AI to solve complex problems or conduct literature reviews will allow for faster results and a different perspective. Choosing the best model to suit their needs would help reduce any common errors that are seen across AI models. As for policymakers, they need to consider the costs of building the infrastructure for AIs along with training costs. For example, the cost of training ChatGPT-4 on the newer H100 GPUs was around $30 million, while the cost of training Deepseek-V3 was about $6 million (Komatsuzaki, 2024). Considering how President Trump announced a $500 billion investment in AI Infrastructure (Holland, 2025), understanding the costs of development for each type of model becomes important.

Our presentation was at the Gopher AI Club at the University of Minnesota. Our audience was students who are interested in AI and may be working towards a degree that may involve AI in the future. These students, who could potentially be future business persons, researchers, or policymakers, can understand the importance of our presentation at the earlier stages and utilize these insights for the future of our society.

1.3 Opportunity Statement

Given how quickly the field of artificial intelligence is developing, open-source models currently offer a compelling opportunity to democratize access to state-of-the-art AI solutions. By systematically evaluating and comparing the performance of open-source AI models against proprietary models across various real-world tasks, this study highlights the practical, cost-effective alternatives that may be applied in industry, training, and research. This gives businesses, researchers, and policymakers a better understanding of how open-source AI may boost innovation, decrease dependency on commercial providers, and increase productivity—all of which are especially helpful in setups with limited funds.

2 Approach and Plan

2.1 High-Level WBS

A Work Breakdown Structure (WBS), which separates the project into discrete phases, defines the top-level structure of our project. The whole WBS can be referenced in **Figure 10.** Each team member received a defined task, and precise deadlines with buffer times to provide for unforeseen circumstances were supplied. The following provides an overview of the WBS which was the planned approach to the project.

Stage 1: Background Project Management and Project Planning

This stage mainly focused on defining the overall structure of the project and creating the supporting documents related to project management. It also involved acquiring the basic resources

required to complete the project. These tasks continued throughout the entire project. These tasks include:

- Planning & Scheduling
  - Define project scope and deliverables
  - Develop Work Breakdown Structure (WBS)
  - Risk management planning
  - Create a resource allocation strategy
- Outreach & Coordination
  - Sponsor and professor outreach
  - Confirm sponsors and seminar date
  - Collect and classify test cases

Stage 2: Setting up Testing and Testing

This stage mainly focused on developing the benchmarks for the testing and going through with testing.

- Design & Development
  - Select AI models
  - Develop and refine the testing framework
  - Define benchmarks for performance comparison
- Testing & Analysis
  - Conduct AI model testing (stress testing, edge cases, scalability, bias analysis, etc.)
  - Evaluate model performance
  - Compare with benchmarks

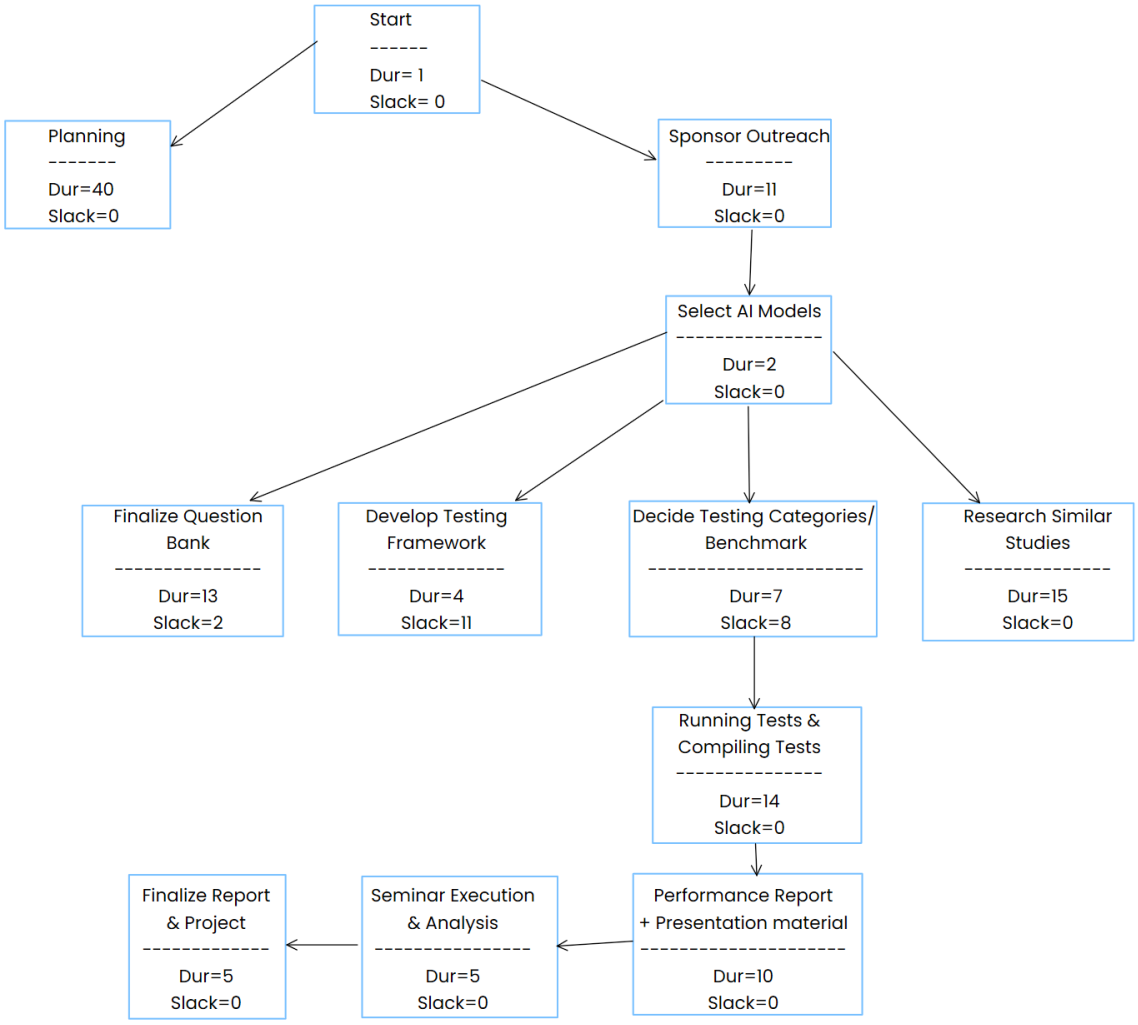
Stage 3: Finalizing results and Delivering Product

This final stage involved finalizing the testing results and compiling them into an understandable format. These results were used to summarize the findings in the seminar which was the product deliverable for Gopher AI Club.

- Documentation & Reporting
  - Compile research and testing results
  - Draft and finalize performance analysis report
  - Create data visuals for findings
- Final Seminar & Project Closure
  - Prepare final presentation materials

- Conduct seminar and present results
- Gather feedback and integrate improvements
- Submit the final project report and poster

Figure 2: Critical Path and Slack



Total Number of Days to Complete the Project: 62

2.2 *Planned vs Actual*

When following this approach in practice, we were able to complete every task on time up until the collection of the test cases. This was mainly due to a lack of urgency we felt as a team towards this project. Many of the tasks defined in the WBS were tasks that could be crashed without any consequence. Additionally, some professors were quite slow in responding to our request for questions that could be used as test cases. As such, we had to significantly decrease the amount of time we could dedicate to testing and compiling the results.

Another issue we ran into was with the benchmarks we would test the AIs with. Given how we have limited knowledge of how AIs are tested and how they are made, many of the benchmarks found through background research were too intricate for us to give accurate results. To sidestep these limitations, we limited the scope of the project to one where the results made sense from the perspective of college students. Rather than conducting complicated tests like scalability and adversarial attacks, we focused on accuracy, relevance, completeness, conciseness, and bias.

One final deviation from the original plan was with how the AIs were to be tested. In the original plan, we planned to develop and refine a testing framework in the form of a graphical user interface (GUI). While this GUI was successfully created as seen in **Figure 3**, we ended up not using it. This was because it would require the use of API keys and given that we had no needed budget for this project, we decided to record the test results in a Google sheet.

Comparison by Stage:

Stage 1 (Project Planning):

- **Planned:** Background project management, scheduling, SoW, and WBS preparation.
- **Actual:** Completed largely on time. Buffer times helped ensure the timely submission of planning documents.

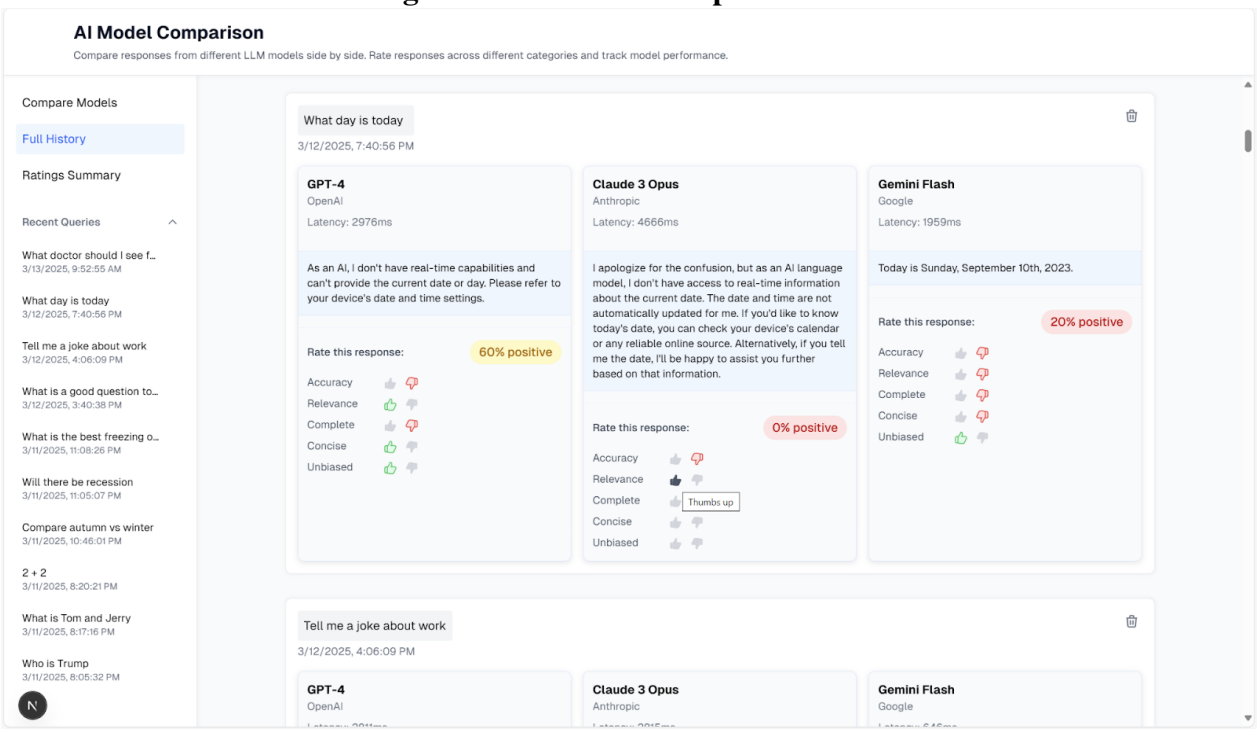
Stage 2 (Testing Setup & Execution):

- **Planned:** Collect test cases early, define benchmark metrics, and run all AIs through a GUI-based framework.
- **Actual:** Delays in test case collection and benchmark complexity forced us to simplify the evaluation metrics and switch from GUI to manual testing. This reduced data uniformity but allowed us to proceed.

Stage 3 (Analysis & Presentation):

- **Planned:** Analyze results and present them in a club seminar.
- **Actual:** Compression in Stage 2 delayed the start of the analysis, but we prioritized core comparisons to prepare effectively for the Gopher AI Club seminar. Despite the reduced dataset, the presentation was delivered on time and well-received.

Figure 3: AI Model Comparison GUI



Screenshot of AI Model Comparison GUI showing prompt and rating interface.

Despite all of the delays and changes in the plan, the overall structure of the project remained unchanged. We gave priority to the main deliverables over the non-essential tasks in the adaptation process.

2.3 *Budget and Duration Variances*

Despite the lack of a budget, our project's duration varied due to resource limitations, particularly the timeliness of question collection and access to commercial APIs. Some of the work, like testing, that was initially planned for two weeks frequently took three or longer. In addition to the risks that were initially identified, we also had to deal with unanticipated ones, like abrupt changes in the performance of the AI model in the middle of the project and model output incompatibilities with particular test formats. Continuity of progress was made possible by our proactive risk reduction, which included using local server resources and simplifying the evaluation approach. Despite deviations from the initial plan, we were able to accomplish our main objectives within the extended period thanks to our flexible modifications.

2.4 *Risk Analysis*

The team came up with multiple risks through discussions in class. First, we discussed potential risks that may arise due to delays in communication with the professors. Then, as we continued, we thought of how we are limited by the usage limits of AIs and how quickly AI progresses. Continuing on this path, with how touchy the government is about foreign competition, even bans on certain

models were considered a risk. There was also a possibility that there would be low attendance at the event and because of this, we might not be able to make the impact that we intended. We took these risks and then put them into a comprehensive risk register while rating each risk's likelihood, impact, and detectability. The risks were then mapped onto the risk severity matrix, as seen in **Figure 4**, to help understand which risks require more immediate action.

Figure 4: Risk Analysis and Contingency Plan

Likelihood						
	5		Because AI Changes quickly , a better model might appear mid-way, potentially making our research/results outdated.			There is a risk of uncontrolled expansion or delay in the testing methods. Due to usage limits.
	4		Limited processing power might slow down testing, making it harder to run large language models efficiently and get complete results. --- ----- Existence of many sub tasks: Many activities dont have predecessors and can be ignored under the line.			
	3		If researchers take too long to provide test cases, the project timeline could be delayed, affecting model evaluation and reporting.			
	2					The AIs get banned in the US
	1	Data quality and relevance of questions may be overly specific to certain fields, potentially limiting their applicability to real-world situations and affecting the performance.				
		1	2	3	4	5
	Impact					

Table 1 describes the risk response plan for each risk that the team came up with. The table outlines what each risk is and how we plan to respond to the risk, whether that be through mitigation or other strategies. The table also describes the potential causes for each risk. We came up with a contingency plan that followed the response strategy and finally described who would be responsible for taking care of the risk should it come to fruition.

In some cases, such as the delay in test case submission, this directly impacted the timeline for finalizing data and preparing for the Gopher AI Club seminar. While we had mitigation strategies in place (regular follow-ups and synthetic questions), the delay was on the critical path and forced us to compress later tasks. Nevertheless, our contingency planning helped ensure that we met the presentation deadline and delivered a complete and functional project.

Table I: Risk Response Plan

No.	Risk Event	Response	Contingency Plan	Trigger	Who is Responsible
1	Researchers delay test case submissions	Mitigate – Set clear deadlines and follow up regularly.	Use synthetic or publicly available test cases.	Missed deadline for test case submission	Team Members/ Testing Leads
2	Limited processing power slows testing	Avoid – Run large models on local or university servers.	Use smaller models or run during off-peak hours.	Hardware or runtime failure	Resource Manager
3	Usage limits cause delays in testing	Mitigate – Simplify evaluation framework and reduce number of tests.	Set test priorities; reduce volume if needed.	API use limits	Testing Lead
4	Questions are too specific or technical	Transfer – Ask professors to simplify/generalize questions.	Ask professors to reframe or simplify questions	Question bank has unclear or domain-specific content	Data Collection Team
5	Certain AIs banned in US	Retain – Use available models; prepare VPN as backup.	Substitute models with similar capabilities.	Model banned for national security risk	Everyone
6	New model released mid-project	Retain – Note limitations; continue with current models.	Explain ongoing relevance or revise scope.	Release of clearly superior AI model	No one (uncontrollable)
7	Too many tasks without dependencies	Mitigate – Streamline task structure.	Remove non-essential tasks from critical path.	Redundancy in task list	Project Managers
8	Low seminar attendance	Retain – Focus on session quality.	Promote event widely; record and distribute.	Low RSVP numbers	Everyone

2.5 Unexpected Risks

During testing, we encountered unexpected issues with the capabilities of various AI models. DeepSeek exhibited extremely long response times and, during peak hours (after 9 PM), experienced significant server congestion, leading to delayed or failed outputs. Claude and Co-Pilot occasionally rejected documents due to size limitations, which constrained our ability to fully test medicine-related prompts. Furthermore, during testing, ChatGPT released a Student Plus Plan, creating discrepancies between free and premium-tier performance, which may have influenced consistency in results.

3 Project Results and Product Deliverables

Our comparison of open-source and proprietary AI models revealed key insights into their performance, cost, and real-world applicability. Subject-specific test questions were collected from students in various fields including Math, History, and Biology. Evaluation metrics included accuracy, relevance, completeness, conciseness, and bias.

Closed-Source Models: ChatGPT, Claude Sonnet, Co-Pilot, Gemini 2.0 Flash
Open-Source Models: Llama 4, DeepSeek R1

Cost analysis, based on 60 words per hour, 1.3 tokens per word, continuous 24-hour usage, and a 3:1 input-output ratio, showed a range from $29.48/month (most expensive) to $0.59/month (least expensive). This is illustrated in **Figure 1**.

Closed-Source AI Strengths: High accuracy, speed, and integration with other tools.
Weaknesses: Costly API access and slower development cycles.

Open-Source AI Strengths: Often free for unlimited use (except Llama).
Weaknesses: Slower responses and server overloads, particularly after 9 PM.

Performance scores were based on subjective ratings (1 point if deserved). Averages were calculated per subject.

Figure 5: Testing Results Based on Performance

Model	Math	History	Biology	Physics	English	Machine Learning	Medicine	Average
ChatGPT	96.00%	85.00%	80.00%	84.00%	80.00%	84.00%	90.00%	85.57%
Claude	100.00%	90.00%	86.67%	96.00%	80.00%	100.00%	-	87.67%
CoPilot	100.00%	80.00%	86.67%	76.00%	86.67%	96.00%	-	86.44%
Gemini	84.00%	90.00%	100.00%	72.00%	100.00%	96.00%	80.00%	88.86%
LLama	84.00%	70.00%	80.00%	92.00%	66.67%	96.00%	95.00%	83.38%
Deepseek	100.00%	95.00%	93.33%	96.00%	100.00%	72.00%	100.00%	86.62%

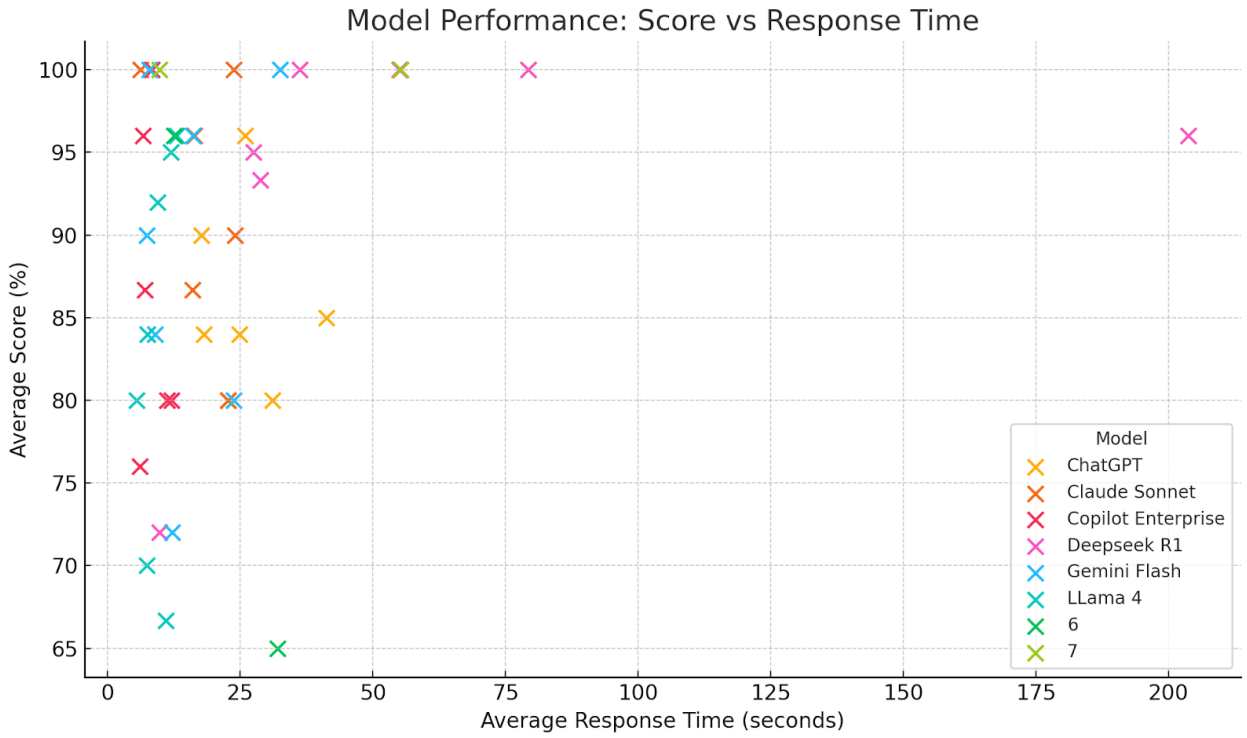
Time-to-response metrics were also collected. DeepSeek had the longest response (8:33.64 in Physics), while Claude had the shortest (3.14 in Machine Learning).

Figure 6: Testing Results Based on Time

Model	Math	History	Biology	Physics	English	Machine Learning	Medicine	Average
ChatGPT	25.97s	41.18s	31.12s	24.86s	22.74s	18.19s	17.62s	25.95s
Claude	23.71s	24.06s	15.95s	16.35s	22.65s	6.21s	-	18.15s
CoPilot	8.38s	12.04s	7.02s	6.02s	11.17s	6.64s	-	8.55s
Gemini	8.90s	7.43s	7.81s	12.20s	32.56s	16.11s	23.73s	15.54s
LLama	7.55s	7.43s	5.51s	9.38s	10.92s	28.09s	11.92s	11.54s
Deepseek	79.28s	27.42s	28.83s	203.75s	36.25s	9.80s	54.98s	64.08s

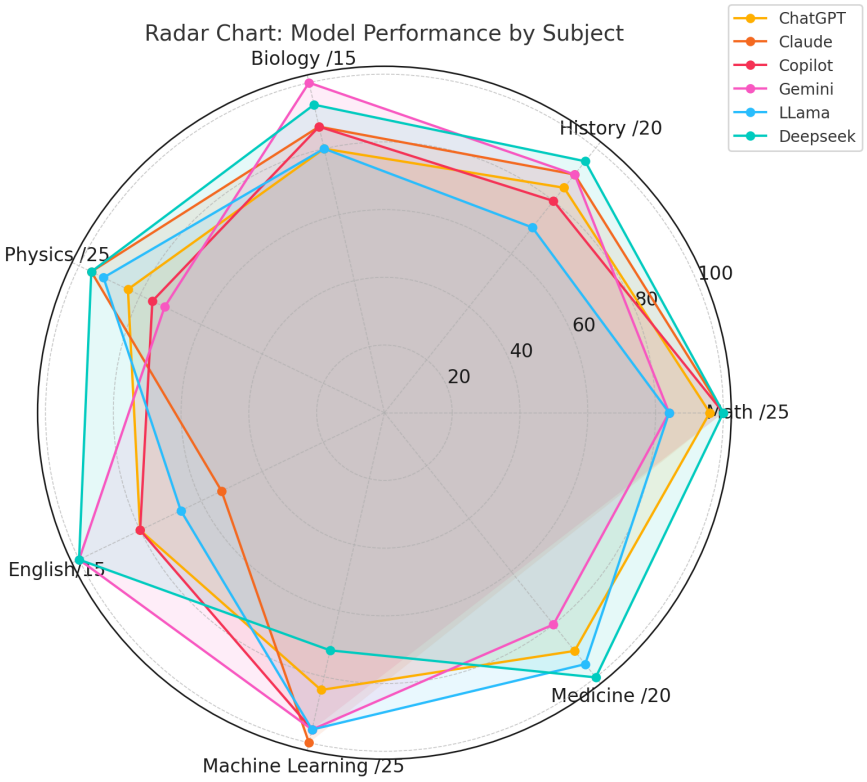
We compared average model scores to their average response times. The scatter plot **(Figure 7)**, maps the average score against the average response time for each model. This plot helps identify trade-offs between performance and efficiency. Ideal models appear in the top-left quadrant, indicating high accuracy and low latency. Models such as Gemini Flash and Claude Sonnet strike a good balance, offering both strong performance and fast responses. Meanwhile, DeepSeek R1, despite its high scores, shows longer response times, suggesting a computational cost for its superior accuracy. Outliers with unusually long latencies or lower scores provide insight into potential limitations for real-time or resource-constrained applications.

Figure 7: Score vs. Response Time Graph



We also evaluated each model's performance by subject. To evaluate and compare the capabilities of different large language models (LLMs) across academic domains, two visualizations were generated. The radar chart **(Figure 8)**, which illustrates the normalized performance of six AI models—ChatGPT, Claude, Copilot, Gemini, LLaMA, and DeepSeek—across seven subjects: Math, History, Biology, Physics, English, Machine Learning, and Medicine. Each axis represents the maximum achievable score in a subject, and the polygons show how each model performs relative to that maximum. Notably, DeepSeek and Gemini exhibit consistently strong performance across all domains, with DeepSeek achieving near-perfect scores in several categories. In contrast, models like Claude and LLaMA display more variability, performing better in technical subjects like Math and Machine Learning but relatively weaker in areas like English and Biology.

Figure 8: Radar Chart of Model Performance by Subject



3.1 Seminar Event Details

A major deliverable was a seminar hosted at the Gopher AI Club on April 22, 2025, in Keller Hall Room 3-180, University of Minnesota. Approximately 25 attendees from computer science, data science, and biomedical engineering participated.

Presenters included Shizra Tariq, Kevin Dai, Niketha Sabesan, and Vivan Nguyen. Topics included AI model overviews, evaluation criteria, performance findings, and cost implications.

The seminar concluded with an engaging Q&A session. Questions focused on model limitations, pricing models, and practical application. Feedback indicated that the seminar was informative and useful for students exploring AI tools.

Advertising was conducted via email through the Gopher AI Club.

Figure 9: Gopher AI Club Seminar Presentation – April 22, 2025

4 Lessons Learned

Our team learned several important project management lessons. First, we recognized the value of stricter internal deadlines and milestone ownership. Although some delays were out of our control (e.g., professors submitting test questions), proactive contingency planning—such as creating synthetic backups—would have helped. Second, relying on informal updates led to missed or delayed tasks. Frequent, structured check-ins would have improved pacing and accountability.

Additionally, we learned the importance of flexible scheduling and early risk identification in rapidly evolving technical projects. Trying to include too many models or test metrics created overhead, which could have been avoided by narrowing scope. Finally, early investments in automation greatly reduced testing time and should be standard in future data-intensive projects.

5 Conclusions

This project showed that while closed-source AI models like ChatGPT and Claude offer superior performance in speed and accuracy, open-source models like DeepSeek and Llama are more cost-effective and flexible. Closed models excel in reliability and tool integration, ideal for professional applications. Open models, however, offer affordability and adaptability—though at the cost of response speed and reliability.

As AI evolves rapidly, strategies must remain adaptable. Our comparison offers students, researchers, and professionals the insights needed to choose between performance and scalability, depending on organizational needs. Both open- and closed-source models have critical roles in shaping the future of AI.

6 Appendix

Figure 10: Full WBS

WBS ID	Status	Task Name	Duration	Start	Finish	Predecessors	Assigned To	Work
		Status Key						
		Not Started						
		Behind						
		In Progress						
		Complete						
		Testing the AI's	87d	02/04/25	05/01/25			
1		**Planning and Scheduling**	73d	02/04/25	04/17/25			
1.1		Start of Project	0	02/04/25	02/04/25		Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	
1.2		SoW	8d	02/04/25	02/11/25	10	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	16
1.3		SoW Revision	9d	02/12/25	02/20/25	11	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	18
1.4		WBS	7d	02/21/25	02/27/25			
1.4.1		Draft WBS	7d	02/21/25	02/27/25	12	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	28
1.4.2		Submit WBS	1d	02/27/25	02/27/25	14FF	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	0
1.5		Risk Management Plan	1d	02/27/25	02/27/25	13FF	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	4
1.6		Project Plan	19d	02/28/25	03/18/25			
1.6.1		Draft Project Plan	19d	02/28/25	03/18/25	12, 14, 16	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	38
		Submit Project Plan	1d	03/18/25	03/18/25	18FF	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	0
1.7		Project plan revision	14d	03/19/25	04/01/25	17	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	28
1.8		Sponsor Outreach	11d	02/12/25	02/22/25			
1.8.1		Develop Sponsorship Proposal	7d	02/12/25	02/18/25	11	Kevin Dai	3
1.8.2		Confirm Sponsors	7d	02/16/25	02/22/25	22SS +4d	Kevin Dai	1
1.8.3		Confirmation of Seminar Date	1d	02/22/25	02/22/25	23FF	Kevin Dai	0
1.9		Resource Allocation Strategy	2d	02/23/25	02/24/25	23	Kevin Dai	2
1.11		Select AI Models	1d	02/24/25	02/24/25	25FF	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	4
1.12		Decide Testing Categories	1d	02/25/25	02/25/25	26	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	4
1.13		Mid-Project Review	3d	04/02/25	04/04/25	45SS +7d, 58SS, 20	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	4
1.14		Status Update Memo	4d	04/05/25	04/08/25			
1.14.1		Draft Status Update Memo	4d	04/05/25	04/08/25	28	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	16
1.14.2		Submit Status Update Memo	1d	04/08/25	04/08/25	30FF	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	0
1.14.3		Send Status Update to Sponsors	1d	04/08/25	04/08/25	30FF	Kevin Dai	0
1.15		Status Update Memo Revision	9d	04/09/25	04/17/25	30	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	16
2		**Running Tests and Compiling Results**	15d	02/24/25	03/10/25			
2.1		Research Similar Studies	7d	02/24/25	03/02/25	10FS +20d	Vivian Nguyen	7
2.2		Summarize Research Findings	4d	02/28/25	03/03/25	35SS +4d	Vivian Nguyen	4
2.3		Develop Testing Framework	4d	03/04/25	03/07/25	36	Vivian Nguyen	4
2.4		**Professor Outreach**	13d	02/26/25	03/10/25			
2.4.1		Contact Professors	13d	02/26/25	03/10/25	27	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	26
2.4.2		Collect Questions from Professors	5d	03/06/25	03/10/25	39SS +8d	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	10
2.4.3		Classify Test Cases	4d	03/07/25	03/10/25	40SS +1d	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	16
2.4.4		Finalize Question Bank	1d	03/10/25	03/10/25	41FF	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	16
2.5		Define Benchmarks for Comparison	7d	03/04/25	03/10/25	36	Vivian Nguyen	7
2.6		**AI Model Testing**	14d	03/11/25	03/24/25			
2.6.1		Plan and assign AI Model testing	1d	03/11/25	03/11/25	26, 37, 43, 42	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	16
2.6.2		Initial Performance Evaluation	5d	03/12/25	03/16/25	45	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	8
2.6.3		Hyperparameter Tuning	7d	03/12/25	03/18/25	45	Kevin Dai	
2.6.4		Stress Testing	4d	03/12/25	03/15/25	45	Kevin Dai	
2.6.5		Cross-Validation	10d	03/12/25	03/21/25	45	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	8
2.6.6		Edge Case Testing	3d	03/12/25	03/14/25	45	Niketha Sabesan	3
2.6.7		Scalability Testing	8d	03/12/25	03/19/25	45	Shizra Tariq	2
2.6.8		Bias and Fairness Testing	5d	03/12/25	03/16/25	45	Niketha Sabesan	2
2.6.9		Robustness to Adversarial Attacks	6d	03/12/25	03/17/25	45	Vivian Nguyen	2
2.6.10		Real-world Test Analysis	10d	03/12/25	03/21/25	45	Vivian Nguyen	2
2.6.11		Long-term Performance Monitoring	13d	03/12/25	03/24/25	45	Shizra Tariq	3
2.6.12		Integration Testing	10d	03/12/25	03/21/25	45	Shizra Tariq	2
3		**Seminar Execution and Analysis**	38d	03/25/25	05/01/25			
3.1		Finalize Testing Results	4d	03/25/25	03/28/25	46, 47, 48, 49, 50,	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	8
3.2		Benchmark Comparison Dataset	8d	03/29/25	04/05/25			
3.2.1		Draft Benchmark Comparison	8d	03/29/25	04/05/25	58	Kevin Dai	
3.2.2		Finalize Benchmark Comparison	1d	04/05/25	04/05/25	60FF	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	8
3.3		Performance Analysis Report	18d	03/29/25	04/15/25			
3.3.1		Draft Performance Analysis Report	18d	03/29/25	04/15/25	58	Niketha Sabesan	5
3.3.2		Finalize Performance Analysis	1d	04/15/25	04/15/25	63FF	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	8
3.4		Data Analysis and Evaluation	6d	03/29/25	04/03/25	58	Niketha Sabesan	3
3.5		Create Data Visuals	5d	04/04/25	04/08/25	65	Niketha Sabesan	6
3.6		Final Report Draft	9d	04/16/25	04/24/25	58, 65, 66, 61, 64	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	8
3.7		Finalize Presentation Material for	7d	04/18/25	04/24/25			
3.7.1		Draft Presentation Material	7d	04/18/25	04/24/25	33, 61	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	4
3.7.2		Finalize Presentation Material	1d	04/24/25	04/24/25	69FF	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	4
3.8		**Present the results of the project to AI**	1d	04/25/25	04/25/25			
3.8.1		Present Results for Seminar	1d	04/25/25	04/25/25	70	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	4
3.8.2		Collect feedback from audience	1d	04/25/25	04/25/25	72SS	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	4
3.9		**Finalize Project**	9d	04/23/25	05/01/25			
3.9.1		Feedback Integration	2d	04/26/25	04/27/25	73	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	8
3.9.2		Final project report	2d	04/26/25	04/27/25	75SS	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	40
3.9.3		Project poster designing	8d	04/23/25	04/30/25	67SS +7d	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	8
3.9.4		Final project poster presentation	1d	05/01/25	05/01/25	77	Kevin Dai, Niketha Sabesan, Shizra Tariq, Vivian Nguyen	4

References

[1] Aran Komatsuzaki [@arankomatsuzaki]. (2025, January 25). Here is our cost estimate for training popular models like GPT-4o, Sonnet, and DeepSeek (w/ H100s)! https://x.com/arankomatsuzaki/status/1884676245922934788?s=46

[2] DeepSeek-AI, Lu, C. D., Wu, B. C., Wang, B. X., Xue, B., Feng, B., & Liu, A. X. (2024, December 27). DeepSeek-V3 Technical Report. *arXiv*. https://arxiv.org/html/2412.19437v1

[3] Holland, S. (2025, January 21). Trump announces private-sector $500 billion investment in AI Infrastructure. *Reuters*. https://www.reuters.com/technology/artificial-intelligence/trump-announce-private-sector-ai-infrastructure-investment-cbs-reports-2025-01-21/

[4] Howarth, J. (2025, February 26). ChatGPT Enterprise Pricing, Features, and Limitations. *Exploding Topics*. https://explodingtopics.com/blog/chatgpt-enterprise

[5] Kaduri, B., & Stansfield, T. (2024, October 8). 10 Most Popular AI Models of 2024. *Orca Security*. https://orca.security/resources/blog/top-10-most-popular-ai-models-2024/