THE GRADING CAPABILITIES OF LARGE LANGUAGE MODELS: A COMPARATIVE STUDY OF OPENAI AND OLLAMA ACROSS PYTHON AND SHORT ANSWERS ASSESSMENTS AND RUBRICS

Christopher Vishnu Kumar
UniSQ Toowoomba

Dr Derek Long
UniSQ Toowoomba

ABSTRACT

The integration of Language Learning Models (LLMs) into educational assessments has shown substantial potential to transform grading processes. However, their application is often hindered by inconsistencies in performance across various tasks and the challenge of aligning automated scoring with manual scoring standards. This study introduces a comprehensive framework designed to evaluate the performance of LLMs using a diverse array of rubrics and metrics such as R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE). Our framework was rigorously applied to both commercial and open-source models, including GPT-4 Turbo-Preview, GPT-4, Llama 3, and Mistral-7B Instruct v0.2, across a series of Python programming and short answer tasks.

Our results underscore the superior performance of GPT-4 Turbo-Preview, particularly when coupled with detailed numeric weighted rubrics, demonstrating its robustness in complex grading scenarios. Similarly, the GPT-4 model consistently showed high accuracy, suggesting its effectiveness in educational applications. In contrast, open-source models like Llama 3 exhibited variable performance, though they performed commendably in specific settings. The study also highlights that while numeric rubrics can yield high accuracy, text-rich and detailed rubrics generally enhance model performance by leveraging LLMs' advanced language processing capabilities.

Furthermore, our findings suggest potential for innovative hybrid rubrics that combine numeric precision with descriptive textual analysis, paving the way for enhanced accuracy and reliability in automated assessments. This research has significant implications for educational technology, providing a robust framework that can guide the development and implementation of LLMs in educational settings, and is of particular interest to educators and curriculum developers looking to integrate AI-driven tools into the learning environment. To support ongoing research and application, all methodologies and resources have been made publicly available on GitHub at https://github.com/christopherkumar/prompt_engineering_test.git.

1. INTRODUCTION

Integrating artificial intelligence (AI) in education can revolutionise how students are evaluated and supported throughout their learning journeys. One of AI's most promising educational applications is using large language models (LLMs) to support educational assessments. LLMs can potentially improve the efficiency and effectiveness of grading, provide personalised feedback, and support teaching and learning.

However, using LLMs in educational settings raises important questions about the role of human judgment in educational assessments, the potential biases of LLMs, and the need for more research on their effectiveness in different educational contexts. This study aims to contribute to understanding the potential benefits and limitations of using LLMs in educational assessments. It focuses on their use in grading Python programming assignments and question types requiring short-answer responses.

The study is driven by the need to ensure that LLM deployment in educational settings is effective and reliable. By evaluating the effectiveness of LLMs in grading educational assessments, this study aims to provide a clearer understanding of their potential to support teaching and learning and to identify areas where further research is needed.

The study is grounded in assessment validity, which emphasises the importance of ensuring that assessments are fair and valid measures of student learning. The study also draws on the importance of rubrics in educational assessments, which provide a clear and transparent framework for evaluating student work.

The methodology involves using a combination of LLMs, rubrics, and assessment pieces to evaluate the effectiveness of LLMs in grading educational assessments. The study does not aim to adjust, modify, tune, or train models but to evaluate their effectiveness in grading educational assessments using a variety of prompts and rubrics. It neither attempts to influence the models' behaviour, having all prompts passed as a zero-shot prompt.

This study can directly inform the development of more effective and efficient assessment methods by providing a deeper understanding of the potential benefits and limitations of using LLMs in educational assessments.

2. Related Work

The landscape of higher education is undergoing a significant transformation, propelled by the latest advancements in artificial intelligence (AI), especially within the domain of generative AI. This evolution is profoundly influencing educational methodologies, particularly through the integration of automated assessment and feedback mechanisms. This section delves into the transformative potential of AI and Large Language Models (LLMs) in refining traditional assessment practices and enhancing student outcomes. We explore their application in grading Python scripts and essays, tackle the technological challenges involved, and envisage future trajectories for the incorporation of AI and LLMs within educational frameworks.

Automated Assessment and Feedback

The role of AI in higher education is becoming increasingly pivotal, primarily due to its capability to deliver immediate and precise feedback, a critical component for fostering student success. The importance of timely feedback is underscored in studies such as those by Bulut (2022) and Hooda (2022), which highlight its significant impact on learning outcomes by addressing students' needs in real time. Furthermore, Tubino (2022) accentuates the development of feedback literacy, which is emerging as an essential skill enabling students to engage with and derive maximum benefit from feedback actively. Additionally, Chang (2023) presents a practical illustration of how AI can be utilised within an automated post-rating system that promotes interdisciplinary knowledge exchange, thereby enriching the educational experience. This shift towards AI-enhanced learning tools reflects a broader trend of integrating cutting-edge technology to cultivate a dynamic and responsive learning environment.

Large Language Models in Education - Grading and Beyond

The integration of Large Language Models (LLMs) such as GPT-3 and Codex has revolutionised the educational grading process, particularly in the assessment of Python scripts and essays. Research undertaken by Balse (2023) and Phung (2023) showcases how LLMs enhance the accuracy and relevance of feedback provided on Python scripts, facilitating a more nuanced understanding of student submissions. Furthermore, studies by Wadhwa (2023) and Ellis (2024) underline the utility of LLMs in improving code quality and offering comparative analyses with solutions generated by humans.

In the realm of essay grading, techniques such as Latent Semantic Analysis (LSA) are employed to assess the written content, as elucidated by Barkaoui (2010) and Valsamidis (2012). LSA is instrumental in evaluating linguistic features and user engagement metrics within Learning Management Systems (LMS). This approach highlights the adaptability of LLMs to comprehend and assess complex student submissions, showcasing their potential to extend beyond simple grading tasks. The integration of these models into educational practices not only enhances the grading process but also contributes to a more holistic and informed educational environment.

AI Innovations in Comprehensive Student Assessment

AI is fundamentally reshaping comprehensive student assessment strategies, extending its influence beyond mere feedback mechanisms. Studies by Rodríguez-Hernández (2021) and Xia (2022) delve into the capabilities of AI to predict and enhance academic performance through the

analysis of data-driven insights. These insights enable educators to tailor instructional strategies effectively, ensuring they meet the unique needs of each student. Such personalised approaches are pivotal in maximising educational outcomes and fostering an environment conducive to individualised learning.

Moreover, the challenge of assessment biases remains a significant concern. Historical discussions by Dennis (1996) and more recent analyses by Gómez-Benito (2010) highlight the importance of addressing these biases. They advocate for the development of bias-free assessment tools that ensure equitable and valid evaluations across diverse student populations. This focus on equity is crucial in realising the full potential of AI in education, as it strives to offer fair and accurate assessments that reflect the true capabilities of all students, irrespective of their backgrounds.

Use Cases and Applications

Large Language Models (LLMs) have proven to be highly versatile and effective in various assessment contexts, underlining their broad applicability across different types of educational evaluations. Chatchai Wangwiwattana and Yuwaree Tongvivat (2023) have developed a method utilising LLMs to assess students' short-answer responses automatically. This method incorporates answer matching, keyword extraction, and clustering techniques, achieving an impressive accuracy rate of 99.03%. Their innovative approach highlights the precision and reliability of LLMs in evaluating student submissions.

In a related study, Wangwiwattana and Tongvivat (2022) introduced a grading approach for short open-ended questions using document clustering techniques coupled with TF-IDF and K-Means algorithms. This method exemplifies how LLMs can be tailored to handle more open-ended and subjective content, providing educators with robust tools to assess student understanding effectively.

Additionally, Olowolayemo (2018) explores the use of LLMs in assessing fill-in-the-blank questions by analysing student responses based on text similarity measurements. This focus on measuring text similarity demonstrates the potential of LLMs to adapt to various question formats, offering a nuanced assessment that can accurately gauge student knowledge and comprehension. Together, these examples showcase the diverse applications of LLMs in educational assessments, ranging from structured to open-ended response formats.

Technological Challenges and Future Directions

While AI and Large Language Models (LLMs) present considerable benefits, they are not without challenges. Research conducted by Li (2023) and Yuan (2023) sheds light on the technical limitations, particularly in quantisation. These studies aim to optimise LLM efficiency and deployment, addressing the need for these models to operate more effectively within resource-constrained environments. This highlights a key area where technological advancements are crucial for the broader application of AI in educational settings.

Moreover, the ethical implications of integrating AI into education must be considered. As discussed by Yan (2023) and Cavojský (2023), ethical considerations must be at the forefront of deploying AI technologies in education. It is not just important but essential to ensure that these technologies align with educational values and uphold equitable practices. The ethical deployment of AI involves careful consideration of how these technologies impact all students,

particularly in terms of accessibility and fairness, providing a reassurance of responsible use.

The integration of AI and LLMs into higher education represents a transformative opportunity that has the potential to revolutionise traditional teaching and assessment methods profoundly. These technologies not only enhance feedback mechanisms and improve assessment accuracy but also provide personalised learning experiences tailored to individual student needs. As the capabilities of AI and LLMs continue to advance, their impact on higher education is anticipated to expand significantly, reshaping the landscape of teaching, and learning in profound ways.

To fully realise the potential of AI in education, ongoing research, ethical oversight, and collaboration between educators and AI specialists will be crucial. This collective effort, which requires the active involvement of educators and researchers, is essential to harness the full capabilities of AI technologies. It ensures that these technologies meet the complex demands of contemporary educational environments and contribute positively to the educational experience, making everyone feel integral to the process.

3. Experimental Validation

Overview

This study aimed to rigorously assess the effectiveness and reliability of various Language Models (LLMs) in automating the grading of educational assessments, specifically Python scripts and short answer tests for undergraduate-level content. The goal was to validate the performance of these models across multiple dimensions and against varied assessment rubrics using a diverse range of LLMs and an extensive evaluation framework.

Background

The integration of LLMs into educational settings, particularly for automated grading, has gained traction due to their potential to provide rapid and consistent evaluations. However, it is imperative to thoroughly evaluate the performance and reliability of these models before implementing them in practical applications.

Setup

Our experimental setup involved using locally hosted models and API-based services to assess their efficacy in automated grading. We selected LLMs from two prominent providers, OpenAI and Ollama, with an additional model downloaded from Hugging Face.

We designed six predefined rubrics in detail to ensure a comprehensive evaluation. The assessments were categorised into two main types: six Python script tasks of increasing complexity and four short answer questions covering various theoretical topics, which were created as the task to be outlined in the rubric.

Models

We selected a diverse set of LLMs from different providers to assess their performance in automated grading:

OpenAI LLMs

We leveraged OpenAI's powerful models, GPT-3.5-turbo, GPT-4-turbo-preview, and GPT-4, by accessing them via API calls. Unlike the Ollama models described in the subsequent section, these models necessitate a stable internet connection and adequate API call or token allowances to prevent crashes during API interactions.

Ollama LLMs

We installed Ollama and utilised their range of models, as detailed in the table below. Ollama's local implementation allowed us direct control over model execution. Additionally, we downloaded the Mistral-7B-Instruct-v0.2 model from Hugging Face and deployed it with Ollama.

Model	Architecture	Parameters	Quantization
Gemma	Gemma	9B	4-bit
Llama2	Llama	7B	4-bit
Llama3	Llama	8B	4-bit
Mistral	Llama	7B	4-bit
WizardLM2	Llama	7B	4-bit
TheBloke/Mistral-7B-Instruct-v0.2-GGUF	Llama	7B	8-bit

Table 1. Open-source models used (deployed using Ollama).

The primary parameters which were kept constant to ensure consistent results were:

(These are the parameters for the OpenAI API. For Ollama-Python, these are sub-parameters under “options”).

- temperature” set to “0.2.”
- “seed” set to “1.”

For Ollama models, the "num_predict" parameter was set to "2048," which pre-empts the model with the number of tokens to generate for the response. This was done to address an issue where the Ollama server continuously generated output tokens greater than the size of the context window.

All the models implemented using Ollama are running completely offline and locally.

Test Data Generation

Synthetic data, consisting of Python scripts and short answers, was generated to mimic the varying levels of quality that undergraduate students could expect. This data was used to test how well LLMs could mark these scripts and answers accurately according to a defined rubric.

Data Generation Process

The data was generated using a prompt designed to produce outputs representing different quality levels, ranging from 1 (poorest) to 5 (best). The prompt was formulated as follows:

> "I need to generate {short answers/Python scripts} which attempt to fulfil the requirements outlined in the rubric. The {short answers | Python scripts} need to be of quality 1 (worst) - 5 (best). The {short answers | Python scripts} need to be written as though an undergraduate student had written them. Write 10 {short answers | Python scripts} of quality {1-5}."

This structured approach allowed us to systematically generate diverse responses that could be used for model evaluation.

Tools and Technologies

The GPT-4 model was used for the generation of Python scripts. We combined multiple advanced language models to generate more complex textual data, such as short answers, to ensure variety and depth. These models included:

- ChatGPT (GPT-4)
- Gemini
- Claude 3 Sonnet
- Perplexity

As well as models available through HuggingChat:

- CohereForAI/c4ai-command-r-plus
- meta-llama/Meta-Llama-3-70B-Instruct
- HuggingFaceH4/zephyr-orpo-141b-A35b-v
- mistralai/Mixtral-8x7B-Instruct-v0.1
- NousResearch/Nous-Hermes-2-Mixtral-8x
- mistralai/Mistral-7B-Instruct-v0.2

The synthetic data generation process we employed is crucial as it enables us to create a controlled yet diverse set of data. This data is instrumental in effectively training and testing language models. By encompassing a variety of responses and ensuring that these responses span a range of defined quality levels, we can better simulate the grading of real student submissions. The diversity of the language models used also ensures that the generated data closely mirrors the variability expected in actual student responses.

RUBRICS

The grading rubrics were carefully designed to evaluate the LLM outputs comprehensively. We structured our rubrics into two main types. These rubrics assessed the depth and relevance of the responses to the posed questions.

TEXT-BASED RUBRICS

- Detailed Rubrics: Provided specific feedback and included comprehensive evaluation criteria.
- Nondetailed Rubrics: Offered a more general assessment.

Scores were assigned using categories such as Poor, OK, Competent, Excellent, and Perfect.

NUMERIC RUBRICS

Numeric rubrics assigned scores based on the presence of key terms and concepts. We developed four subtypes of numeric rubrics:

- Detailed Numeric Rubrics: Provided extensive information justifying each score.
- Nondetailed Numeric Rubrics: Offered concise explanations.
- Weighted Rubrics: Considered the relative importance of each criterion.
- Nonweighted Rubrics: Assigned scores on a simple scale (1-5) before applying weights based on specific criteria.

SYSTEM PROMPT

This methodology aims to pass prompts to the LLMs as a single, initial prompt without keeping the context in memory. No candidate prompts were passed to pre-empt the model into giving more favourable responses. A combination of one of the six rubric formats, with the Python script or short answer appended to it was created before being passed as a prompt to the LLMs.

The models themselves were given a "system" prompt which would tell the model how to behave. This was the primary testing parameter, as we wanted to test the efficacy of prompting only.

> You are an assistant tasked with evaluating Python scripts using a specified rubric. Upon receiving a Python script and the rubric, your role is to:
>
> Parse the provided rubric, which specifies scoring criteria such as Functionality, Logic, Code Quality, User Input Handling, and Documentation. Each criterion has its own potential score.
>
> Assess the Python script against each criterion listed in the rubric.
>
> Calculate the score for each category based on the assessment and the criteria in the rubric.
>
> Remember to:
>
> Avoid explaining why a particular score was assigned.
>
> Do not include any comments, definitions, explanations, or calculations beyond the scores.
>
> If the rubric contains a numeric scoring system, assign a numeric score.

If the rubric contains a text scoring system, assign a text score.

Provide the results in a list format as follows:

<start of response> Functionality: {score}, Logic: {score}, Code Quality: {score}, User Input Handling: {score}, Documentation: {score} </end of response>

Figure 1. System prompt given to the LLMs whilst marking scripts.

You are an assistant tasked with evaluating Python scripts using a specified rubric. Upon receiving a Python script and the rubric, your role is to:

Parse the provided rubric, which specifies scoring criteria such as Understanding of the Topic, Argumentation and Evidence, Organisation and Clarity. Each criterion has its own potential score.

Assess the Python script against each criterion listed in the rubric.

Calculate the score for each category based on the assessment and the criteria in the rubric.

Remember to:

Avoid explaining why a particular score was assigned.

Do not include any comments, definitions, explanations, or calculations beyond the scores.

If the rubric contains a numeric scoring system, assign a numeric score.

If the rubric contains a text scoring system, assign a text score.

Provide the results in a list format as follows:

<start of response> Understanding of the Topic: {score}, Argumentation and Evidence: {score}, Organisation and Clarity: {score}</end of response>

Figure 2. System prompt given to the LLMs whilst marking short answers.

DATA COLLECTION

We collected and compared the scores from both the automated and manual grading processes to evaluate each LLM's performance and consistency. We employed key statistical metrics, including:

- R-squared,
- Mean Squared Error (MSE),
- Mean Absolute Error (MAE), and
- Scatter plots,

to analyse the results and quantify the accuracy and errors in the automated scoring system.

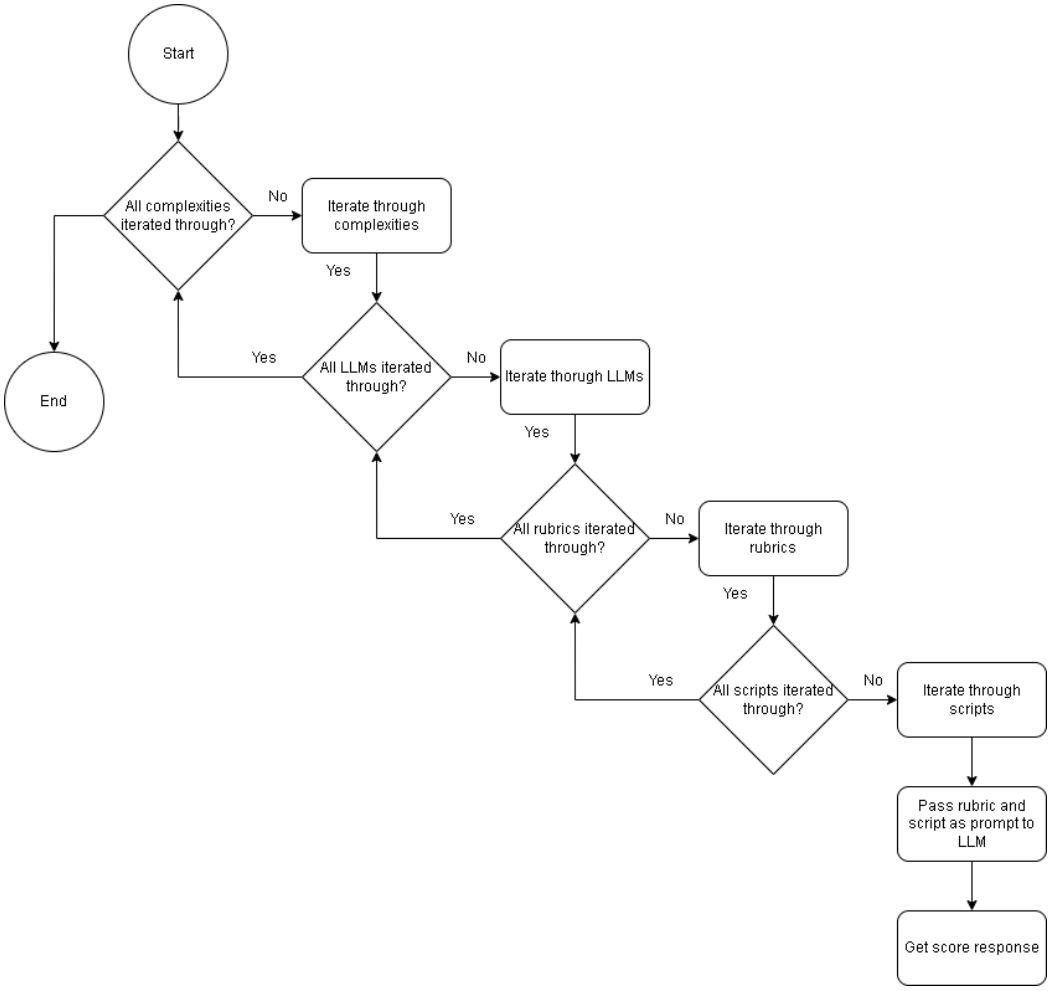

Figure 3. Overview of response collection methodology used

ANALYSIS

We analysed the collected data statistically to identify patterns and outliers in LLM performance. This analysis guided the refinement of scoring algorithms and the adjustment of rubrics, leading to improved accuracy and reliability. Our goal was to enhance the overall grading process by understanding model strengths and weaknesses.

Tools and Resources

Hardware Configuration

The experiments were conducted on a high-performance desktop computer with the following specifications:

- Processor: 11th-generation Intel Core i7-11700
- Clock Speed: 2.50GHz
- Graphics Processing Unit (GPU): NVIDIA GeForce RTX 3060 with 12GB of video RAM (VRAM)
- Random Access Memory (RAM): 16GB

Software Environment

The experiments utilised custom Python scripts developed using Visual Studio Code, a popular integrated development environment (IDE). The scripts were written in Python version 3.10.0, and all necessary packages and dependencies are listed in the requirements.txt file, available in the GitHub repository at [https://github.com/christopherkumar/prompt_engineering_test.git].

Large Language Models (LLMs)

Local LLMs (Offline)

- Gemma
- Llama2
- Llama3
- Mistral
- WizardLM2
- TheBloke/Mistral-7B-Instruct-v0.2-GGUF

OpenAI LLMs (API)

- GPT-3.5-Turbo
- GPT-4
- GPT-4-Turbo-Preview

APIs and Libraries

The experiments leveraged the following APIs and libraries:

- OpenAI platform: The official OpenAI documentation [https://platform.openai.com/docs/introduction] and the OpenAI-Python repository [https://github.com/openai/openai-python] were referenced for implementation details specific to OpenAI's API and Python library.
- Ollama LLM: The official documentation [https://github.com/ollama/ollama] and the Ollama-Python repository [https://github.com/ollama/ollama-python] were consulted for guidance on implementation and usage.

RUBRIC FORMATS

*See Appendix A.1 for the full rubric format templates.

- numeric_detailed_nonweighted
- numeric_detailed_weighted
- numeric_nondetailed_weighted
- numeric_nondetailed_nonweighted
- text_detailed
- text_nondetailed

SCRIPT RUBRICS

*See Appendix A.2 for the full script questions

The six rubric formats were used to create rubrics for each of the following script questions.

Script Questions

- Sphere surface area and volume calculations.
- 5 number-summary of a given array.
- Create a basic calculator.
- Zip files from directory A to B.
- Print out the word frequency in a text passage.
- Create a basic contact book.

SHORT ANSWER RUBRICS

*See Appendix A.3 for the full short answer questions

The six rubric formats were used to create rubrics for each of the following short answer questions.

- Using drones vs satellites for image capture.
- Using small, specialised robots instead of a large boom for pesticide/weedicide application.
- Using commercial satellites vs non-commercial satellites for image capture.
- Why did the first generation of sensors for data capture have hoods?

4. Assessment of Open-Source Models' Competitiveness

While GPT-4 Turbo-Preview and GPT-4 from OpenAI consistently rank among the best performers, it is crucial to evaluate the competitiveness of open-source models like Gemma, Llama2, Llama3, Mistral, and WizardLM2, as well as the Hugging Face-hosted Mistral-7B-Instruct-v0.2.

Current Competitiveness

Currently, the performance of open-source models lags behind that of their OpenAI counterparts in complex assessment scenarios. For instance, models like Llama2 and Gemma have shown substantial underperformance, particularly in environments requiring nuanced language processing and sophisticated evaluation criteria, as demonstrated by their negative R-squared values in many rubric tests. These results suggest difficulties in effectively integrating detailed textual analysis, which is critical for handling complex grading scenarios.

Potential for Competitiveness

Despite their current shortcomings, there is significant potential for these open-source models to become more competitive. This potential hinges on several factors:

Enhancements in Model Training and Architecture: Increasing the parameter count and improving quantization methods could substantially boost the performance of these models. For example, the 4-bit quantization used in most Llama models might need to be improved to handle the fine-grained nuances required by detailed rubrics. Transitioning to 8-bit or higher could enhance their processing capabilities.

Customized Rubric Design: Tailoring rubrics better to match the specific strengths and weaknesses of these models can optimize their performance. For instance, simplifying text-based rubrics or enhancing numeric rubrics with more contextual cues could help models like Llama2 and WizardLM2 perform better.

Community and Collaborative Improvements: Open-source models benefit from community-driven development, which can accelerate improvements and innovation. By leveraging a broader base of contributors, these models can quickly incorporate advanced AI research and techniques that could close the gap with commercial models.

Focused Research on Specific Applications: Directing research efforts towards specific educational applications, like automated feedback systems or adaptive learning environments, can help these models excel in niche areas and become more competitive overall.

5. Model and Rubric Combination Review

In the upcoming section of the report, we systematically explore the effectiveness of various language models and rubric combinations across all complexities for both Python scripts and short answers. This detailed analysis is structured to provide a thorough breakdown for each level of complexity, ensuring a nuanced understanding of performance dynamics in automated grading systems.

For each complexity level, the analysis will include:

- Scatter Plots: Visual representations through scatter plots will showcase the distribution and correlation of scores achieved using different model and rubric combinations, highlighting the top performers.
- Top 5 Best Performing Combinations: Identification of the top five model and rubric combinations that have demonstrated the highest efficiency and accuracy in grading based on metrics such as R-squared values.
- Bottom 5 Worst Performing Models: Conversely, the analysis will also spotlight the five least effective combinations, underscoring potential areas for improvement or adaptation in model or rubric design.
- Best Overall Model: This segment will pinpoint the single model that consistently performs well across various rubrics, offering insights into its robustness and versatility.
- Best Overall Rubric: Similarly, the rubric that yields the most reliable and accurate results across different models will be identified, suggesting its effectiveness in capturing essential assessment criteria.
- Notable Trends: Finally, the section will discuss discernible trends and patterns observed in the data, such as the impact of rubric detail or model complexity on grading accuracy, providing valuable insights for future applications and research in the field of educational technology.

*See Appendix A.4 for all R-squared scores for model and rubric combinations.

Analysis for scripts-complexity1 (Python Scripts, Complexity Level 1, sphere)

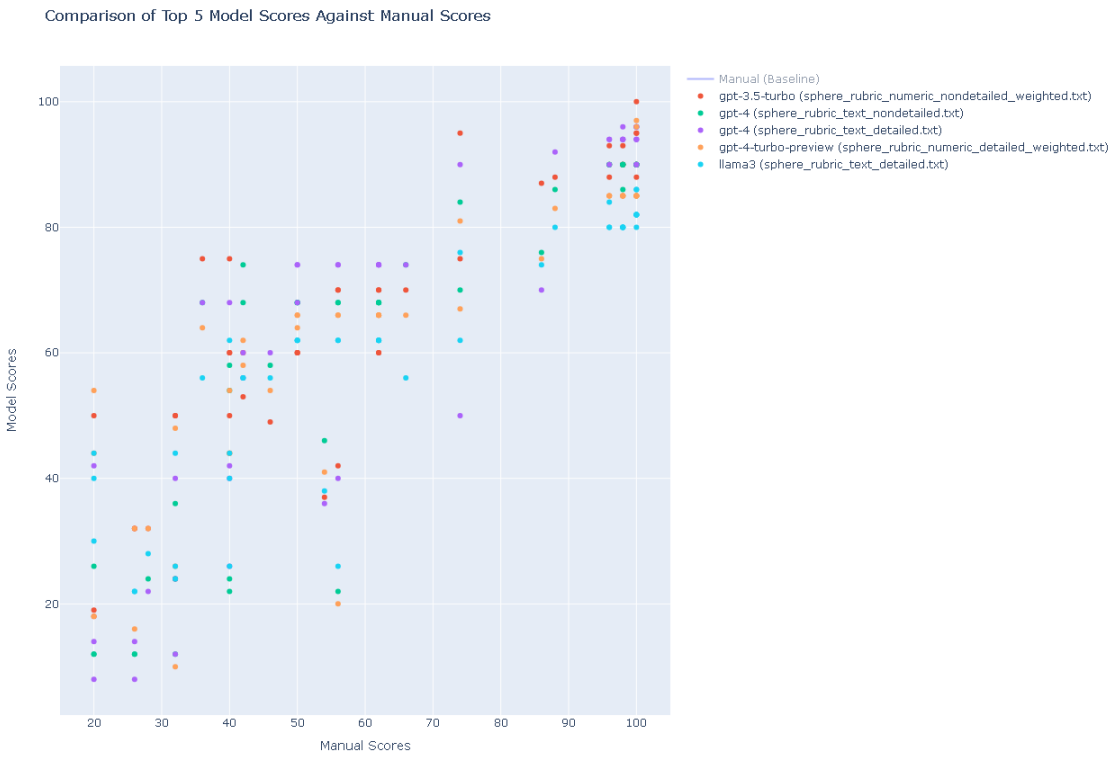

Figure 4. Achieved scores using different combinations of LLM and rubric format for scripts-complexity1-sphere.

Top 5 Best Performing Model + Rubric Combinations

1. GPT-3.5-Turbo with rubric sphere_rubric_numeric_nondetailed_weighted.txt (R^2 = 0.7839)
2. GPT-4 with rubric sphere_rubric_text_nondetailed.txt (R^2 = 0.7691)
3. GPT-4 with rubric sphere_rubric_text_detailed.txt (R^2 = 0.7666)
4. GPT-4-Turbo-Preview with rubric sphere_rubric_numeric_detailed_weighted.txt (R^2 = 0.7529)
5. LLama3 with rubric sphere_rubric_text_detailed.txt (R^2 = 0.7488)

Bottom 5 Worst Performing Model + Rubric Combinations

1. Llama2 with rubric sphere_rubric_text_detailed.txt (R^2 = -2.5066)
2. Llama2 with rubric sphere_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -0.7136)
3. Gemma with rubric sphere_rubric_numeric_detailed_nonweighted.txt (R^2 = -0.4962)
4. WizardLM2 with rubric sphere_rubric_numeric_detailed_weighted.txt (R^2 = -0.4912)
5. Gemma with rubric sphere_rubric_text_detailed.txt (R^2 = -0.3813)

Best Overall Model

The best overall model, based on average R-squared across rubrics, is GPT-4- turbo-preview (R^2 = 0.7055)

Best Overall Rubric

The best overall rubric, based on average R-squared across models, is sphere_rubric_text_nondetailed.txt $R^2 = 0.5167$).

Notable Trends

Model Trend: Among the top performers, GPT-4-turbo-preview consistently shows high performance across different rubrics, indicating robustness in handling varied evaluation criteria.

Rubric Trend: Text-based rubrics, particularly those detailed or nondetailed, are associated with higher R-squared values compared to numeric rubrics, suggesting that they might capture aspects of the responses better in this dataset.

Analysis for scripts-complexity2 (Python Scripts, Complexity Level 2, numsummary)

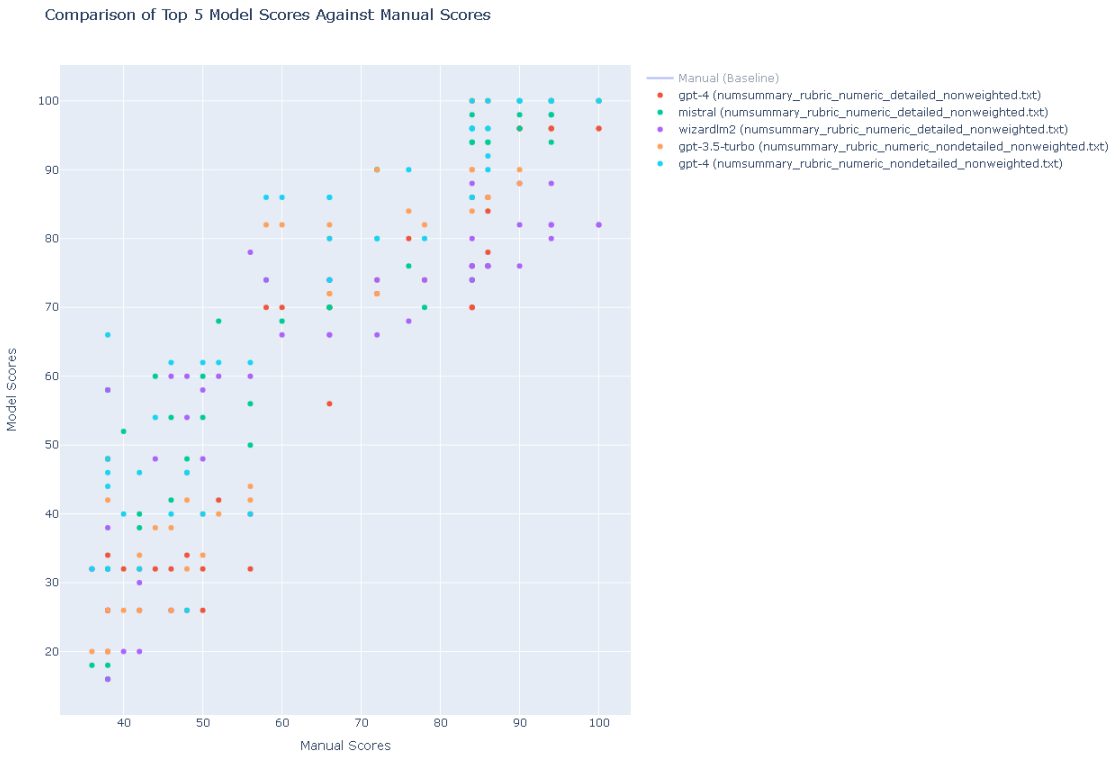

Figure 5. Achieved scores using different combinations of LLM and rubric format for scripts-complexity2-numsummary.

Top 5 Best Performing Model + Rubric Combinations

1. GPT-4 with rubric numsummary_rubric_numeric_detailed_nonweighted.txt ($R^2 = 0.7233$)
2. Mistral with rubric numsummary_rubric_numeric_detailed_nonweighted.txt ($R^2 = 0.7152$)
3. WizardLM2 with rubric numsummary_rubric_numeric_detailed_nonweighted.txt ($R^2 = 0.6921$)
4. GPT-3.5-Turbo with rubric numsummary_rubric_numeric_nondetailed_nonweighted.txt ($R^2 = 0.6827$)
5. GPT-4 with rubric numsummary_rubric_numeric_nondetailed_nonweighted.txt ($R^2 =$

0.6689)

Bottom 5 Worst Performing Model + Rubric Combinations

1. Llama2 with rubric numsummary_rubric_text_detailed.txt (R^2 = -10.2955)
2. Llama2 with rubric numsummary_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -1.0753)
3. Gemma with rubric numsummary_rubric_text_detailed.txt (R^2 = -0.9059)
4. GPT-4-Turbo-Preview with rubric numsummary_rubric_text_nondetailed.txt (R^2 = -0.7335)
5. Llama2 with rubric numsummary_rubric_text_nondetailed.txt (R^2 = -0.5577)

Best Overall Model

The best overall model, based on average R-squared across rubrics, is GPT-4.

Best Overall Rubric

The best overall rubric, based on average R-squared across models, is numsummary_rubric_numeric_detailed_nonweighted.txt.

Notable Trends

Model Trend: GPT-4 continues to perform well, suggesting consistency across different complexity levels in script analysis.

Rubric Trend: Numeric rubrics, especially those detailed, appear more effective at higher complexity levels, potentially due to their ability to capture specific numerical accuracies needed in the responses.

Analysis for scripts-complexity3 (Python Scripts, Complexity Level 3, calculator)

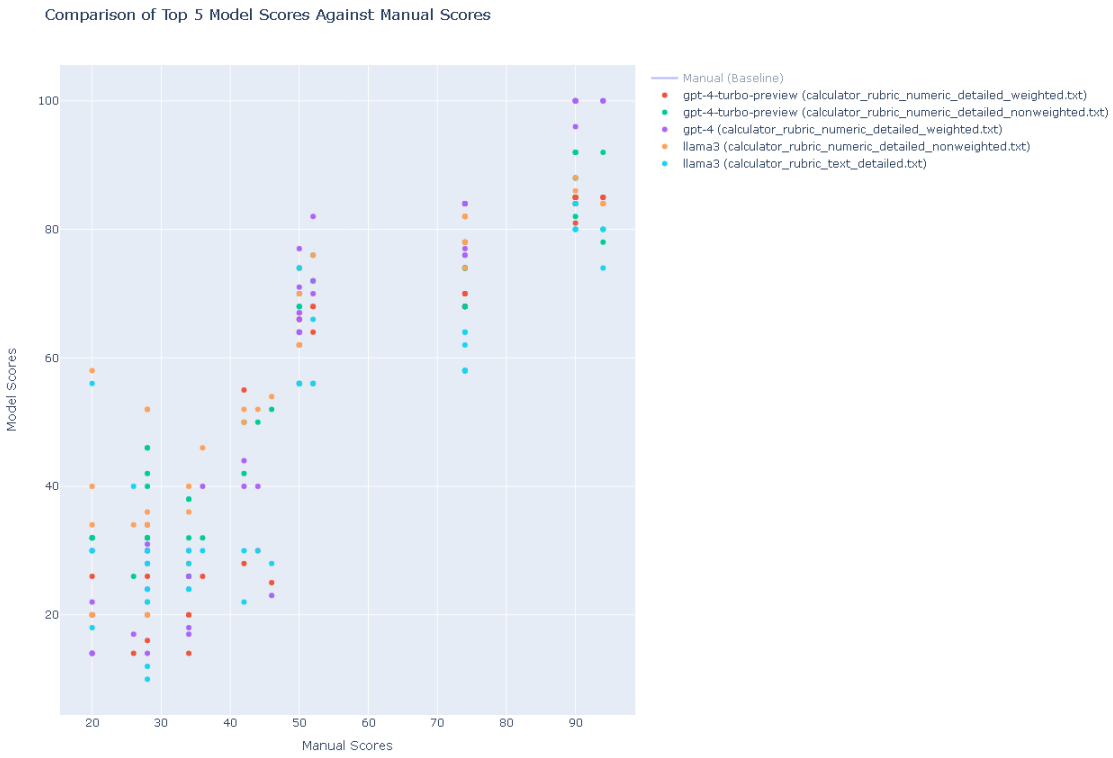

Figure 6. Achieved scores using different combinations of LLM and rubric format for scripts-complexity3-calculator.

Top 5 Best Performing Model + Rubric Combinations

1. GPT-4-Turbo-Preview with rubric calculator_rubric_numeric_detailed_weighted.txt (R^2 = 0.8210)
2. GPT-4-Turbo-Preview with rubric calculator_rubric_numeric_detailed_nonweighted.txt (R^2 = 0.7755)
3. GPT-4 with rubric calculator_rubric_numeric_detailed_weighted.txt (R^2 = 0.7590)
4. Llama3 with rubric calculator_rubric_numeric_detailed_nonweighted.txt (R^2 = 0.7415)
5. Llama3 with rubric calculator_rubric_text_detailed.txt (R^2 = 0.7261)

Bottom 5 Worst Performing Model + Rubric Combinations

1. Llama2 with rubric calculator_rubric_text_detailed.txt (R^2 = -2.6268)
2. Llama2 with rubric calculator_rubric_text_nondetailed.txt (R^2 = -1.1582)
3. WizardLM2 with rubric calculator_rubric_numeric_detailed_weighted.txt (R^2 = -0.9944)
4. Llama2 with rubric calculator_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -0.9578)
5. Llama2 with rubric calculator_rubric_numeric_detailed_nonweighted.txt (R^2 = -0.4386)

Best Overall Model

The best overall model, based on average R-squared across rubrics, is GPT-4-Turbo-Preview.

Best Overall Rubric

The best overall rubric, based on average R-squared across models, is calculator_rubric_numeric_nondetailed_weighted.txt.

Notable Trends

Model Trend: GPT-4-Turbo-Preview stands out in this complexity level, suggesting its potential for handling more complex scripting tasks with detailed numeric rubrics effectively.

Rubric Trend: Numeric rubrics, particularly those with detailed weightings, tend to perform better, indicating their suitability in accurately evaluating complex scripts that possibly involve calculations or data manipulations.

Analysis for scripts-complexity4 (Python Scripts, Complexity Level 4, zip)

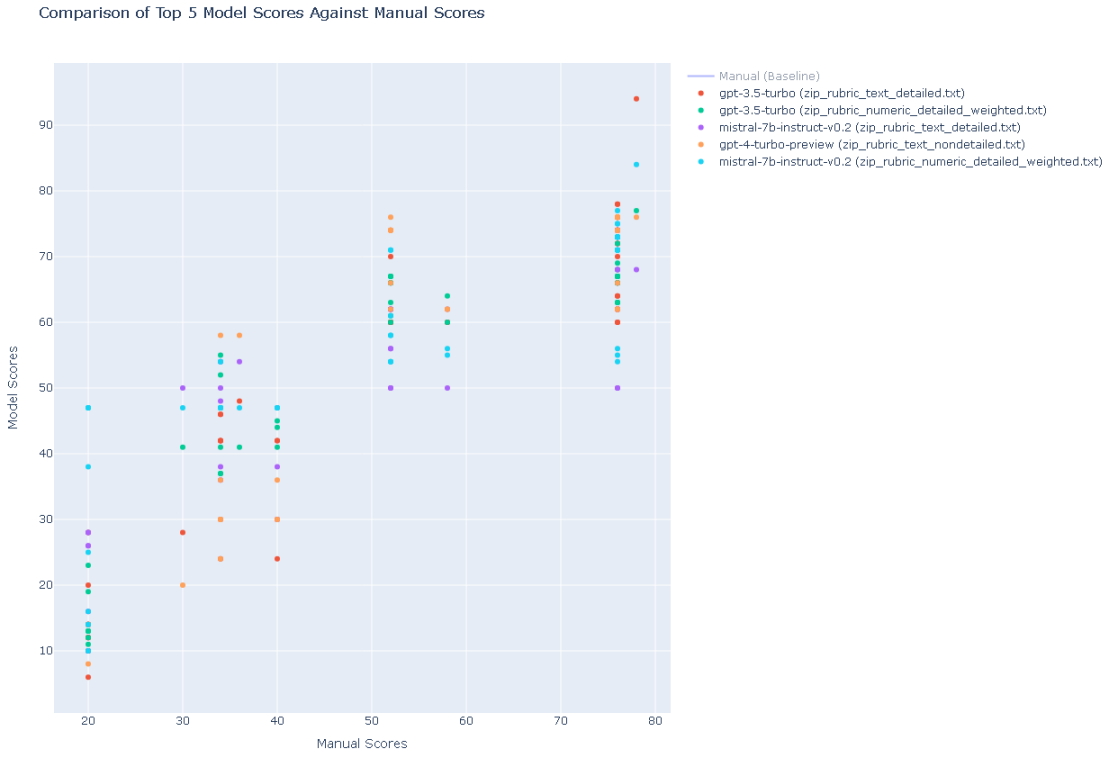

Figure 7. Achieved scores using different combinations of LLM and rubric format for scripts-complexity4-zip.

Top 5 Best Performing Model + Rubric Combinations

1. GPT-3.5-Turbo with rubric zip_rubric_text_detailed.txt (R^2 = 0.8407)
2. GPT-3.5-Turbo with rubric zip_rubric_numeric_detailed_weighted.txt (R^2 = 0.8111)
3. Mistral-7B-Instruct-V0.2 with rubric zip_rubric_text_detailed.txt (R^2 = 0.7608)
4. GPT-4-Turbo-Preview with rubric zip_rubric_text_nondetailed.txt (R^2 = 0.7340)
5. Mistral-7B-Instruct-V0.2 with rubric zip_rubric_numeric_detailed_nonweighted.txt (R^2 = 0.7001)

Bottom 5 Worst Performing Model + Rubric Combinations

1. Llama2 with rubric zip_rubric_text_detailed.txt (R^2 = -5.6874)
2. Gemma with rubric zip_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -2.7762)
3. Mistral-7B-Instruct-V0.2 with rubric zip_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -2.4797)
4. Llama2 with rubric zip_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -2.3901)
5. Mistral with rubric zip_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -2.3289)

Best Overall Model

The best overall model, based on average R-squared across rubrics, is GPT-3.5-Turbo.

Best Overall Rubric

The best overall rubric, based on average R-squared across models, is zip_rubric_text_nondetailed.txt.

Notable Trends

Model Trend: GPT-3.5-Turbo shows strong performance, suggesting its efficiency in handling complex scripting tasks that might involve file handling or operations reflected by the "zip" rubric context.

Rubric Trend: Text-based rubrics, especially those that are detailed, show superior performance, indicating their effectiveness in providing nuanced assessments that might capture broader criteria or qualitative aspects better in this complexity level.

Analysis for scripts-complexity5 (Python Scripts, Complexity Level 5, wordfreq)

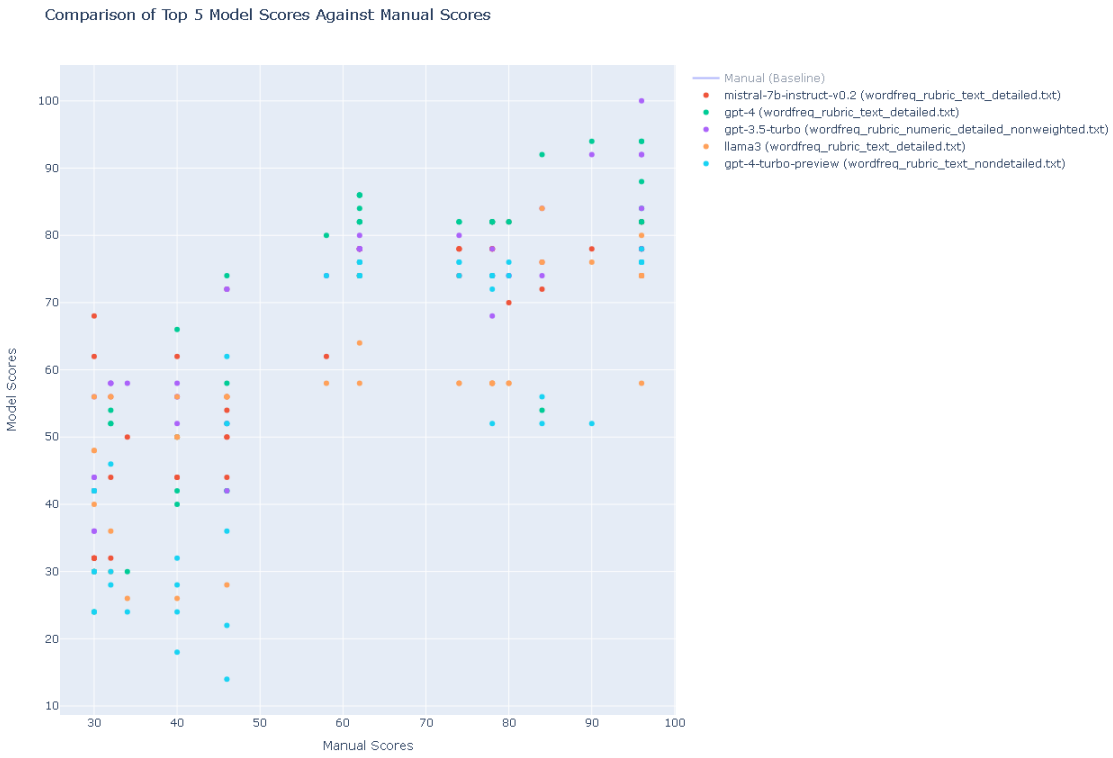

Figure 8. Achieved scores using different combinations of LLM and rubric format for scripts-complexity5-wordfreq.

Top 5 Best Performing Model + Rubric Combinations

1. Mistral-7B-Instruct-V0.2 with rubric wordfreq_rubric_text_detailed.txt (R^2 = 0.6296)
2. GPT-4 with rubric wordfreq_rubric_text_detailed.txt (R^2 = 0.5919)
3. GPT-3.5-Turbo with rubric wordfreq_rubric_numeric_detailed_nonweighted.txt (R^2 = 0.5884)
4. Llama3 with rubric wordfreq_rubric_text_detailed.txt (R^2 = 0.5384)
5. GPT-4-Turbo-Preview with rubric wordfreq_rubric_text_nondetailed.txt (R^2 = 0.5325)

Bottom 5 Worst Performing Model + Rubric Combinations

1. Llama2 with rubric wordfreq_rubric_text_detailed.txt (R^2 = -7.104)
2. Mistral-7B-Instruct-V0.2 with rubric wordfreq_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -2.2416)
3. Mistral-7B-Instruct-V0.2 with rubric wordfreq_rubric_numeric_nondetailed_weighted.txt (R^2 = -1.9908)
4. Mistral with rubric wordfreq_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -1.9225)
5. Mistral with rubric wordfreq_rubric_numeric_nondetailed_weighted.txt (R^2 = -1.8433)

Best Overall Model

The best overall model, based on average R-squared across rubrics, is GPT-4.

Best Overall Rubric

The best overall rubric, based on average R-squared across models, is wordfreq_rubric_text_nondetailed.txt.

Notable Trends

Model Trend: GPT-4 demonstrates strong performance, highlighting its robustness in handling complex scripts that involve frequency analysis or operations reflected by the "wordfreq" rubric context.

Rubric Trend: Text-based rubrics again appear to perform well, especially in more detailed evaluations, suggesting their effectiveness in capturing more nuanced aspects of script analysis.

Analysis for scripts-complexity6 (Python Scripts, Complexity Level 6, contactbook)

Figure 9. Achieved scores using different combinations of LLM and rubric format for scripts-complexity6-contactbook.

Top 5 Best Performing Model + Rubric Combinations

1. Llama3 with rubric contactbook_rubric_text_detailed.txt ($R^2 = 0.7843$)
2. GPT-3.5-Turbo with rubric contactbook_rubric_numeric_detailed_weighted.txt ($R^2 = 0.7385$)
3. GPT-4-Turbo-Preview with rubric contactbook_rubric_text_detailed.txt ($R^2 = 0.5974$)
4. Mistral with rubric contactbook_rubric_numeric_detailed_weighted.txt ($R^2 = 0.5712$)
5. Mistral-7B-Instruct-V0.2 with rubric contactbook_rubric_text_nondetailed.txt ($R^2 = 0.5707$)

Bottom 5 Worst Performing Model + Rubric Combinations

1. Mistral-7B-Instruct-V0.2 with rubric contactbook_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -3.8631)
2. Mistral with rubric contactbook_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -3.6408)
3. Llama2 with rubric contactbook_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -3.5022)
4. Gemma with rubric contactbook_rubric_numeric_nondetailed_nonweighted.txt (R^2 = -3.2103)
5. Llama2 with rubric contactbook_rubric_numeric_detailed_nonweighted.txt (R^2 = -3.0569)

Best Overall Model

The best overall model, based on average R-squared across rubrics, is GPT-4.

Best Overall Rubric

The best overall rubric, based on average R-squared across models, is contactbook_rubric_text_nondetailed.txt.

Notable Trends

Model Trend: Llama3 and GPT-3.5-Turbo perform well at this highest complexity level, indicating their strength in handling intricate script tasks that likely involve sophisticated operations or data handling.

Rubric Trend: Text-based rubrics, especially those that are detailed, continue to show good performance, suggesting their effectiveness in capturing complex interactions within scripts better than numeric rubrics.

Analysis for short_answers-complexity1 (Short Answers, Complexity Level 1, dronevsatellite)

Figure 10. Achieved scores using different combinations of LLM and rubric format for short_answeers-complexity1-dronevsatellite.

Top 5 Best Performing Model + Rubric Combinations

1. Mistral-7B-Instruct-V0.2 with rubric dronevsatellite_rubric_numeric_detailed_weighted.txt (R^2 = 0.7732)
2. GPT-4-Turbo-Preview with rubric dronevsatellite_rubric_numeric_detailed_nonweighted.txt (R^2 = 0.7565)
3. Llama3 with rubric dronevsatellite_rubric_numeric_detailed_nonweighted.txt (R^2 = 0.7483)
4. GPT-3.5-Turbo with rubric dronevsatellite_rubric_numeric_nondetailed_weighted.txt (R^2 = 0.7279)
5. WizardLM2 with rubric dronevsatellite_rubric_numeric_detailed_weighted.txt (R^2 = 0.7151)

Bottom 5 Worst Performing Model + Rubric Combinations

1. Llama2 with rubric dronevsatellite_rubric_text_detailed.txt (R^2 = -5.4198)
2. Llama2 with rubric dronevsatellite_rubric_numeric_detailed_nonweighted.txt (R^2 = -2.0140)
3. WizardLM2 with rubric dronevsatellite_rubric_numeric_nondetailed_weighted.txt (R^2 = -0.8983)
4. Gemma with rubric dronevsatellite_rubric_numeric_detailed_nonweighted.txt (R^2 = -0.6676)

5. GPT-3.5-Turbo with rubric dronevsatellite_rubric_numeric_detailed_weighted.txt (R^2 = -0.6609)

Best Overall Model

The best overall model, based on average R-squared across rubrics, is GPT-4-Turbo-Preview.

Best Overall Rubric

The best overall rubric, based on average R-squared across models, is dronevsatellite_rubric_numeric_nondetailed_nonweighted.txt.

Notable Trends

Model Trend: GPT-4-Turbo-Preview shows a strong performance across rubrics, suggesting it handles nuanced assessments well in short-answer formats at this complexity level.

Rubric Trend: Numeric rubrics, both detailed and non-detailed, perform well in evaluating short answers, indicating their effectiveness in assessing specific, quantifiable elements of responses.

Analysis for short_answers-complexity2 (Short Answers, Complexity Level 2, satellitecomvnoncom)

Figure 11. Achieved scores using different combinations of LLM and rubric format for short_answeers-complexity2-satellitecomvnoncom.

Top 5 Best Performing Model + Rubric Combinations

1. Llama3 with rubric satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt (R^2 = 0.9031)
2. GPT-4 with rubric satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt (R^2 =

0.8851)
3. GPT-4-Turbo-Preview with rubric satellitecomvnoncom_rubric_numeric_nondetailed_nonweighted.txt (R^2 = 0.8380)
4. GPT-4-Turbo-Preview with rubric satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt (R^2 = 0.8320)
5. GPT-4 with rubric satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt (R^2 = 0.8312)

BOTTOM 5 WORST PERFORMING MODEL + RUBRIC COMBINATIONS

1. Llama2 with rubric satellitecomvnoncom_rubric_text_detailed.txt (R^2 = -4.0410)
2. Llama2 with rubric satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt (R^2 = -1.2361)
3. GPT-3.5-Turbo with rubric satellitecomvnoncom_rubric_text_detailed.txt (R^2 = -0.7406)
4. WizardLM2 with rubric satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt (R^2 = -0.4891)
5. Llama2 with rubric satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt (R^2 = -0.1375)

BEST OVERALL MODEL

The best overall model, based on average R-squared across rubrics, is GPT-4.

BEST OVERALL RUBRIC

The best overall rubric, based on average R-squared across models, is satellitecomvnoncom_rubric_numeric_nondetailed_nonweighted.txt.

NOTABLE TRENDS

Model Trend: GPT-4 and its Turbo Preview version show strong performance across different rubrics, indicating their effectiveness in accurately assessing short-answer responses at a moderate complexity level.

Rubric Trend: Numeric rubrics, especially those that are nondetailed, tend to perform well, reflecting their suitability for efficiently capturing the essence of responses in short-answer assessments.

Analysis for short_answers-complexity3 (Short Answers, Complexity Level 3, robotsvboom)

Figure 12. Achieved scores using different combinations of LLM and rubric format for short_answeers-complexity3-robotsvboom.

Top 5 Best Performing Model + Rubric Combinations

1. GPT-4 with rubric robotsvboom_rubric_numeric_detailed_nonweighted.txt (R^2 = 0.9419)
2. Llama3 with rubric robotsvboom_rubric_numeric_detailed_nonweighted.txt (R^2 = 0.9323)
3. GPT-4-Turbo-Preview with rubric robotsvboom_rubric_numeric_nondetailed_nonweighted.txt (R^2 = 0.9282)
4. GPT-4 with rubric robotsvboom_rubric_numeric_nondetailed_weighted.txt (R^2 = 0.9081)
5. GPT-4-Turbo-Preview with rubric robotsvboom_rubric_numeric_nondetailed_weighted.txt (R^2 = 0.9041)

Bottom 5 Worst Performing Model + Rubric Combinations

1. Llama2 with rubric robotsvboom_rubric_text_detailed.txt (R^2 = -3.5528)
2. Llama2 with rubric robotsvboom_rubric_numeric_detailed_nonweighted.txt (R^2 = -0.2671)
3. Llama2 with rubric robotsvboom_rubric_numeric_nondetailed_weighted.txt (R^2 = -0.0329)
4. GPT-3.5-Turbo with rubric robotsvboom_rubric_text_detailed.txt (R^2 = -0.0013)
5. WizardLM2 with rubric robotsvboom_rubric_numeric_nondetailed_weighted.txt (R^2 = 0.1377)

Best Overall Model

The best overall model, based on average R-squared across rubrics, is GPT-4.

Best Overall Rubric

The best overall rubric, based on average R-squared across models, is robotsvboom_rubric_numeric_nondetailed_nonweighted.txt.

Notable Trends

Model Trend: GPT-4 and its Turbo Preview version consistently show strong performance, demonstrating their advanced capabilities in accurately assessing short-answer responses at higher complexity levels.

Rubric Trend: Numeric rubrics, particularly those that are non-detailed and non-weighted, tend to perform better, suggesting that they are effective at capturing essential aspects of responses without the need for detailed grading criteria.

Analysis for short_answers-complexity4 (Short Answers, Complexity Level 4, sensorhood)

Figure 13. Achieved scores using different combinations of LLM and rubric format for short_answeers-complexity4-sensorhood.

Top 5 Best Performing Model + Rubric Combinations

1. GPT-4-Turbo-Preview with rubric sensorhood_rubric_numeric_nondetailed_nonweighted.txt ($R^2 = 0.8206$)
2. Llama3 with rubric sensorhood_rubric_numeric_detailed_nonweighted.txt ($R^2 = 0.8147$)
3. GPT-4-Turbo-Preview with rubric sensorhood_rubric_numeric_detailed_nonweighted.txt ($R^2 = 0.7876$)
4. Mistral-7B-Instruct-V0.2 with rubric sensorhood_rubric_numeric_detailed_weighted.txt

($R^2 = 0.7801$)
5. Mistral with rubric sensorhood_rubric_numeric_detailed_weighted.txt ($R^2 = 0.7483$)

Bottom 5 Worst Performing Model + Rubric Combinations

1. Llama2 with rubric sensorhood_rubric_text_detailed.txt ($R^2 = -1.6257$)
2. Llama2 with rubric sensorhood_rubric_numeric_detailed_nonweighted.txt ($R^2 = -0.9138$)
3. Llama2 with rubric sensorhood_rubric_numeric_nondetailed_weighted.txt ($R^2 = -0.5567$)
4. GPT-3.5-Turbo with rubric sensorhood_rubric_text_detailed.txt ($R^2 = -0.3065$)
5. WizardLM2 with rubric sensorhood_rubric_numeric_nondetailed_weighted.txt ($R^2 = -0.0819$)

Best Overall Model

The best overall model, based on average R-squared across rubrics, is GPT-4-Turbo-Preview.

Best Overall Rubric

The best overall rubric, based on average R-squared across models, is sensorhood_rubric_numeric_nondetailed_nonweighted.txt.

Notable Trends

Model Trend: GPT-4-Turbo-Preview shows superior performance across various rubrics, suggesting its robustness in evaluating high-complexity short-answer responses.

Rubric Trend: Numeric rubrics, especially those that are non-detailed and non-weighted, demonstrate strong performance, highlighting their effectiveness in capturing essential aspects of complex short answers.

6. LIMITATIONS

HARDWARE CONSTRAINTS

- Storage constraints: Disk space became an issue when running multiple models locally, which limited the number of models that could be stored on disk.
- VRAM constraints: Running models trained on higher parameters (>10B) would have significantly increased response collection time and could potentially not run, limiting the scope of models that could be evaluated.

TECHNICAL LIMITATIONS

- Model performance: larger models typically behave better than their smaller counterparts, which may impact the accuracy of the results.
- Quantisation: All local models, except for "mistral-7b-instruct-v0.2", are 4-bit quantised, while the latter is 8-bit quantised. Lower quantisation can significantly constrain the models' performance (Li, 2023).
- Model response format: It was challenging to get the models to respond in a fixed format. Only the OpenAI models responded as requested. Other models' responses could easily be parsed into usable data. Some model and rubric combinations required each score to be manually entered after reading each response, defeating the purpose of this study. The average Levenshtein distance was found for all responses in each results file to identify which model and rubric combinations were guilty of this.

API AND LIBRARY LIMITATIONS

- OpenAI API usage limitations: OpenAI imposed API usage limitations, which limited the number of tests that could be run.
- OpenAI API limitations: Using OpenAI's API, it was impossible to ensure that each LLM prompt was passed as a brand-new prompt without the context of the previous prompt.
- Ollama API limitations: An attempt to use "format: JSON" (implemented uniquely in OpenAI-Python and Ollama-Python) was considered to produce a consistent result. However, this proved to cause the Python application to hang when getting Ollama models to use this format. This is discussed further in the thread [https://github.com/ollama/ollama/issues/3154]. Implementing fixes proved to be futile, as the issue behaved too inconsistently to debug after attempting different solutions to remedy the issue.

Bugs and Errors

- Irregular application hang: Another potential error was caused by irregular application hang. The solution to this was setting the "num_predict" parameter to a reasonable value, 2048, in our use case (for Ollama models). Documented in the thread [https://github.com/ollama/ollama/issues/2805], the model would hallucinate output tokens greater than the size of the context window, causing an endless context shift in an attempt to reallocate memory. Also discussed in the threads [https://github.com/ollama/ollama/issues/1863] and [https://github.com/ggerganov/llama.cpp/issues/3969].
- Context issue: Ensuring each LLM prompt (rubric and assessment piece) was passed as a brand-new prompt, i.e., did not have the context of the previous prompt, was challenging. There was no identifiable way to do this at the time using OpenAI's API. For Ollama models, the best way to do this was by removing the loaded model from memory. This was done by setting the "keep_alive" parameter to "0". Outlined in the Ollama/docs/API.md [https://github.com/ollama/ollama/blob/main/docs/api.md] and discovered from the thread [https://github.com/ollama/ollama/issues/2343].

Human Error and Bias

- Potential human error/bias: Manual assessment of test assessments may introduce subjective metrics and bias. Due to this human element, there is no way to ensure an unbiased assessment, as different people will have subjective metrics for what they consider a given quality. To minimise this, we had two passes for the manual score allocation, one to assign and the other to verify.

Cost and Resource Constraints

- Cost constraints: Much testing could not be conducted using OpenAI's higher-end models, as it would prove costly. Test runs had to remain limited due to cost concerns.
- Additional testing using other closed source LLMs such as Google’s Gemini models and Anthropic’s Claude models.

7. SUMMARY AND FUTURE DIRECTIONS

BEST PERFORMER FOR EACH QUESTION

Question	Model	Rubric	R^2
Scripts			
Q1	gpt-3.5-turbo	sphere_rubric_numeric_nondetailed_weighted.txt	0.7839
Q2	gpt-4	numsummary_rubric_numeric_detailed_nonweighted.txt	0.7233
Q3	gpt-4-turbo-preview	calculator_rubric_numeric_detailed_weighted.txt	0.8210
Q4	gpt-3.5-turbo	zip_rubric_text_detailed.txt	0.8410
Q5	mistral-7b-instruct-v0.2	wordfreq_rubric_text_detailed.txt	0.6296
Q6	llama3	contactbook_rubric_text_detailed.txt	0.7843
Short Answers			
Q1	mistral-7b-instruct-v0.2	dronevsatellite_rubric_numeric_detailed_weighted.txt	0.7732
Q2	llama3	satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt	0.9031
Q3	gpt-4	robotsvboom_rubric_numeric_detailed_nonweighted.txt	0.9419
Q4	gpt-4-turbo-preview	sensorhood_rubric_numeric_nondetailed_nonweighted.txt	0.8206

Table 2. Best model and rubric combination for each question

Summary

This comprehensive evaluation delves into the nuanced performance of various models and rubrics in automated assessment systems, focusing on Python scripts across six complexity levels and short-answer responses across four. Among the top performers, GPT-4 Turbo-Preview stands out for its robust handling of complex grading scenarios, especially with detailed numeric rubrics. This model, alongside GPT-4, exhibits high adaptability and effectiveness across various rubrics, suggesting an advanced ability to integrate linguistic analysis with quantitative evaluation criteria.

However, the study also highlights significant variability in performance, particularly among non-OpenAI models. Notably, models like Gemma and the Llama series (Llama2, Llama3, and Mistral) showed less consistent results. For instance, Llama2 consistently underperformed, especially with text-based rubrics, possibly due to its lower parameter count (7B) and basic 4-bit quantization, limiting its processing capabilities compared to more robust configurations. This resulted in the model consistently providing responses which were unable to be parsed, using the robust methods implemented. Comparatively, the other open source models, while also not responding as desired were still able to be parsed to gather the model scores.

The analysis also reveals that while numeric rubrics generally provide precise evaluations, they lack the descriptive depth that text-rich rubrics offer, which can significantly enhance a model's performance by leveraging its natural language processing capabilities. This is evident in the superior performance of text-based rubrics in more complex grading scenarios, which require a nuanced understanding of context and content beyond mere numeric accuracy.

Discussion on Model Performance

The best performing models, notably GPT-4 Turbo-Preview, often benefit from high parameter counts and sophisticated training regimens, although exact numbers for OpenAI's models are not publicly disclosed. These models excel in tasks requiring both depth of knowledge and breadth of language understanding, making them well-suited for complex educational assessments.

Conversely, the worst performers, such as Llama2 and Gemma, struggle with detailed evaluations, particularly in text-based assessments. Their lower performance could be correlated with their architecture's inherent limitations, such as fewer parameters and simpler quantization, which may restrict their ability to process and generate nuanced language effectively.

Future Directions

The path forward includes the development of hybrid rubrics that integrate the precision of numeric evaluations with the comprehensive analysis capabilities of text-based assessments. Such innovations could bridge the existing gaps between quantitative and linguistic evaluations, enhancing both the accuracy and reliability of automated assessments.

Further, there is a clear need for targeted improvements in model training, especially focusing on enhancing models' abilities to process numeric data and interpret complex text inputs. Tailoring rubrics to complement specific model capabilities can also optimize performance, particularly for non-OpenAI models that might benefit from customized assessment strategies to mitigate their inherent limitations.

This detailed analysis not only informs the strategic selection of model-rubric combinations for educational assessments but also highlights the importance of ongoing research into model

capabilities and rubric design to ensure that automated systems can approximate the nuanced judgments characteristic of skilled human evaluators.

ACKNOWLEDGEMENTS

I want to express my sincere gratitude to several individuals and institutions whose contributions were invaluable to completing this technical report.

Firstly, I want to highlight the significant contributions of Derek Long, my supervisor, who not only provided me with the opportunity to carry out this project during my work experience but also played a pivotal role in guiding its direction. His insights and our extensive discussions have been fundamental in shaping the methodology and the results presented in this report.

I am also grateful to Bernadette McCabe, who assisted with administrative tasks such as granting me elevated access to the necessary machines to install any required software.

My time at the University of Southern Queensland (UniSQ), specifically within the Agricultural Engineering Department, was productive and enjoyable thanks to the facilities provided, including a dedicated cubicle where I conducted much of my work. The environment and resources available at UniSQ were essential for completing my project.

I must also acknowledge the invaluable resources provided by the documentation writers at OpenAI and the creators of the Ollama documentation. These documents served as a foundational tool and guide throughout the project.

Lastly, I am thankful for the countless forums, blog posts, and discussions on large language models (LLMs) available online. These community resources have been a wellspring of knowledge and have significantly contributed to the success of this project.

Each contribution has been instrumental in bringing this project to fruition, and I am immensely thankful for the support and guidance I have received.

8. ETHICS STATEMENT

This study conducts an introductory analysis of the efficacy of using Large Language Models (LLMs) to automate assessment evaluation. It is essential to emphasise that the primary intention of this research is not to replace the evaluations conducted by qualified academic markers but rather to provide a supplementary tool to expedite the process. The methodology employed in this study is task-agnostic, meaning that the models used are not tailored or optimised for a particular function or application. Therefore, no foreseeable ethical implications are anticipated from the methods outlined in this report. However, as Gomez-Benito (2010) highlighted, it is crucial to acknowledge the potential for bias in manual marking. To mitigate this, a more meticulous process can be employed to minimise bias when marking test assessment pieces. Furthermore, the rapid development and availability of new LLMs pose a challenge in documenting them all. As the tools for utilising these models become more accessible and easier to integrate, understanding their capabilities and limitations will become increasingly important.

9. REPRODUCIBILITY STATEMENT

In the GitHub repository listed in the Abstract, the following materials are included:

- All scripts to conduct LLM response collection, score parsing, graphing, and analyses.
- The results obtained from these scripts, as well as all supplementary graphs/results.
- A README file providing sample commands and information on how to run all scripts. Any changes made will also be outlined here.

REFERENCES

Agaci, R 2017, 'Learning management systems in higher education', viewed <date-here>, https://api.semanticscholar.org/CorpusID:150707305.

Almeida, PS, Novais, P, Costa, E, Rodrigues, M & Neves, J 2008, 'Artificial intelligence tools for student learning assessment in professional schools', viewed <date-here>, https://api.semanticscholar.org/CorpusID:16663347.

Balse, R, Valaboju, B, Singhal, S, Warriem, JM & Prasad, P 2023, 'Investigating the potential of GPT-3 in providing feedback for programming assessments', Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1, viewed <date-here>, https://api.semanticscholar.org/CorpusID:259297695.

Barkaoui, K 2010, 'Explaining ESL essay holistic scores: a multilevel modeling approach', Language Testing, vol. 27, pp. 515-535, viewed <date-here>, https://api.semanticscholar.org/CorpusID:145070909.

Bulut, O & Wongvorachan, T 2022, 'Feedback generation through artificial intelligence', The Open/Technology in Education, Society, and Scholarship Association Conference, viewed <date-here>, https://api.semanticscholar.org/CorpusID:255122053.

Cai, B 2020, 'Intelligent marking system based on STM32', 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), pp. 961-965, viewed <date-here>, https://api.semanticscholar.org/CorpusID:222219936.

Caines, A, Benedetto, L, Taslimipoor, S, Davis, C, Gao, Y, Andersen, O, Yuan, Z, Elliott, M, Moore, R, Bryant, C, Rei, M, Yannakoudakis, H, Mullooly, A, Nicholls, D & Buttery, P 2023, 'On the application of large language models for language teaching and assessment technology', LLM@AIED, viewed <date-here>, https://api.semanticscholar.org/CorpusID:259937556.

Cavojský, M, Bugár, G, Kormaňík, T & Hasin, M 2023, 'Exploring the capabilities and possible applications of large language models for education', 2023 21st International Conference on Emerging eLearning Technologies and Applications (ICETA), pp. 91-98, viewed <date-here>, https://api.semanticscholar.org/CorpusID:266196844.

Chakraborty, S, Zhou, X, Hafeez-Baig, A, Gururajan, R, Paul, M, Mandal, A, Chacko, AE & Barua, PD 2016, 'Objective analysis of marker bias in higher education', 2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), pp. 453-457, viewed <date-here>, https://api.semanticscholar.org/CorpusID:13135526.

Chang, CK, Chen, PC, Chen, ZS & Kuo, TM 2023, 'Developing AI-based automated post-rating system to scaffold interdisciplinary knowledge-sharing', 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 239-241, viewed <date-here>, https://api.semanticscholar.org/CorpusID:263229552.Cunningham, HM 1984, 'Instruments bias in assessment centers', viewed <date-here>, https://api.semanticscholar.org/CorpusID:151917381.

Dennis, I, Newstead, SE & Wright, D 1996, 'A new approach to exploring biases in educational assessment', British Journal of Psychology, vol. 87 (Pt 4), pp. 515-34, viewed <date-here>,

https://api.semanticscholar.org/CorpusID:24207988.

Ellis, ME, Casey, KM & Hill, G 2024, 'ChatGPT and Python programming homework', Decision Sciences Journal of Innovative Education, viewed <date-here>, https://api.semanticscholar.org/CorpusID:267076352.

Gan, W, Qi, Z, Wu, J & Lin, CW 2023, 'Large language models in education: Vision and opportunities', 2023 IEEE International Conference on Big Data (BigData), pp. 4776-4785, viewed <date-here>, https://api.semanticscholar.org/CorpusID:265352038.

Gómez-Benito, J, Hidalgo, MD, Guilera, G & University of Barcelona 2010, 'Bias in measurement instruments. Fair tests' viewed <date-here>, https://api.semanticscholar.org/CorpusID:146279790.

Hooda, M, Rana, C, Dahiya, O, Rizwan, A & Hossain, MS 2022, 'Artificial intelligence for assessment and feedback to enhance student success in higher education', Mathematical Problems in Engineering, viewed <date-here>, https://api.semanticscholar.org/CorpusID:248609337.

Horoshko, YV, Shevchenko, T, Tsybko, H, Kostiuchenko, A & Tsybko, H 2019, 'Methodological approaches to the selection of learning management systems for use in the educational', viewed <date-here>, https://api.semanticscholar.org/CorpusID:201914218.

Kim, S, Hooper, C, Gholami, A, Dong, Z, Li, X, Shen, S, Mahoney, MW & Keutzer, K 2023, 'SqueezeLLM: Dense-and-sparse quantization', ArXiv, vol. abs/2306.07629, viewed <date-here>, https://api.semanticscholar.org/CorpusID:259144954.

Kocmi, T & Federmann, C 2023, 'Large language models are state-of-the-art evaluators of translation quality', European Association for Machine Translation Conferences/Workshops, viewed <date-here>, https://api.semanticscholar.org/CorpusID:257232490.

Koike, H, Akama, K, Morita, H & Miura, K 2006, 'Using an automatic marking system for programming courses', Proceedings of the 34th annual ACM SIGUCCS fall conference: expanding the boundaries, viewed <date-here>, https://api.semanticscholar.org/CorpusID:16399704.

Konstantinova, LV, Vorozhikhin, VV, Petrov, AM, Titova, ES & Shtykhno, D 2023, 'Generative artificial intelligence in education: Discussions and forecasts', Open Education, viewed <date-here>, https://api.semanticscholar.org/CorpusID:258561584.

Li, J, Li, S, Xu, J, Huang, S, Lian, Y, Liu, J, Wang, Y, Dai, G 2023, 'Enabling fast 2-bit LLM on GPUs: Memory alignment, sparse outlier, and asynchronous dequantization', ArXiv, vol. abs/2311.16442, viewed <date-here>, https://api.semanticscholar.org/CorpusID:265466134.

Miller, T 2007, 'Essay assessment 1 running head: essay assessment essay assessment with latent semantic analysis', viewed <date-here>, https://api.semanticscholar.org/CorpusID:267909409.

Naismith, B, Mulcaire, P & Burstein, J 2023, 'Automated evaluation of written discourse coherence using GPT-4', Workshop on Innovative Use of NLP for Building Educational Applications, viewed <date-here>, https://api.semanticscholar.org/CorpusID:259376861.

Ndukwe, IG, Amadi, CE, Nkomo, LM & Daniel, BK 2020, 'Automatic grading system using Sentence-BERT network', Artificial Intelligence in Education, vol. 12164, pp. 224-227, viewed <date-here>, https://api.semanticscholar.org/CorpusID:220365332.

Olowolayemo, A, Nawi, SD & Mantoro, T 2018, 'Short answer scoring in English grammar using text similarity measurement', 2018 International Conference on Computing, Engineering, and Design (ICCED), pp. 131-136, viewed <date-here>, https://api.semanticscholar.org/CorpusID:117729499.

Paľová, D 2016, 'Experience with usage of LMS moodle not only for the educational purposes at the educational institution', 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 901-906, viewed <date-here>, https://api.semanticscholar.org/CorpusID:15163676.

Phung, T, Cambronero, JP, Gulwani, S, Kohn, T, Majumdar, R, Singla, AK & Soares, G 2023, 'Generating high-precision feedback for programming syntax errors using large language models', ArXiv, vol. abs/2302.04662, viewed <date-here>, https://api.semanticscholar.org/CorpusID:256697110.

Ramamurthy, M & Krishnamurthi, I 2016, 'An automated assessment system for evaluation of students' answers using novel similarity measures', Research Journal of Applied Sciences, Engineering and Technology, vol. 12, pp. 258-263, viewed <date-here>, https://api.semanticscholar.org/CorpusID:56349529.

Rehder, B, Schreiner, ME, Wolfe, MBW, Laham, D, Landauer, TK & Kintsch, W 1998, 'Using latent semantic analysis to assess knowledge: some technical considerations', Discourse Processes, vol. 25, pp. 337-354, viewed <date-here>, https://api.semanticscholar.org/CorpusID:262556310.

Rodríguez-Hernández, CF, Musso, MF, Kyndt, E & Cascallar, EC 2021, 'Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation', Comput. Educ. Artif. Intell., vol. 2, pp. 100018, viewed <date-here>, https://api.semanticscholar.org/CorpusID:234941340.

Sampson, DG & Zervas, P 2012, 'Mobile learning management systems in higher education', viewed <date-here>, https://api.semanticscholar.org/CorpusID:107708907.

Sánchez Prieto, JC, Gamazo, A, Cruz-Benito, J, Therón, R & García-Peñalvo, FJ 2020, 'AI-driven assessment of students: Current uses and research trends', Interacción, viewed <date-here>, https://api.semanticscholar.org/CorpusID:220527744.

Shi, X 2002, 'Design of automatic marking device', Mechanical & Electrical Engineering Magazine, viewed <date-here>, https://api.semanticscholar.org/CorpusID:112388027.

Siva Balan, RV, Senthilnathan, T, Gobinath, R & Gondkar, RR 2023, 'Assessing academic performance using ensemble machine learning models', 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), pp. 917-924, viewed <date-here>, https://api.semanticscholar.org/CorpusID:262131632.

Subasri, R & Meenakumari, R 2023, 'AI based automatic mark entry system', viewed <date-here>, https://api.semanticscholar.org/CorpusID:259311055.

Sun, Q, Wu, J, Rong, W & Liu, W 2019, 'Formative assessment of programming language

learning based on peer code review: Implementation and experience report', Tsinghua Science and Technology, viewed <date-here>, https://api.semanticscholar.org/CorpusID:88498788.

Tubino, L & Adachi, C 2022, 'Developing feedback literacy capabilities through an AI automated feedback tool', ASCILITE Publications, viewed <date-here>, https://api.semanticscholar.org/CorpusID:253866057.

Valsamidis, SI, Kazanidis, I, Kontogiannis, S & Karakos, AS 2012, 'An approach for LMS assessment', International Journal of Technology Enhanced Learning, vol. 4, pp. 265-283, viewed <date-here>, https://api.semanticscholar.org/CorpusID:62155003.

Wadhwa, N, Pradhan, J, Sonwane, A, Sahu, SPS, Natarajan, N, Kanade, A, Parthasarathy, S & Rajamani, SK 2023, 'Frustrated with code quality issues? LLMs can help!', ArXiv, vol. abs/2309.12938, viewed <date-here>, https://api.semanticscholar.org/CorpusID:262216950.

Wangwiwattana, C & Tongvivat, Y 2022, 'Semi-automatic short-answer grading tools for Thai language using natural language processing', Proceedings of the 2022 5th International Conference on Education Technology Management, viewed <date-here>, https://api.semanticscholar.org/CorpusID:258508782.

Wangwiwattana, C & Tongvivat, Y 2023, 'Automating academic assessment: A large language model approach', 2023 7th International Conference on Information Technology (InCIT), pp. 330-334, viewed <date-here>, https://api.semanticscholar.org/CorpusID:267339050.

Xia, XT & Li, X 2022, 'Artificial intelligence for higher education development and teaching skills', Wireless Communications and Mobile Computing, viewed <date-here>, https://api.semanticscholar.org/CorpusID:248459594.

Yan, L, Sha, L, Zhao, L, Li, Y, Martínez-Maldonado, R, Chen, G, Li, X, Jin, Y & Gašević, D 2023, 'Practical and ethical challenges of large language models in education: A systematic scoping review', Br. J. Educ. Technol., vol. 55, pp. 90-112, viewed <date-here>, https://api.semanticscholar.org/CorpusID:260125084.

Yuan, Z, Niu, L, Liu, JW, Liu, W, Wang, X, Shang, Y, Sun, G, Wu, Q, Wu, JX & Wu, B 2023, 'RPTQ: Reorder-based post-training quantization for large language models', ArXiv, vol. abs/2304.01089, viewed <date-here>, https://api.semanticscholar.org/CorpusID:257913374.

Zapata-González, A & Luis, JL, 'The impact of generative artificial intelligence in higher education: a focus on ethics and academic integrity', viewed <date-here>, https://api.semanticscholar.org/CorpusID:266327916.

Zhang, L, Fei, W, Wu, W, He, Y, Lou, Z & Zhou, H 2023, 'Dual grained quantization: Efficient fine-grained quantization for LLM', ArXiv, vol. abs/2310.04836, viewed <date-here>, https://api.semanticscholar.org/CorpusID:263829729.

Zhao, WX, Zhou, K, Li, J, Tang, T, Wang, X, Hou, Y, Min, Y, Zhang, B, Zhang, J, Dong, Z, Du, Y, Yang, C, Chen, Y, Chen, Z, Jiang, J, Ren, R, Li, Y, Tang, X, Liu, Z, Liu, P, Nie, J & Wen, JR 2023, 'A survey of large language models', ArXiv, vol. abs/2303.18223, viewed <date-here>, https://api.semanticscholar.org/CorpusID:257900969.

A. APPENDIX

A.1. RUBRIC FORMAT TEMPLATES

NUMERIC_DETAILED_NONWEIGHTED

<table>
<tr><td colspan="2">{Script Question {1:6}} | {Short Answer Question {1:6}}

Rubric:</td></tr>
<tr><td>{script criteria}
{additional information provided for each criteria interval based on question}
Functionality: 1, 2, 3, 4, 5
Logic: 1, 2, 3, 4, 5
Code Quality: 1, 2, 3, 4, 5
User Input Handling: 1, 2, 3, 4, 5
Documentation: 1, 2, 3, 4, 5</td><td>{rubric criteria}
{additional information provided for each criteria interval based on question}
Understanding of the Topic: 1, 2, 3, 4, 5
Argumentation and Evidence: 1, 2, 3, 4, 5
Organization and Clarity: 1, 2, 3, 4, 5</td></tr>
<tr><td colspan="2">The {script | short answer} to be assessed is below:</td></tr>
</table>

Table 3. Rubric format for numeric, detailed, nonweighted rubrics.

NUMERIC_DETAILED_WEIGHTED

{Script Question {1:6}} | {Short Answer Question {1:6}} Rubric:	
{script criteria} {additional information provided for each criteria interval based on question} Functionality: 0-2, 3-4, 5-6, 7-8, 9-10 Logic: 0-6, 7-12, 13-18, 19-24, 25-30 Code Quality: 0-6, 7-12, 13-18, 19-24, 25-30 User Input Handling: 0-4, 5-8, 9-12, 13-16, 17-20 Documentation: 0-2, 3-4, 5-6, 7-8, 9-10	{rubric criteria} {additional information provided for each criteria interval based on question} Understanding of the Topic: 0-6, 7-12, 13-18, 18-24, 25-30 Argumentation and Evidence: 0-12, 13-24, 25-36, 37-48, 49-60 Organization and Clarity: 0-2, 3-4, 5-6, 7-8, 9-10
The {script | short answer} to be assessed is below:	

Table 4. Rubric format for numeric, detailed, weighted rubrics.

NUMERIC_NONDETAILED_NONWEIGHTED

{Script Question {1:6}} | {Short Answer Question {1:6}} Rubric:	
{script criteria} {no additional information provided aside from that below} Functionality: 1-5 Logic: 1-5 Code Quality: 1-5 User Input Handling: 1-5 Documentation: 1-5	{rubric criteria} {no additional information provided aside from that below} Understanding of the Topic: 1-5 Argumentation and Evidence: 1-5 Organization and Clarity: 1-5
The {script | short answer} to be assessed is below:	

Table 5. Rubric format for numeric, nondetailed, nonweighted rubrics.

NUMERIC_NONDETAILED_WEIGHTED

<table>
<tr><td colspan="2">{Script Question {1:6}} | {Short Answer Question {1:6}}

Rubric:</td></tr>
<tr><td>{script criteria}
{no additional information provided aside from that below}
Functionality: 0-10
Logic: 0-30
Code Quality: 0-30
User Input Handling: 0-20
Documentation: 0-10</td><td>{rubric criteria}
{no additional information provided aside from that below}
Understanding of the Topic: 0-30
Argumentation and Evidence: 0-60
Organization and Clarity: 0-10</td></tr>
<tr><td colspan="2">The {script | short answer} to be assessed is below:</td></tr>
</table>

Table 6. Rubric format for numeric, nondetailed, weighted rubrics.

TEXT_DETAILED

<table>
<tr><td colspan="2">{Script Question {1:6}} | {Short Answer Question {1:6}}

Rubric:</td></tr>
<tr><td>{script criteria}
{additional information provided for each criteria interval based on question}
Functionality: Poor, OK, Competent, Excellent, Perfect
Logic: Poor, OK, Competent, Excellent, Perfect
Code Quality: Poor, OK, Competent, Excellent, Perfect
User Input Handling: Poor, OK, Competent, Excellent, Perfect
Documentation: Poor, OK, Competent, Excellent, Perfect</td><td>{rubric criteria}
{additional information provided for each criteria interval based on question}
Understanding of the Topic: Poor, OK, Competent, Excellent, Perfect
Argumentation and Evidence: Poor, OK, Competent, Excellent, Perfect
Organization and Clarity: Poor, OK, Competent, Excellent, Perfect</td></tr>
<tr><td colspan="2">The {script | short answer} to be assessed is below:</td></tr>
</table>

Table 7. Rubric format for text, detailed rubrics.

TEXT_NONDETAILED

{Script Question {1:6}} | {Short Answer Question {1:6}} Rubric:	
{script criteria} {no additional information provided aside from that below} Functionality: Poor, OK, Competent, Excellent, Perfect Logic: Poor, OK, Competent, Excellent, Perfect Code Quality: Poor, OK, Competent, Excellent, Perfect User Input Handling: Poor, OK, Competent, Excellent, Perfect Documentation: Poor, OK, Competent, Excellent, Perfect	{rubric criteria} {no additional information provided aside from that below} Understanding of the Topic: Poor, OK, Competent, Excellent, Perfect Argumentation and Evidence: Poor, OK, Competent, Excellent, Perfect Organization and Clarity: Poor, OK, Competent, Excellent, Perfect
The {script | short answer} to be assessed is below:	

Table 8. Rubric format for text, nondetailed rubrics.

A.2. SCRIPT QUESTIONS

This section contains the questions used to create the synthetic scripts and short answers, as well as the rubrics used to assess them. “Complexities” is an artefact carried over from the initial preparation of this project; only scripts of increasing complexity were assessed.

In this assignment a Python script was to be developed which was capable of calculating the surface area and volume of a sphere given its radius.

Figure 14. Question for scripts-complexity1_sphere.

In this assignment a Python script was to be developed which creates a subroutine that returns the 5-number summary of an input array. The input array should be [9 27 81 86 23 30 57 31 53 0]. The script should contain the array as well as the subroutine

Figure 15. Question for scripts-complexity2_numsummary.

In this assignment a Python script was to be developed which creates a basic calculator program that performs operations like addition, subtraction, multiplication, and division based on user input.

Figure 16. Question for scripts-complexity3_calculator.

In this assignment a Python script was to be developed which creates subroutine that zips every folder from an input directory into a separate output directory. The files in directory "D:/pretendfolder/pretenddata" need to be zipped to the directory "D:/pretendfolder/pretendzippeddata"

Figure 17. Question for scripts-complexity4_zip.

In this assignment a Python script was to be developed which reads a text file, counts the occurrence of each word, and prints the most frequent words in descending order. The input text file, "wordfreq_input.txt" should be in the same directory as your script.

Figure 18. Question for scripts-complexity5_wordfreq.

In this assignment a Python script was to be developed which implemented a simple contact book application that allows users to add, delete, update, and search for contacts.

Figure 19. Question for scripts-complexity6_contactbook.

A.3. SHORT ANSWER QUESTIONS

In this assignment, a short answer response needs to be given for the following topic:

What are the advantages of using drone sensing over satellite sensing in agriculture?

The essay should be 1-2 paragraphs.

Primary factors to be addressed:

- Drones are capable of providing higher resolution images, reason being they are physically closer to the site.
- Drones are better able to meet the timing needs of image capture than of a satellite, i.e., the revisit capability.
- Drones do not need to be equipped with specific sensors.

Figure 20. Question for short_answers-complexity1_dronevsatellite.

In this assignment, a short answer response needs to be given for the following topic:

What are the advantages of using commercial satellite data over non-commercial satellite data in agriculture?

The essay should be 1-2 paragraphs.

Primary factors to be addressed:

- Commercial satellites are more cost effective.
- Commercial satellites are better able to meet the timing needs of image capture, i.e., revisit capability.
- Commercial satellites have better satellite resolution.

Figure 21. Question for short_answers-complexity2_satellitecomvnoncom.

In this assignment, a short answer response needs to be given for the following topic:

What are the advantages of using small specialty robots over a large boom sprayer in agriculture?

The essay should be 1-2 paragraphs.

Primary factors to be addressed:

- Using small specialty robots reduce soil compaction in the field.
- Using small specialty robots reduces labour use.
- Using small specialty robots increases savings on chemical use.

Figure 22. Question for short_answers-complexity3_robotsvboom.

In this assignment, a short answer response needs to be given for the following topic:

Why did all first generation weed spot and spray sensors have a hood?

The essay should be 1-2 paragraphs.

Ideal answer:

"All first-generation spot and spray sensors had a hood to ensure consistent lighting conditions regardless of time of day. The eased software challenges faced when training early models, as it reduced the number of images by eliminating the need to capture images at different times of day."

Figure 23. Question for short_answers-complexity4_sensorhood.

A.4. R^2 SCORES

Model	Rubric	R-Squared Score
gemma	sphere_rubric_numeric_detailed_nonweighted.txt	-0.477
gemma	sphere_rubric_numeric_detailed_weighted.txt	-0.2927
gemma	sphere_rubric_numeric_nondetailed_nonweighted.txt	0.0507
gemma	sphere_rubric_numeric_nondetailed_weighted.txt	0.3373
gemma	sphere_rubric_text_detailed.txt	-0.3621
gemma	sphere_rubric_text_nondetailed.txt	0.2261
gpt-3.5-turbo	sphere_rubric_numeric_detailed_nonweighted.txt	0.7016
gpt-3.5-turbo	sphere_rubric_numeric_detailed_weighted.txt	0.6076
gpt-3.5-turbo	sphere_rubric_numeric_nondetailed_nonweighted.txt	0.5442
gpt-3.5-turbo	sphere_rubric_numeric_nondetailed_weighted.txt	0.7854
gpt-3.5-turbo	sphere_rubric_text_detailed.txt	0.6588
gpt-3.5-turbo	sphere_rubric_text_nondetailed.txt	0.7415
gpt-4	sphere_rubric_numeric_detailed_nonweighted.txt	0.6723
gpt-4	sphere_rubric_numeric_detailed_weighted.txt	0.7332
gpt-4	sphere_rubric_numeric_nondetailed_nonweighted.txt	0.4639
gpt-4	sphere_rubric_numeric_nondetailed_weighted.txt	0.3492
gpt-4	sphere_rubric_text_detailed.txt	0.7702
gpt-4	sphere_rubric_text_nondetailed.txt	0.77
gpt-4-turbo-preview	sphere_rubric_numeric_detailed_nonweighted.txt	0.7505
gpt-4-turbo-preview	sphere_rubric_numeric_detailed_weighted.txt	0.7533
gpt-4-turbo-preview	sphere_rubric_numeric_nondetailed_nonweighted.txt	0.6554
gpt-4-turbo-preview	sphere_rubric_numeric_nondetailed_weighted.txt	0.6958
gpt-4-turbo-preview	sphere_rubric_text_detailed.txt	0.723
gpt-4-turbo-preview	sphere_rubric_text_nondetailed.txt	0.6652
llama2	sphere_rubric_numeric_detailed_nonweighted.txt	-0.3674
llama2	sphere_rubric_numeric_detailed_weighted.txt	0.1737
llama2	sphere_rubric_numeric_nondetailed_nonweighted.txt	-0.6971
llama2	sphere_rubric_numeric_nondetailed_weighted.txt	0.3644
llama2	sphere_rubric_text_detailed.txt	-2.4818
llama2	sphere_rubric_text_nondetailed.txt	0.1838
llama3	sphere_rubric_numeric_detailed_nonweighted.txt	0.4715
llama3	sphere_rubric_numeric_detailed_weighted.txt	0.6917
llama3	sphere_rubric_numeric_nondetailed_nonweighted.txt	0.2372
llama3	sphere_rubric_numeric_nondetailed_weighted.txt	0.2388
llama3	sphere_rubric_text_detailed.txt	0.7487
llama3	sphere_rubric_text_nondetailed.txt	0.5894
mistral	sphere_rubric_numeric_detailed_nonweighted.txt	0.4277
mistral	sphere_rubric_numeric_detailed_weighted.txt	0.5405
mistral	sphere_rubric_numeric_nondetailed_nonweighted.txt	-0.0226
mistral	sphere_rubric_numeric_nondetailed_weighted.txt	0.3545
mistral	sphere_rubric_text_detailed.txt	0.6256
mistral	sphere_rubric_text_nondetailed.txt	0.5182
mistral-7b-instruct-v0.2	sphere_rubric_numeric_detailed_nonweighted.txt	-0.0099
mistral-7b-instruct-v0.2	sphere_rubric_numeric_detailed_weighted.txt	0.5135
mistral-7b-instruct-v0.2	sphere_rubric_numeric_nondetailed_nonweighted.txt	-0.0537
mistral-7b-instruct-v0.2	sphere_rubric_numeric_nondetailed_weighted.txt	0.0782
mistral-7b-instruct-v0.2	sphere_rubric_text_detailed.txt	0.6995
mistral-7b-instruct-v0.2	sphere_rubric_text_nondetailed.txt	0.5827
wizardlm2	sphere_rubric_numeric_detailed_nonweighted.txt	-0.183
wizardlm2	sphere_rubric_numeric_detailed_weighted.txt	-0.4805
wizardlm2	sphere_rubric_numeric_nondetailed_nonweighted.txt	-0.1855
wizardlm2	sphere_rubric_numeric_nondetailed_weighted.txt	0.1567
wizardlm2	sphere_rubric_text_detailed.txt	0.2031
wizardlm2	sphere_rubric_text_nondetailed.txt	0.3808

Table 9. R-squared scores for scripts-complexity1_sphere.

Model	Rubric	R-Squared Score
gemma	numsummary_rubric_numeric_detailed_nonweighted.txt	-0.2793
gemma	numsummary_rubric_numeric_detailed_weighted.txt	0.2323
gemma	numsummary_rubric_numeric_nondetailed_nonweighted.txt	0.2158
gemma	numsummary_rubric_numeric_nondetailed_weighted.txt	0.5744
gemma	numsummary_rubric_text_detailed.txt	-0.8985
gemma	numsummary_rubric_text_nondetailed.txt	0.5332
gpt-3.5-turbo	numsummary_rubric_numeric_detailed_nonweighted.txt	0.5199
gpt-3.5-turbo	numsummary_rubric_numeric_detailed_weighted.txt	-0.1974
gpt-3.5-turbo	numsummary_rubric_numeric_nondetailed_nonweighted.txt	0.6844
gpt-3.5-turbo	numsummary_rubric_numeric_nondetailed_weighted.txt	0.3307
gpt-3.5-turbo	numsummary_rubric_text_detailed.txt	-0.1039
gpt-3.5-turbo	numsummary_rubric_text_nondetailed.txt	0.042
gpt-4	numsummary_rubric_numeric_detailed_nonweighted.txt	0.7229
gpt-4	numsummary_rubric_numeric_detailed_weighted.txt	0.4068
gpt-4	numsummary_rubric_numeric_nondetailed_nonweighted.txt	0.6766
gpt-4	numsummary_rubric_numeric_nondetailed_weighted.txt	0.5936
gpt-4	numsummary_rubric_text_detailed.txt	0.012
gpt-4	numsummary_rubric_text_nondetailed.txt	0.2781
gpt-4-turbo-preview	numsummary_rubric_numeric_detailed_nonweighted.txt	0.454
gpt-4-turbo-preview	numsummary_rubric_numeric_detailed_weighted.txt	0.0466
gpt-4-turbo-preview	numsummary_rubric_numeric_nondetailed_nonweighted.txt	0.5878
gpt-4-turbo-preview	numsummary_rubric_numeric_nondetailed_weighted.txt	0.3701
gpt-4-turbo-preview	numsummary_rubric_text_detailed.txt	-0.3539
gpt-4-turbo-preview	numsummary_rubric_text_nondetailed.txt	-0.718
llama2	numsummary_rubric_numeric_detailed_nonweighted.txt	0.0029
llama2	numsummary_rubric_numeric_detailed_weighted.txt	0.1407
llama2	numsummary_rubric_numeric_nondetailed_nonweighted.txt	-1.0717
llama2	numsummary_rubric_numeric_nondetailed_weighted.txt	-0.4921
llama2	numsummary_rubric_text_detailed.txt	-10.073
llama2	numsummary_rubric_text_nondetailed.txt	-0.5548
llama3	numsummary_rubric_numeric_detailed_nonweighted.txt	0.5943
llama3	numsummary_rubric_numeric_detailed_weighted.txt	0.416
llama3	numsummary_rubric_numeric_nondetailed_nonweighted.txt	0.4247
llama3	numsummary_rubric_numeric_nondetailed_weighted.txt	0.3657
llama3	numsummary_rubric_text_detailed.txt	-0.211
llama3	numsummary_rubric_text_nondetailed.txt	-0.2684
mistral	numsummary_rubric_numeric_detailed_nonweighted.txt	0.714
mistral	numsummary_rubric_numeric_detailed_weighted.txt	-0.1767
mistral	numsummary_rubric_numeric_nondetailed_nonweighted.txt	0.4286
mistral	numsummary_rubric_numeric_nondetailed_weighted.txt	0.0857
mistral	numsummary_rubric_text_detailed.txt	-0.5316
mistral	numsummary_rubric_text_nondetailed.txt	0.141
mistral-7b-instruct-v0.2	numsummary_rubric_numeric_detailed_nonweighted.txt	0.5406
mistral-7b-instruct-v0.2	numsummary_rubric_numeric_detailed_weighted.txt	-0.0001
mistral-7b-instruct-v0.2	numsummary_rubric_numeric_nondetailed_nonweighted.txt	0.3366
mistral-7b-instruct-v0.2	numsummary_rubric_numeric_nondetailed_weighted.txt	-0.0541
mistral-7b-instruct-v0.2	numsummary_rubric_text_detailed.txt	-0.4082
mistral-7b-instruct-v0.2	numsummary_rubric_text_nondetailed.txt	0.3178
wizardlm2	numsummary_rubric_numeric_detailed_nonweighted.txt	0.6939
wizardlm2	numsummary_rubric_numeric_detailed_weighted.txt	-0.3528
wizardlm2	numsummary_rubric_numeric_nondetailed_nonweighted.txt	0.5165
wizardlm2	numsummary_rubric_numeric_nondetailed_weighted.txt	0.3662
wizardlm2	numsummary_rubric_text_detailed.txt	0.4734
wizardlm2	numsummary_rubric_text_nondetailed.txt	0.2548

Table 10. R-squared scores for scripts-complexity2_numsummary.

Model	Rubric	R-Squared Score
gemma	calculator_rubric_numeric_detailed_nonweighted.txt	0.0253
gemma	calculator_rubric_numeric_detailed_weighted.txt	-0.3315
gemma	calculator_rubric_numeric_nondetailed_nonweighted.txt	0.3009
gemma	calculator_rubric_numeric_nondetailed_weighted.txt	0.4424
gemma	calculator_rubric_text_detailed.txt	0.0576
gemma	calculator_rubric_text_nondetailed.txt	0.6522
gpt-3.5-turbo	calculator_rubric_numeric_detailed_nonweighted.txt	0.6243
gpt-3.5-turbo	calculator_rubric_numeric_detailed_weighted.txt	0.5732
gpt-3.5-turbo	calculator_rubric_numeric_nondetailed_nonweighted.txt	0.3746
gpt-3.5-turbo	calculator_rubric_numeric_nondetailed_weighted.txt	0.6597
gpt-3.5-turbo	calculator_rubric_text_detailed.txt	0.6061
gpt-3.5-turbo	calculator_rubric_text_nondetailed.txt	0.6351
gpt-4	calculator_rubric_numeric_detailed_nonweighted.txt	0.6029
gpt-4	calculator_rubric_numeric_detailed_weighted.txt	0.7736
gpt-4	calculator_rubric_numeric_nondetailed_nonweighted.txt	0.2941
gpt-4	calculator_rubric_numeric_nondetailed_weighted.txt	0.1743
gpt-4	calculator_rubric_text_detailed.txt	0.625
gpt-4	calculator_rubric_text_nondetailed.txt	0.6174
gpt-4-turbo-preview	calculator_rubric_numeric_detailed_nonweighted.txt	0.7946
gpt-4-turbo-preview	calculator_rubric_numeric_detailed_weighted.txt	0.8294
gpt-4-turbo-preview	calculator_rubric_numeric_nondetailed_nonweighted.txt	0.7251
gpt-4-turbo-preview	calculator_rubric_numeric_nondetailed_weighted.txt	0.7235
gpt-4-turbo-preview	calculator_rubric_text_detailed.txt	0.7041
gpt-4-turbo-preview	calculator_rubric_text_nondetailed.txt	0.6155
llama2	calculator_rubric_numeric_detailed_nonweighted.txt	-0.4239
llama2	calculator_rubric_numeric_detailed_weighted.txt	0.3565
llama2	calculator_rubric_numeric_nondetailed_nonweighted.txt	-0.9005
llama2	calculator_rubric_numeric_nondetailed_weighted.txt	-0.0783
llama2	calculator_rubric_text_detailed.txt	-2.6172
llama2	calculator_rubric_text_nondetailed.txt	-1.1583
llama3	calculator_rubric_numeric_detailed_nonweighted.txt	0.7607
llama3	calculator_rubric_numeric_detailed_weighted.txt	0.7325
llama3	calculator_rubric_numeric_nondetailed_nonweighted.txt	0.5204
llama3	calculator_rubric_numeric_nondetailed_weighted.txt	0.649
llama3	calculator_rubric_text_detailed.txt	0.7452
llama3	calculator_rubric_text_nondetailed.txt	0.3346
mistral	calculator_rubric_numeric_detailed_nonweighted.txt	0.686
mistral	calculator_rubric_numeric_detailed_weighted.txt	0.4399
mistral	calculator_rubric_numeric_nondetailed_nonweighted.txt	0.2532
mistral	calculator_rubric_numeric_nondetailed_weighted.txt	0.6708
mistral	calculator_rubric_text_detailed.txt	0.6091
mistral	calculator_rubric_text_nondetailed.txt	0.5773
mistral-7b-instruct-v0.2	calculator_rubric_numeric_detailed_nonweighted.txt	0.5938
mistral-7b-instruct-v0.2	calculator_rubric_numeric_detailed_weighted.txt	0.6602
mistral-7b-instruct-v0.2	calculator_rubric_numeric_nondetailed_nonweighted.txt	0.0196
mistral-7b-instruct-v0.2	calculator_rubric_numeric_nondetailed_weighted.txt	0.4717
mistral-7b-instruct-v0.2	calculator_rubric_text_detailed.txt	0.5833
mistral-7b-instruct-v0.2	calculator_rubric_text_nondetailed.txt	0.6471
wizardlm2	calculator_rubric_numeric_detailed_nonweighted.txt	0.2902
wizardlm2	calculator_rubric_numeric_detailed_weighted.txt	-0.9552
wizardlm2	calculator_rubric_numeric_nondetailed_nonweighted.txt	0.4317
wizardlm2	calculator_rubric_numeric_nondetailed_weighted.txt	0.6235
wizardlm2	calculator_rubric_text_detailed.txt	0.5619
wizardlm2	calculator_rubric_text_nondetailed.txt	0.6692

Table 11. R-squared scores for scripts-complexity3_calculator.

Model	Rubric	R-Squared Score
gemma	zip_rubric_numeric_detailed_nonweighted.txt	-2.1067
gemma	zip_rubric_numeric_detailed_weighted.txt	-1.4704
gemma	zip_rubric_numeric_nondetailed_nonweighted.txt	-2.7096
gemma	zip_rubric_numeric_nondetailed_weighted.txt	-0.9466
gemma	zip_rubric_text_detailed.txt	-1.7816
gemma	zip_rubric_text_nondetailed.txt	-0.3836
gpt-3.5-turbo	zip_rubric_numeric_detailed_nonweighted.txt	0.2957
gpt-3.5-turbo	zip_rubric_numeric_detailed_weighted.txt	0.8205
gpt-3.5-turbo	zip_rubric_numeric_nondetailed_nonweighted.txt	-0.0909
gpt-3.5-turbo	zip_rubric_numeric_nondetailed_weighted.txt	0.2217
gpt-3.5-turbo	zip_rubric_text_detailed.txt	0.8434
gpt-3.5-turbo	zip_rubric_text_nondetailed.txt	0.6912
gpt-4	zip_rubric_numeric_detailed_nonweighted.txt	-0.2004
gpt-4	zip_rubric_numeric_detailed_weighted.txt	-0.1111
gpt-4	zip_rubric_numeric_nondetailed_nonweighted.txt	-1.0153
gpt-4	zip_rubric_numeric_nondetailed_weighted.txt	-1.3362
gpt-4	zip_rubric_text_detailed.txt	0.3597
gpt-4	zip_rubric_text_nondetailed.txt	0.5327
gpt-4-turbo-preview	zip_rubric_numeric_detailed_nonweighted.txt	0.3622
gpt-4-turbo-preview	zip_rubric_numeric_detailed_weighted.txt	0.4237
gpt-4-turbo-preview	zip_rubric_numeric_nondetailed_nonweighted.txt	-0.3578
gpt-4-turbo-preview	zip_rubric_numeric_nondetailed_weighted.txt	-0.2943
gpt-4-turbo-preview	zip_rubric_text_detailed.txt	0.7188
gpt-4-turbo-preview	zip_rubric_text_nondetailed.txt	0.7541
llama2	zip_rubric_numeric_detailed_nonweighted.txt	-0.9452
llama2	zip_rubric_numeric_detailed_weighted.txt	0.1203
llama2	zip_rubric_numeric_nondetailed_nonweighted.txt	-2.342
llama2	zip_rubric_numeric_nondetailed_weighted.txt	-1.2216
llama2	zip_rubric_text_detailed.txt	-5.5748
llama2	zip_rubric_text_nondetailed.txt	-0.3025
llama3	zip_rubric_numeric_detailed_nonweighted.txt	0.0512
llama3	zip_rubric_numeric_detailed_weighted.txt	-0.1445
llama3	zip_rubric_numeric_nondetailed_nonweighted.txt	-0.5615
llama3	zip_rubric_numeric_nondetailed_weighted.txt	-0.7177
llama3	zip_rubric_text_detailed.txt	0.5196
llama3	zip_rubric_text_nondetailed.txt	0.5567
mistral	zip_rubric_numeric_detailed_nonweighted.txt	0.0811
mistral	zip_rubric_numeric_detailed_weighted.txt	0.6625
mistral	zip_rubric_numeric_nondetailed_nonweighted.txt	-2.2483
mistral	zip_rubric_numeric_nondetailed_weighted.txt	-1.0607
mistral	zip_rubric_text_detailed.txt	0.4001
mistral	zip_rubric_text_nondetailed.txt	0.4604
mistral-7b-instruct-v0.2	zip_rubric_numeric_detailed_nonweighted.txt	-0.3696
mistral-7b-instruct-v0.2	zip_rubric_numeric_detailed_weighted.txt	0.7016
mistral-7b-instruct-v0.2	zip_rubric_numeric_nondetailed_nonweighted.txt	-2.3991
mistral-7b-instruct-v0.2	zip_rubric_numeric_nondetailed_weighted.txt	-1.2098
mistral-7b-instruct-v0.2	zip_rubric_text_detailed.txt	0.765
mistral-7b-instruct-v0.2	zip_rubric_text_nondetailed.txt	0.2535
wizardlm2	zip_rubric_numeric_detailed_nonweighted.txt	-0.3825
wizardlm2	zip_rubric_numeric_detailed_weighted.txt	0.3268
wizardlm2	zip_rubric_numeric_nondetailed_nonweighted.txt	-1.2085
wizardlm2	zip_rubric_numeric_nondetailed_weighted.txt	-1.0007
wizardlm2	zip_rubric_text_detailed.txt	0.5127
wizardlm2	zip_rubric_text_nondetailed.txt	0.5056

Table 12. R-squared scores for scripts-complexity4_zip.

Model	Rubric	R-Squared Score
gemma	wordfreq_rubric_numeric_detailed_nonweighted.txt	-0.588
gemma	wordfreq_rubric_numeric_detailed_weighted.txt	-0.7751
gemma	wordfreq_rubric_numeric_nondetailed_nonweighted.txt	-1.5887
gemma	wordfreq_rubric_numeric_nondetailed_weighted.txt	-0.7422
gemma	wordfreq_rubric_text_detailed.txt	-0.6679
gemma	wordfreq_rubric_text_nondetailed.txt	-0.2174
gpt-3.5-turbo	wordfreq_rubric_numeric_detailed_nonweighted.txt	0.5982
gpt-3.5-turbo	wordfreq_rubric_numeric_detailed_weighted.txt	0.2737
gpt-3.5-turbo	wordfreq_rubric_numeric_nondetailed_nonweighted.txt	0.3482
gpt-3.5-turbo	wordfreq_rubric_numeric_nondetailed_weighted.txt	0.395
gpt-3.5-turbo	wordfreq_rubric_text_detailed.txt	0.3098
gpt-3.5-turbo	wordfreq_rubric_text_nondetailed.txt	0.1589
gpt-4	wordfreq_rubric_numeric_detailed_nonweighted.txt	0.1832
gpt-4	wordfreq_rubric_numeric_detailed_weighted.txt	0.4483
gpt-4	wordfreq_rubric_numeric_nondetailed_nonweighted.txt	-0.5121
gpt-4	wordfreq_rubric_numeric_nondetailed_weighted.txt	-0.7683
gpt-4	wordfreq_rubric_text_detailed.txt	0.6073
gpt-4	wordfreq_rubric_text_nondetailed.txt	0.4997
gpt-4-turbo-preview	wordfreq_rubric_numeric_detailed_nonweighted.txt	0.4891
gpt-4-turbo-preview	wordfreq_rubric_numeric_detailed_weighted.txt	0.4223
gpt-4-turbo-preview	wordfreq_rubric_numeric_nondetailed_nonweighted.txt	-0.1075
gpt-4-turbo-preview	wordfreq_rubric_numeric_nondetailed_weighted.txt	-0.1191
gpt-4-turbo-preview	wordfreq_rubric_text_detailed.txt	0.4816
gpt-4-turbo-preview	wordfreq_rubric_text_nondetailed.txt	0.54
llama2	wordfreq_rubric_numeric_detailed_nonweighted.txt	-0.3088
llama2	wordfreq_rubric_numeric_detailed_weighted.txt	0.1623
llama2	wordfreq_rubric_numeric_nondetailed_nonweighted.txt	-1.2381
llama2	wordfreq_rubric_numeric_nondetailed_weighted.txt	-0.4645
llama2	wordfreq_rubric_text_detailed.txt	-6.9566
llama2	wordfreq_rubric_text_nondetailed.txt	-0.0266
llama3	wordfreq_rubric_numeric_detailed_nonweighted.txt	0.3797
llama3	wordfreq_rubric_numeric_detailed_weighted.txt	0.4666
llama3	wordfreq_rubric_numeric_nondetailed_nonweighted.txt	-0.1713
llama3	wordfreq_rubric_numeric_nondetailed_weighted.txt	-0.2036
llama3	wordfreq_rubric_text_detailed.txt	0.544
llama3	wordfreq_rubric_text_nondetailed.txt	-0.005
mistral	wordfreq_rubric_numeric_detailed_nonweighted.txt	-0.2142
mistral	wordfreq_rubric_numeric_detailed_weighted.txt	0.1849
mistral	wordfreq_rubric_numeric_nondetailed_nonweighted.txt	-1.8728
mistral	wordfreq_rubric_numeric_nondetailed_weighted.txt	-1.7937
mistral	wordfreq_rubric_text_detailed.txt	0.4481
mistral	wordfreq_rubric_text_nondetailed.txt	0.2302
mistral-7b-instruct-v0.2	wordfreq_rubric_numeric_detailed_nonweighted.txt	-0.2914
mistral-7b-instruct-v0.2	wordfreq_rubric_numeric_detailed_weighted.txt	0.3494
mistral-7b-instruct-v0.2	wordfreq_rubric_numeric_nondetailed_nonweighted.txt	-2.192
mistral-7b-instruct-v0.2	wordfreq_rubric_numeric_nondetailed_weighted.txt	-1.9411
mistral-7b-instruct-v0.2	wordfreq_rubric_text_detailed.txt	0.6394
mistral-7b-instruct-v0.2	wordfreq_rubric_text_nondetailed.txt	-0.2719
wizardlm2	wordfreq_rubric_numeric_detailed_nonweighted.txt	-0.6484
wizardlm2	wordfreq_rubric_numeric_detailed_weighted.txt	-0.1458
wizardlm2	wordfreq_rubric_numeric_nondetailed_nonweighted.txt	-1.3905
wizardlm2	wordfreq_rubric_numeric_nondetailed_weighted.txt	-0.4954
wizardlm2	wordfreq_rubric_text_detailed.txt	0.1111
wizardlm2	wordfreq_rubric_text_nondetailed.txt	0.4985

Table 13. R-squared scores for scripts-complexity5_wordfreq.

Model	Rubric	R-Squared Score
gemma	contactbook_rubric_numeric_detailed_nonweighted.txt	-2.891
gemma	contactbook_rubric_numeric_detailed_weighted.txt	-2.3371
gemma	contactbook_rubric_numeric_nondetailed_nonweighted.txt	-3.1067
gemma	contactbook_rubric_numeric_nondetailed_weighted.txt	-1.8838
gemma	contactbook_rubric_text_detailed.txt	-2.5824
gemma	contactbook_rubric_text_nondetailed.txt	-0.8648
gpt-3.5-turbo	contactbook_rubric_numeric_detailed_nonweighted.txt	-0.0533
gpt-3.5-turbo	contactbook_rubric_numeric_detailed_weighted.txt	0.7422
gpt-3.5-turbo	contactbook_rubric_numeric_nondetailed_nonweighted.txt	-0.7768
gpt-3.5-turbo	contactbook_rubric_numeric_nondetailed_weighted.txt	-0.5744
gpt-3.5-turbo	contactbook_rubric_text_detailed.txt	0.4321
gpt-3.5-turbo	contactbook_rubric_text_nondetailed.txt	0.5094
gpt-4	contactbook_rubric_numeric_detailed_nonweighted.txt	-0.4322
gpt-4	contactbook_rubric_numeric_detailed_weighted.txt	-0.058
gpt-4	contactbook_rubric_numeric_nondetailed_nonweighted.txt	-1.2573
gpt-4	contactbook_rubric_numeric_nondetailed_weighted.txt	-1.4188
gpt-4	contactbook_rubric_text_detailed.txt	0.0122
gpt-4	contactbook_rubric_text_nondetailed.txt	0.2484
gpt-4-turbo-preview	contactbook_rubric_numeric_detailed_nonweighted.txt	0.3967
gpt-4-turbo-preview	contactbook_rubric_numeric_detailed_weighted.txt	0.4653
gpt-4-turbo-preview	contactbook_rubric_numeric_nondetailed_nonweighted.txt	-0.4442
gpt-4-turbo-preview	contactbook_rubric_numeric_nondetailed_weighted.txt	-0.1765
gpt-4-turbo-preview	contactbook_rubric_text_detailed.txt	0.6032
gpt-4-turbo-preview	contactbook_rubric_text_nondetailed.txt	0.4788
llama2	contactbook_rubric_numeric_detailed_nonweighted.txt	-2.9808
llama2	contactbook_rubric_numeric_detailed_weighted.txt	0.0224
llama2	contactbook_rubric_numeric_nondetailed_nonweighted.txt	-3.4174
llama2	contactbook_rubric_numeric_nondetailed_weighted.txt	-1.613
llama2	contactbook_rubric_text_detailed.txt	-2.5928
llama2	contactbook_rubric_text_nondetailed.txt	-0.1408
llama3	contactbook_rubric_numeric_detailed_nonweighted.txt	0.1178
llama3	contactbook_rubric_numeric_detailed_weighted.txt	0.439
llama3	contactbook_rubric_numeric_nondetailed_nonweighted.txt	-0.4704
llama3	contactbook_rubric_numeric_nondetailed_weighted.txt	-0.447
llama3	contactbook_rubric_text_detailed.txt	0.7881
llama3	contactbook_rubric_text_nondetailed.txt	0.4224
mistral	contactbook_rubric_numeric_detailed_nonweighted.txt	-1.208
mistral	contactbook_rubric_numeric_detailed_weighted.txt	0.5863
mistral	contactbook_rubric_numeric_nondetailed_nonweighted.txt	-3.5054
mistral	contactbook_rubric_numeric_nondetailed_weighted.txt	-1.0158
mistral	contactbook_rubric_text_detailed.txt	0.528
mistral	contactbook_rubric_text_nondetailed.txt	0.3906
mistral-7b-instruct-v0.2	contactbook_rubric_numeric_detailed_nonweighted.txt	-1.4244
mistral-7b-instruct-v0.2	contactbook_rubric_numeric_detailed_weighted.txt	0.5897
mistral-7b-instruct-v0.2	contactbook_rubric_numeric_nondetailed_nonweighted.txt	-3.7277
mistral-7b-instruct-v0.2	contactbook_rubric_numeric_nondetailed_weighted.txt	-1.2545
mistral-7b-instruct-v0.2	contactbook_rubric_text_detailed.txt	0.31
mistral-7b-instruct-v0.2	contactbook_rubric_text_nondetailed.txt	-0.1059
wizardlm2	contactbook_rubric_numeric_detailed_nonweighted.txt	-1.018
wizardlm2	contactbook_rubric_numeric_detailed_weighted.txt	-0.0338
wizardlm2	contactbook_rubric_numeric_nondetailed_nonweighted.txt	-1.9556
wizardlm2	contactbook_rubric_numeric_nondetailed_weighted.txt	-0.0293
wizardlm2	contactbook_rubric_text_detailed.txt	-0.1662
wizardlm2	contactbook_rubric_text_nondetailed.txt	0.4045

Table 14. R-squared scores for scripts-complexity6_contactbook.

Model	Rubric	R-Squared Score
gemma	dronevsatellite_rubric_numeric_detailed_nonweighted.txt	-0.6676
gemma	dronevsatellite_rubric_numeric_detailed_weighted.txt	-0.0993
gemma	dronevsatellite_rubric_numeric_nondetailed_nonweighted.txt	0.6668
gemma	dronevsatellite_rubric_numeric_nondetailed_weighted.txt	0.4277
gemma	dronevsatellite_rubric_text_detailed.txt	0.3601
gemma	dronevsatellite_rubric_text_nondetailed.txt	0.3862
gpt-3.5-turbo	dronevsatellite_rubric_numeric_detailed_nonweighted.txt	0.5172
gpt-3.5-turbo	dronevsatellite_rubric_numeric_detailed_weighted.txt	-0.6609
gpt-3.5-turbo	dronevsatellite_rubric_numeric_nondetailed_nonweighted.txt	0.604
gpt-3.5-turbo	dronevsatellite_rubric_numeric_nondetailed_weighted.txt	0.7279
gpt-3.5-turbo	dronevsatellite_rubric_text_detailed.txt	-0.6477
gpt-3.5-turbo	dronevsatellite_rubric_text_nondetailed.txt	0.1051
gpt-4	dronevsatellite_rubric_numeric_detailed_nonweighted.txt	0.6006
gpt-4	dronevsatellite_rubric_numeric_detailed_weighted.txt	0.5135
gpt-4	dronevsatellite_rubric_numeric_nondetailed_nonweighted.txt	0.6846
gpt-4	dronevsatellite_rubric_numeric_nondetailed_weighted.txt	0.5877
gpt-4	dronevsatellite_rubric_text_detailed.txt	0.5543
gpt-4	dronevsatellite_rubric_text_nondetailed.txt	0.2602
gpt-4-turbo-preview	dronevsatellite_rubric_numeric_detailed_nonweighted.txt	0.7565
gpt-4-turbo-preview	dronevsatellite_rubric_numeric_detailed_weighted.txt	0.4605
gpt-4-turbo-preview	dronevsatellite_rubric_numeric_nondetailed_nonweighted.txt	0.6829
gpt-4-turbo-preview	dronevsatellite_rubric_numeric_nondetailed_weighted.txt	0.5992
gpt-4-turbo-preview	dronevsatellite_rubric_text_detailed.txt	-0.2828
gpt-4-turbo-preview	dronevsatellite_rubric_text_nondetailed.txt	0.1383
llama2	dronevsatellite_rubric_numeric_detailed_nonweighted.txt	-2.014
llama2	dronevsatellite_rubric_numeric_detailed_weighted.txt	0.0157
llama2	dronevsatellite_rubric_numeric_nondetailed_nonweighted.txt	0.349
llama2	dronevsatellite_rubric_numeric_nondetailed_weighted.txt	-0.2443
llama2	dronevsatellite_rubric_text_detailed.txt	-5.4198
llama2	dronevsatellite_rubric_text_nondetailed.txt	0.1933
llama3	dronevsatellite_rubric_numeric_detailed_nonweighted.txt	0.7483
llama3	dronevsatellite_rubric_numeric_detailed_weighted.txt	0.6954
llama3	dronevsatellite_rubric_numeric_nondetailed_nonweighted.txt	0.4634
llama3	dronevsatellite_rubric_numeric_nondetailed_weighted.txt	0.5029
llama3	dronevsatellite_rubric_text_detailed.txt	0.2204
llama3	dronevsatellite_rubric_text_nondetailed.txt	0.3359
mistral	dronevsatellite_rubric_numeric_detailed_nonweighted.txt	0.112
mistral	dronevsatellite_rubric_numeric_detailed_weighted.txt	0.692
mistral	dronevsatellite_rubric_numeric_nondetailed_nonweighted.txt	-0.2488
mistral	dronevsatellite_rubric_numeric_nondetailed_weighted.txt	-0.2459
mistral	dronevsatellite_rubric_text_detailed.txt	0.3299
mistral	dronevsatellite_rubric_text_nondetailed.txt	0.3153
mistral-7b-instruct-v0.2	dronevsatellite_rubric_numeric_detailed_nonweighted.txt	0.1928
mistral-7b-instruct-v0.2	dronevsatellite_rubric_numeric_detailed_weighted.txt	0.7732
mistral-7b-instruct-v0.2	dronevsatellite_rubric_numeric_nondetailed_nonweighted.txt	-0.1462
mistral-7b-instruct-v0.2	dronevsatellite_rubric_numeric_nondetailed_weighted.txt	0.4693
mistral-7b-instruct-v0.2	dronevsatellite_rubric_text_detailed.txt	0.5366
mistral-7b-instruct-v0.2	dronevsatellite_rubric_text_nondetailed.txt	0.2473
wizardlm2	dronevsatellite_rubric_numeric_detailed_nonweighted.txt	-0.2357
wizardlm2	dronevsatellite_rubric_numeric_detailed_weighted.txt	0.7151
wizardlm2	dronevsatellite_rubric_numeric_nondetailed_nonweighted.txt	0.0529
wizardlm2	dronevsatellite_rubric_numeric_nondetailed_weighted.txt	-0.8983
wizardlm2	dronevsatellite_rubric_text_detailed.txt	0.6193
wizardlm2	dronevsatellite_rubric_text_nondetailed.txt	0.6178

Table 15. R-squared scores for short_answers-complexity1_dronevsatellite.

Model	Rubric	R-Squared Score
gemma	satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt	0.3169
gemma	satellitecomvnoncom_rubric_numeric_detailed_weighted.txt	0.4727
gemma	satellitecomvnoncom_rubric_numeric_nondetailed_nonweighted.txt	0.7676
gemma	satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt	0.5949
gemma	satellitecomvnoncom_rubric_text_detailed.txt	0.291
gemma	satellitecomvnoncom_rubric_text_nondetailed.txt	0.7695
gpt-3.5-turbo	satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt	0.7049
gpt-3.5-turbo	satellitecomvnoncom_rubric_numeric_detailed_weighted.txt	-0.0098
gpt-3.5-turbo	satellitecomvnoncom_rubric_numeric_nondetailed_nonweighted.txt	0.8158
gpt-3.5-turbo	satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt	0.8161
gpt-3.5-turbo	satellitecomvnoncom_rubric_text_detailed.txt	-0.7406
gpt-3.5-turbo	satellitecomvnoncom_rubric_text_nondetailed.txt	0.0505
gpt-4	satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt	0.8312
gpt-4	satellitecomvnoncom_rubric_numeric_detailed_weighted.txt	0.6869
gpt-4	satellitecomvnoncom_rubric_numeric_nondetailed_nonweighted.txt	0.8851
gpt-4	satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt	0.7849
gpt-4	satellitecomvnoncom_rubric_text_detailed.txt	0.7416
gpt-4	satellitecomvnoncom_rubric_text_nondetailed.txt	0.2921
gpt-4-turbo-preview	satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt	0.7753
gpt-4-turbo-preview	satellitecomvnoncom_rubric_numeric_detailed_weighted.txt	0.5586
gpt-4-turbo-preview	satellitecomvnoncom_rubric_numeric_nondetailed_nonweighted.txt	0.838
gpt-4-turbo-preview	satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt	0.832
gpt-4-turbo-preview	satellitecomvnoncom_rubric_text_detailed.txt	-0.0055
gpt-4-turbo-preview	satellitecomvnoncom_rubric_text_nondetailed.txt	0.2968
llama2	satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt	-1.2361
llama2	satellitecomvnoncom_rubric_numeric_detailed_weighted.txt	-0.1131
llama2	satellitecomvnoncom_rubric_numeric_nondetailed_nonweighted.txt	0.3434
llama2	satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt	-0.1375
llama2	satellitecomvnoncom_rubric_text_detailed.txt	-4.041
llama2	satellitecomvnoncom_rubric_text_nondetailed.txt	0.2979
llama3	satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt	0.9031
llama3	satellitecomvnoncom_rubric_numeric_detailed_weighted.txt	0.7286
llama3	satellitecomvnoncom_rubric_numeric_nondetailed_nonweighted.txt	0.6739
llama3	satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt	0.588
llama3	satellitecomvnoncom_rubric_text_detailed.txt	0.459
llama3	satellitecomvnoncom_rubric_text_nondetailed.txt	0.5492
mistral	satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt	0.3246
mistral	satellitecomvnoncom_rubric_numeric_detailed_weighted.txt	0.7647
mistral	satellitecomvnoncom_rubric_numeric_nondetailed_nonweighted.txt	0.3243
mistral	satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt	0.5662
mistral	satellitecomvnoncom_rubric_text_detailed.txt	0.6084
mistral	satellitecomvnoncom_rubric_text_nondetailed.txt	0.5174
mistral-7b-instruct-v0.2	satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt	0.5173
mistral-7b-instruct-v0.2	satellitecomvnoncom_rubric_numeric_detailed_weighted.txt	0.7896
mistral-7b-instruct-v0.2	satellitecomvnoncom_rubric_numeric_nondetailed_nonweighted.txt	0.3334
mistral-7b-instruct-v0.2	satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt	0.6916
mistral-7b-instruct-v0.2	satellitecomvnoncom_rubric_text_detailed.txt	0.6706
mistral-7b-instruct-v0.2	satellitecomvnoncom_rubric_text_nondetailed.txt	0.4765
wizardlm2	satellitecomvnoncom_rubric_numeric_detailed_nonweighted.txt	0.3787
wizardlm2	satellitecomvnoncom_rubric_numeric_detailed_weighted.txt	0.7338
wizardlm2	satellitecomvnoncom_rubric_numeric_nondetailed_nonweighted.txt	0.4413
wizardlm2	satellitecomvnoncom_rubric_numeric_nondetailed_weighted.txt	-0.4891
wizardlm2	satellitecomvnoncom_rubric_text_detailed.txt	0.7262
wizardlm2	satellitecomvnoncom_rubric_text_nondetailed.txt	0.6428

Table 16. R-squared scores for short_answers-complexity2_satellitecomvnoncom.

Model	Rubric	R-Squared Score
gemma	robotsvboom_rubric_numeric_detailed_nonweighted.txt	0.6398
gemma	robotsvboom_rubric_numeric_detailed_weighted.txt	0.5923
gemma	robotsvboom_rubric_numeric_nondetailed_nonweighted.txt	0.8452
gemma	robotsvboom_rubric_numeric_nondetailed_weighted.txt	0.7651
gemma	robotsvboom_rubric_text_detailed.txt	0.6326
gemma	robotsvboom_rubric_text_nondetailed.txt	0.8148
gpt-3.5-turbo	robotsvboom_rubric_numeric_detailed_nonweighted.txt	0.7345
gpt-3.5-turbo	robotsvboom_rubric_numeric_detailed_weighted.txt	0.173
gpt-3.5-turbo	robotsvboom_rubric_numeric_nondetailed_nonweighted.txt	0.8998
gpt-3.5-turbo	robotsvboom_rubric_numeric_nondetailed_weighted.txt	0.8751
gpt-3.5-turbo	robotsvboom_rubric_text_detailed.txt	-0.0013
gpt-3.5-turbo	robotsvboom_rubric_text_nondetailed.txt	0.6008
gpt-4	robotsvboom_rubric_numeric_detailed_nonweighted.txt	0.9419
gpt-4	robotsvboom_rubric_numeric_detailed_weighted.txt	0.8734
gpt-4	robotsvboom_rubric_numeric_nondetailed_nonweighted.txt	0.8693
gpt-4	robotsvboom_rubric_numeric_nondetailed_weighted.txt	0.9081
gpt-4	robotsvboom_rubric_text_detailed.txt	0.7149
gpt-4	robotsvboom_rubric_text_nondetailed.txt	0.427
gpt-4-turbo-preview	robotsvboom_rubric_numeric_detailed_nonweighted.txt	0.8819
gpt-4-turbo-preview	robotsvboom_rubric_numeric_detailed_weighted.txt	0.6044
gpt-4-turbo-preview	robotsvboom_rubric_numeric_nondetailed_nonweighted.txt	0.9282
gpt-4-turbo-preview	robotsvboom_rubric_numeric_nondetailed_weighted.txt	0.9041
gpt-4-turbo-preview	robotsvboom_rubric_text_detailed.txt	0.5068
gpt-4-turbo-preview	robotsvboom_rubric_text_nondetailed.txt	0.3377
llama2	robotsvboom_rubric_numeric_detailed_nonweighted.txt	-0.2671
llama2	robotsvboom_rubric_numeric_detailed_weighted.txt	0.1522
llama2	robotsvboom_rubric_numeric_nondetailed_nonweighted.txt	0.4527
llama2	robotsvboom_rubric_numeric_nondetailed_weighted.txt	-0.0329
llama2	robotsvboom_rubric_text_detailed.txt	-3.5528
llama2	robotsvboom_rubric_text_nondetailed.txt	0.5444
llama3	robotsvboom_rubric_numeric_detailed_nonweighted.txt	0.9323
llama3	robotsvboom_rubric_numeric_detailed_weighted.txt	0.834
llama3	robotsvboom_rubric_numeric_nondetailed_nonweighted.txt	0.8435
llama3	robotsvboom_rubric_numeric_nondetailed_weighted.txt	0.6514
llama3	robotsvboom_rubric_text_detailed.txt	0.4922
llama3	robotsvboom_rubric_text_nondetailed.txt	0.2301
mistral	robotsvboom_rubric_numeric_detailed_nonweighted.txt	0.8371
mistral	robotsvboom_rubric_numeric_detailed_weighted.txt	0.8611
mistral	robotsvboom_rubric_numeric_nondetailed_nonweighted.txt	0.6713
mistral	robotsvboom_rubric_numeric_nondetailed_weighted.txt	0.6418
mistral	robotsvboom_rubric_text_detailed.txt	0.4369
mistral	robotsvboom_rubric_text_nondetailed.txt	0.3586
mistral-7b-instruct-v0.2	robotsvboom_rubric_numeric_detailed_nonweighted.txt	0.8349
mistral-7b-instruct-v0.2	robotsvboom_rubric_numeric_detailed_weighted.txt	0.8703
mistral-7b-instruct-v0.2	robotsvboom_rubric_numeric_nondetailed_nonweighted.txt	0.7587
mistral-7b-instruct-v0.2	robotsvboom_rubric_numeric_nondetailed_weighted.txt	0.8583
mistral-7b-instruct-v0.2	robotsvboom_rubric_text_detailed.txt	0.5407
mistral-7b-instruct-v0.2	robotsvboom_rubric_text_nondetailed.txt	0.3859
wizardlm2	robotsvboom_rubric_numeric_detailed_nonweighted.txt	0.5407
wizardlm2	robotsvboom_rubric_numeric_detailed_weighted.txt	0.726
wizardlm2	robotsvboom_rubric_numeric_nondetailed_nonweighted.txt	0.7191
wizardlm2	robotsvboom_rubric_numeric_nondetailed_weighted.txt	0.1377
wizardlm2	robotsvboom_rubric_text_detailed.txt	0.672
wizardlm2	robotsvboom_rubric_text_nondetailed.txt	0.6818

Table 17. R-squared scores for short_answers-complexity3_robotsvboom.

Model	Rubric	R-Squared Score
gemma	sensorhood_rubric_numeric_detailed_nonweighted.txt	0.3624
gemma	sensorhood_rubric_numeric_detailed_weighted.txt	0.4195
gemma	sensorhood_rubric_numeric_nondetailed_nonweighted.txt	0.6479
gemma	sensorhood_rubric_numeric_nondetailed_weighted.txt	0.5618
gemma	sensorhood_rubric_text_detailed.txt	0.4059
gemma	sensorhood_rubric_text_nondetailed.txt	0.6546
gpt-3.5-turbo	sensorhood_rubric_numeric_detailed_nonweighted.txt	0.7464
gpt-3.5-turbo	sensorhood_rubric_numeric_detailed_weighted.txt	0.1551
gpt-3.5-turbo	sensorhood_rubric_numeric_nondetailed_nonweighted.txt	0.5507
gpt-3.5-turbo	sensorhood_rubric_numeric_nondetailed_weighted.txt	0.6987
gpt-3.5-turbo	sensorhood_rubric_text_detailed.txt	-0.3065
gpt-3.5-turbo	sensorhood_rubric_text_nondetailed.txt	-0.0795
gpt-4	sensorhood_rubric_numeric_detailed_nonweighted.txt	0.6255
gpt-4	sensorhood_rubric_numeric_detailed_weighted.txt	0.6383
gpt-4	sensorhood_rubric_numeric_nondetailed_nonweighted.txt	0.7343
gpt-4	sensorhood_rubric_numeric_nondetailed_weighted.txt	0.739
gpt-4	sensorhood_rubric_text_detailed.txt	0.7176
gpt-4	sensorhood_rubric_text_nondetailed.txt	0.1634
gpt-4-turbo-preview	sensorhood_rubric_numeric_detailed_nonweighted.txt	0.7876
gpt-4-turbo-preview	sensorhood_rubric_numeric_detailed_weighted.txt	0.4838
gpt-4-turbo-preview	sensorhood_rubric_numeric_nondetailed_nonweighted.txt	0.8206
gpt-4-turbo-preview	sensorhood_rubric_numeric_nondetailed_weighted.txt	0.7026
gpt-4-turbo-preview	sensorhood_rubric_text_detailed.txt	0.4614
gpt-4-turbo-preview	sensorhood_rubric_text_nondetailed.txt	0.222
llama2	sensorhood_rubric_numeric_detailed_nonweighted.txt	-0.9138
llama2	sensorhood_rubric_numeric_detailed_weighted.txt	-0.0497
llama2	sensorhood_rubric_numeric_nondetailed_nonweighted.txt	0.3548
llama2	sensorhood_rubric_numeric_nondetailed_weighted.txt	-0.5567
llama2	sensorhood_rubric_text_detailed.txt	-1.6257
llama2	sensorhood_rubric_text_nondetailed.txt	0.4239
llama3	sensorhood_rubric_numeric_detailed_nonweighted.txt	0.8147
llama3	sensorhood_rubric_numeric_detailed_weighted.txt	0.7087
llama3	sensorhood_rubric_numeric_nondetailed_nonweighted.txt	0.534
llama3	sensorhood_rubric_numeric_nondetailed_weighted.txt	0.6044
llama3	sensorhood_rubric_text_detailed.txt	0.5978
llama3	sensorhood_rubric_text_nondetailed.txt	-0.0051
mistral	sensorhood_rubric_numeric_detailed_nonweighted.txt	0.4913
mistral	sensorhood_rubric_numeric_detailed_weighted.txt	0.7483
mistral	sensorhood_rubric_numeric_nondetailed_nonweighted.txt	0.4028
mistral	sensorhood_rubric_numeric_nondetailed_weighted.txt	0.5325
mistral	sensorhood_rubric_text_detailed.txt	0.5494
mistral	sensorhood_rubric_text_nondetailed.txt	0.4982
mistral-7b-instruct-v0.2	sensorhood_rubric_numeric_detailed_nonweighted.txt	0.4766
mistral-7b-instruct-v0.2	sensorhood_rubric_numeric_detailed_weighted.txt	0.7801
mistral-7b-instruct-v0.2	sensorhood_rubric_numeric_nondetailed_nonweighted.txt	0.3546
mistral-7b-instruct-v0.2	sensorhood_rubric_numeric_nondetailed_weighted.txt	0.6166
mistral-7b-instruct-v0.2	sensorhood_rubric_text_detailed.txt	0.5355
mistral-7b-instruct-v0.2	sensorhood_rubric_text_nondetailed.txt	0.6215
wizardlm2	sensorhood_rubric_numeric_detailed_nonweighted.txt	0.4779
wizardlm2	sensorhood_rubric_numeric_detailed_weighted.txt	0.6147
wizardlm2	sensorhood_rubric_numeric_nondetailed_nonweighted.txt	0.2475
wizardlm2	sensorhood_rubric_numeric_nondetailed_weighted.txt	-0.0819
wizardlm2	sensorhood_rubric_text_detailed.txt	0.427
wizardlm2	sensorhood_rubric_text_nondetailed.txt	0.5381

Table 18. R-squared scores for short_answers-complexity4_sensorhood