Evaluating Large Language Models for Statistical Test Selection and Explanation

Anh Nguyen
Advisor: Heather Pon-Barry

Abstract:

Research study designs are complex, requiring knowledge about different types of variables, models, and assumptions. As a result, it can be difficult to identify the appropriate statistical test. Large language models (LLMs) can help by providing tailored suggestions and explanations. I aim to create an LLM-based assistant that recommends the appropriate statistical test for a study scenario and explains the reasoning behind the choice. However, especially for domain-specific tasks, because model explanations can drift from facts, retrieval can help anchor them to reliable sources. I compare three versions of the system: two small, locally-hosted Llama models, a state-of-the-art Gemini model, and a retrieval-augmented generation (RAG) variant that uses a statistics textbook as an external information source. Using a diverse set of example scenarios and prompts, I evaluate accuracy of test selection and explanations for each model. Gemini is the best performing model, followed by Llama 8B, and finally Llama 3B (the smallest model). Introducing retrieval substantially improves performance. Llama 3B + RAG can sometimes match or exceed larger non-RAG models. These findings show RAG's potential to strengthen classification and explanation for models that require less computation resources. Furthermore, this study highlights LLMs' potential to support effective decision-making in education and research settings.

Keywords: LLM, RAG, prompt engineering, classification and generated text evaluation, statistics education

I. Introduction

By integrating broad contextual knowledge with flexible, generative explanations, large language models (LLMs) have substantially advanced automated reasoning and question answering. Their strong performance across diverse Q&A benchmarks has motivated interest in applying them to educational and scientific tasks (Brown et al., 2020). One such task is selecting appropriate statistical tests, a process that requires knowledge about variable types, study design, and mathematical assumptions. Although structured decision guides exist, many learners still struggle to identify which test fits a given research scenario (Najmi et al., 2021). LLMs offer potential support by providing both recommendations and accompanying explanations.

However, their performance depends on factors such as model size, prompting methods, and whether they can access external reference materials. Several techniques have emerged as methods to improve reliability, such as prompt engineering and Retrieval-Augmented Generation (RAG), which supplements models with valuable domain knowledge through passages retrieved from external sources. To better understand the benefits of these strategies, this study evaluates several LLM configurations: two small, local Llama models, a state-of-the-art Gemini model, and a RAG system that utilizes an external statistics textbook. The goal is to examine how model size and retrieval influence accuracy and explanation quality in the context of statistical test selection.

II. Related Works

LLMs have shown strong performance on both general and domain-specific established Q&A benchmarks such as Natural Questions (Kwiatkowski et al., 2019) and PubMedQA (Jin et al., 2019). As advanced models such as GPT and Llama have expanded in scale and training complexity, Q&A systems have shifted from multi-component architectures toward end-to-end models capable of both retrieval and generation (Guu et al., 2020).

Retrieval-augmented generation (RAG) has become a central strategy for improving factual reliability in LLMs, especially for domain-specific Q&A tasks. This technique has demonstrated that incorporating retrieved context from an external reliable source can increase accuracy and reduce hallucination (Lewis et al., 2020). Applications in education, and specifically Mathematics education, have also grown. Levonian et al. (2023) showed that textbook retrieval can improve groundedness in math explanations.

There are also works that have explored LLMs' statistical reasoning ability specifically. StatQA (Zhu et al., 2024) and StatLLM (Song et al., 2025) offer large-scale datasets and benchmarks for evaluating LLMs on applied statistics tasks, including short answer questions and programming. These works highlight the potential of LLMs to support statistics workflows but also reveal current limitations regarding consistency and accuracy.

A closely related pilot study by Mondal et al. (2024) evaluated GPT-4's statistical test recommendations across 27 research scenarios. Although accuracy was reasonably high, explanations were often incomplete or inconsistent, revealing a gap between correct classification and correct reasoning. This project extends prior work by comparing multiple model sizes and architectures and examining whether adding retrieval can improve the quality of model explanations.

Recent research has also advanced frameworks for evaluating LLM reasoning quality in addition to classification accuracy. Approaches range from rubric-based human evaluation to automated evaluation metrics such as BLEU, ROUGE, or BERTScore (Zhang et al., 2019). For retrieval-augmented systems, measures like RAGAS can assess how faithfully a model's answer is grounded in the retrieved evidence (Es et al., 2024). In high-stakes settings such as medicine, expert evaluation remains indispensable (Tam et al., 2024). Model reliability is commonly evaluated by how closely the model's responses align with human experts' ratings.

III. Dataset

Due to the lack of existing datasets that are well suited for this project, I constructed a dataset specifically designed to evaluate statistical test selection from short descriptions of research scenarios. The dataset contains 77 different study descriptions and the appropriate test for each. Although the dataset is modest in size, a lot of thought was put into balance and diversity to best mitigate limitations related to scale. The scenarios consist of 7 commonly used statistical tests (one-sample t-test, paired t-test, two-sample t-test, chi-square test, anova, multiple linear regression, logistic regression). Scenarios were drawn from a wide range of domains, including psychology, biology, and education. The distribution of the tests is adequately balanced, as shown in Figure 1.

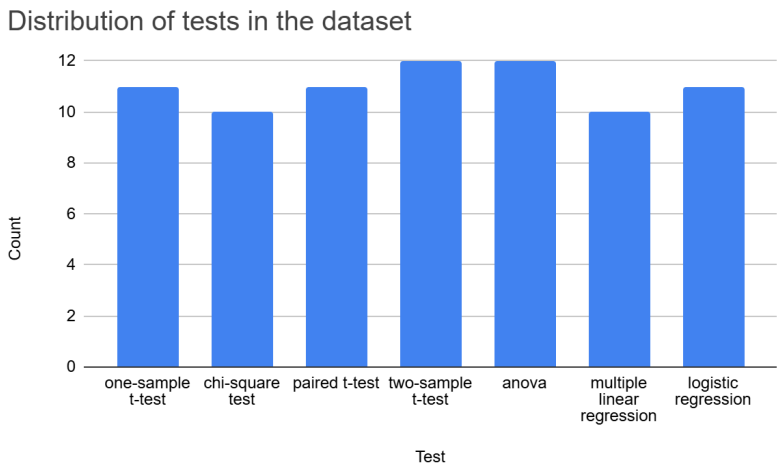

Figure 1: Bar graph showing distribution of tests in the dataset. The distribution is balanced, with all tests appearing 10-12 times in the dataset.

Table 1 shows some examples of rows in the dataset.

Scenario	Test
New York is known as “the city that never sleeps”. A random sample of 35 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. The point estimate suggests New Yorkers sleep less than 8 hours a night on average. Is the result statistically significant?	one-sample t-test
Researchers interested in the relationship between absenteeism from school and certain demographic characteristics of children (ethnicity, sex, etc.) collected data from 146 randomly sampled students in rural New South Wales, Australia, in a particular school year. They want to predict the average number of days absent based on ethnic background, sex, and learner status.	multiple linear regression
The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents. We can consider educational attainment levels for all 1,172 respondents at once. The distributions of hours worked by educational attainment and relevant summary statistics are given. What test can be used for evaluating whether the average number of hours worked varies across the five groups?	anova

Table 1: Examples of dataset entries. Each row has 2 columns. One contains a short description of a research scenario that includes context, research question, and information about response and explanatory variables. The other has the correct test label.

The primary source of scenarios and correct answers is the “OpenIntro Statistics” textbook (Diez et al., 2019), supplemented by online materials from the Oxford College of Emory University’s Department of Mathematics and Computer Science. All scenarios were compiled, paraphrased, and adapted to ensure clarity and suitability for this specific task.

IV. Methods

The full dataset, workflow, and code can be found in the following GitHub repository: https://github.com/PAnhHNguyen/stats-llm-assistant-independent-study

1. Models

This study evaluates three large language models that vary in size, computational requirements, and expected reasoning capabilities. The first two models, Llama 3.2 3B Instruct (Llama 3B) and Llama 3 8B Instruct (Llama 8B), are small, locally-hosted models. The names reflect their parameter counts, approximately 3 billion and 8 billion parameters respectively. Both of these models can be run on the average laptop. I used LMStudio as the hosting application, and the models take up 1.8 and 4.37 GB respectively. Smaller models such as these are attractive for local or resource-constrained applications but generally provide less robust reasoning compared to larger models.

The third model, Gemini Flash 2.5, represents a state-of-the-art, high-capacity system with substantially more parameters and training data than either Llama variant. Although specific parameter counts for proprietary models are not publicly disclosed, Gemini operates at a scale far larger than single-digit-billion-parameter models and is engineered for complex reasoning tasks. Gemini can provide notably higher performance but requires cloud-based access and greater computational resources.

Together, these models allow for a comparison between small, local models and a high-capacity, state-of-the-art model. This provides insight into how model scale affects statistical test selection and explanation quality.

2. Prompt engineering

Prompt design plays an important role in ensuring consistent evaluation across models. Because the task requires both classification and explanation, all prompts ask models to identify key features of each study scenario, justify the selected statistical test, and list all assumptions required for that test. Two prompt variants are used for every model to assess sensitivity to prompt formulation.

Both prompt variants utilize chain-of-thought (CoT) prompting, an approach in which the model is instructed to explicitly write out its reasoning steps before providing a final answer. CoT supports the model in working through problems step by step, and is especially helpful for smaller models. The two prompt variants differ only in whether models receive an explicit multiple-choice constraint. The first prompt (MCQ prompt) includes a list of the 7 common statistical tests to choose from, while the second provides no predefined list of tests. The prompt template is shown below. The underlined line is only included in the MCQ prompt.
Prompt:
" You are a helpful assistant that suggests the appropriate statistical test based on the provided research scenario.
TASK: Given the following research scenario, suggest the most appropriate statistical test.
<u>The test must be one of the following: one-sample t-test, two-sample t-test, ANOVA, Chi-square test, paired t-test, multiple linear regression, logistic regression.</u>

SCENARIO: {scenario}
NOTE: Your response should include consideration of assumptions and conditions for each test. Show your step-by-step reasoning and then on its own line, output exactly:
'Final Answer: {test name}'. "

The unconstrained non-MCQ prompt more closely reflects real-world usage, where models must independently identify suitable tests. However, using a multiple-choice format can stabilize performance by limiting the range of acceptable answers and effectively converting the task into a more conventional classification task. Comparing these two formats can show how constraints interact with CoT prompting to influence accuracy and explanation quality across models of different sizes.

3. RAG

Retrieval-augmented generation (RAG) is incorporated to examine whether grounding a smaller model, in this case Llama 3B, in external statistical knowledge can improve its ability to select appropriate tests and justify its reasoning. A RAG workflow has two components: a retrieval step that identifies relevant information from a knowledge source, and a generation step in which the model incorporates that information into its response.

Specifically, the PDF of the OpenIntro Statistics textbook is first segmented into smaller chunks. Then, each text chunk is converted into a numerical representation using a sentence-transformer model, which is a type of embedding model that turns sentences into vectors so that similar meanings are placed close together. This allows the system to perform semantic similarity search, meaning it can retrieve passages that match the scenario's meaning rather than just its wording. In parallel, a BM25 index was created to support keyword-based retrieval, which ranks passages based on overlapping words and their importance. These two retrieval methods, semantic (dense) and keyword-based (sparse), are then combined using reciprocal rank fusion, a method that merges the rankings from both systems to produce a more reliable final list. The top two retrieved passages are used as context for the model.

These retrieved excerpts are then inserted into the prompt as context. The model is instructed to use the retrieved material only if it is relevant. The remainder of the prompt is exactly the same as non-RAG versions. The result enables evaluation of whether retrieval improves accuracy and reduces errors in comparison to non-RAG workflows.

4. Evaluation

Model performance is evaluated in two main ways: classification accuracy (whether the model selects the correct statistical test) and how strong the accompanying justification is. For the classification component, predictions were compared with the correct test labels for all 77 scenarios. Standard classification evaluation metrics are computed, including overall accuracy, precision, recall, and F1-score. Confusion matrices are also generated to identify systematic patterns of misclassification such as confusion between paired and independent t-tests.

Justification quality is evaluated using both manual and automated approaches. A subset of ten scenarios is randomly sampled to allow closer inspection of the reasoning produced by each model. Manual evaluation follows a structured rubric that assesses research question and variables identification, correctness and completeness of assumptions, and overall reasoning and clarity. The rubric used is shown in Table 2.

	Description	0	1	2
Identifying problem & data	Does the justification state the research question, dependent/independent variables, and whether samples are independent or dependent?	No clear identification of variables or research questions.	Mentions variables or questions, but vague or incomplete.	Clearly identifies dependent & independent variables, specifies categorical vs. numeric, and clarifies whether samples are independent or dependent.
Checking assumptions & conditions	Does the justification note assumptions or conditions of the test (e.g., normality, equal variance, independence)?	None mentioned.	Does not mention all necessary assumptions or is too vague.	States key assumptions/conditions correctly and clearly.
Reasoning & Clarity	Is the explanation logical and easy to follow, connecting the problem to the test choice?	No reasoning.	Some reasoning but unclear or incomplete.	Step-by-step or logically structured reasoning, easy to understand.

Table 2: The rubric used for manual justification evaluation assigns scores from 0 to 6, composed of 0-2 possible points for 3 different categories. The criterias are understanding of data given, understanding of the test suggested, and overall clarity of language.

Automated evaluation is also conducted on the same sample. For each scenario, a “golden” reference explanation is created to represent an ideal justification. Model explanations are then compared to these gold answers using two text-similarity metrics. Cosine similarity is calculated using the sentence-transformer embeddings, which represent each explanation as a vector. This metric measures how close the meanings of two explanations are by comparing the angle between their vectors. In addition, BERTScore is also used, which compares explanations based on contextual token embeddings from a pretrained language model. Unlike simple word-overlap methods, BERTScore evaluates how similar the words are in meaning within their sentence context, giving a more detailed measure of semantic alignment. Both metrics range from 0 to 1,

with values closer to 1 indicating stronger similarity. Together, these automated metrics offer a more scalable way to assess explanation quality.

V. Results

1. Classification results

Classification performance was assessed across all 77 scenarios for each model and prompt variant. Table 3 summarizes overall accuracy.

Model	Prompt	Classification accuracy
Llama 3B	MCQ	0.712
Llama 3B	not MCQ	0.739
Llama 8B	MCQ	0.804
Llama 8B	not MCQ	0.413
Gemini Flash 2.5	MCQ	0.913
Gemini Flash 2.5	not MCQ	0.845
Llama 3B + RAG	MCQ	0.72
Llama 3B + RAG	not MCQ	0.54

Table 3: The classification accuracy for each model under both prompts is recorded. Gemini Flash 2.5 achieved the highest accuracy, reaching 0.913 with the MCQ prompt and 0.845 with the non-MCQ version. Llama 8B showed moderate performance with the MCQ prompt (0.804) but a substantial drop with the non-MCQ version (0.413). Llama 3B achieved accuracies of 0.712 (MCQ) and 0.739 (non-MCQ). The RAG-augmented Llama 3B achieved an accuracy of 0.72 under MCQ but only 0.54 under non-MCQ.

These results show strong benefits of model scale and modest gains from retrieval. Overall, the classification results highlight a clear hierarchy of model ability: Gemini > Llama 8B > Llama 3B, with RAG providing selective improvements for the smallest model. This order is somewhat expected. However, in order to understand Llama 8B's performance drop with the non-MCQ prompt and the specific conditions under which RAG can provide value, more in-depth evaluation of false classifications and inaccurate justifications is required.

To complement the summary statistics, detailed confusion matrices are also created for each model. These visualizations make it possible to identify specific patterns of misclassification. For each confusion matrix, the diagonal shows correct classifications and the off-diagonal squares show misclassifications. The true labels are on the vertical axis and the model's predictions are on the horizontal axis. All 8 confusion matrices for all models under both prompts can be found in the Github repository. Figures 2-4 compare and contrast 3 of the most representative matrices.

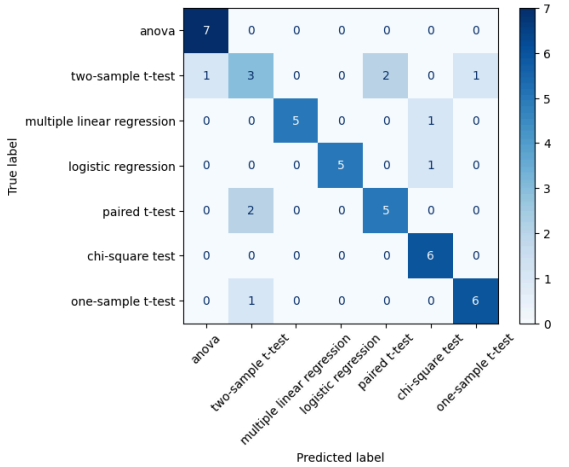

Figure 2: Gemini's confusion matrix (MCQ prompt)
Most off-diagonal squares are 0, showing strong performance. The only errors are misclassifying two-sample t-tests and missclassifying logistic regression as chi-square test.

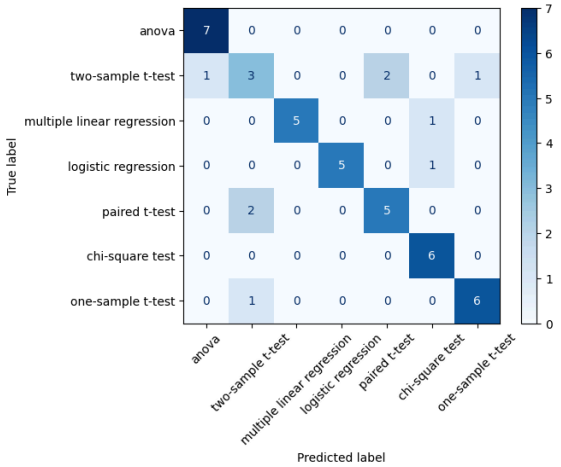

Figure 3: Llama 8B's confusion matrix (MCQ prompt)
The diagonal is weaker than Gemini's, showing worse performance, but still ~80% classification accuracy. Errors are relatively spread out, but most are confusion between two-sample t-tests and paired t-tests.

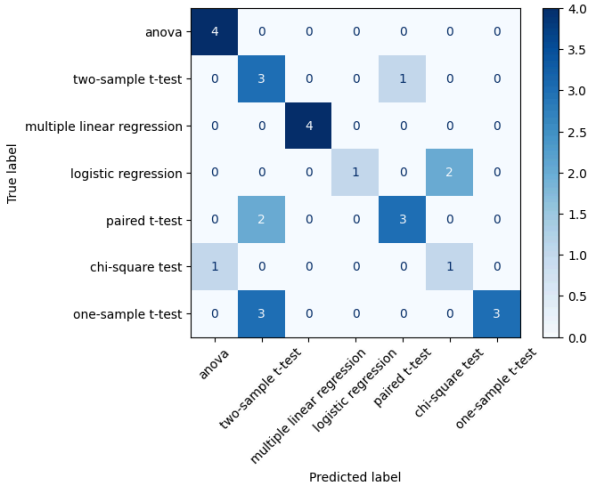

Figure 4: Llama 8B's confusion matrix (non-MCQ prompt)
This is the worst-performing model-prompt combination. Widespread misclassifications can be seen.

Both Llama 3B and Llama 8B under both prompts most often confuse between paired t-tests and two-sample t-tests and between logistic regression and chi-square test. Gemini makes the fewest number of classification errors but if it does the mistake is usually confusing between logistic regression and chi-square test. Llama 8B shows decent performance under MCQ prompting but exhibits widespread confusion with the non-MCQ prompt. Gemini's matrices show the highest consistency, with most classifications aligning to the diagonal. With non-MCQ prompts, there is also a higher likelihood of models suggesting much more complex, incorrect tests that are outside of the 7 true labels. This is the biggest factor in Llama 8B's drop in accuracy. The model sometimes suggests tests such as the Wilcoxon rank-sum test or the Mann-Whitney u test,

which are more advanced statistical tests not required of these introductory statistics scenarios.

2. Justification evaluation results

A sample of 10 randomly selected, correctly classified justifications make up the data for justification evaluation. The sample includes at least 1 of each of the 7 tests. All evaluation results (manual and automated) are shown in tables 4-6.

Manual scores						
Scenario	3B MCQ score	3B No MCQ score	8B MCQ score	8B No MCQ score	Gemini No MCQ score	3B + RAG score
Undergraduate students taking an introductory	5	5	5	5	6	5
A study of reaction times was conducted with 1	4	6	6	6	6	5
A survey of 55 Duke University students asked	5	4	5	5	6	6
Subjects from Central Prison in Raleigh, NC, v	5	6	6	6	6	5
Microhabitat factors associated with forage and	5	5	6	6	6	5
Raina and Luke survey a random sample of 40	6	5	5	5	6	5
Real estate researchers aim to understand hov	5	6	5	6	6	5
A professor who teaches a large introductory s	4	6	5	6	6	6
Participation in physical activity often declines	5	5	4	6	6	4
A university wellness survey asked students w	4	4	5	4	6	5

Table 4: Manual evaluation scores for all models. Gemini receives consistently higher scores, Llama 8B receives mixed results, and Llama 3B has the lowest scores on average.

Automated scores (Cosine similarity)						
Scenario	3B MCQ score	3B No MCQ score	8B MCQ score	8B No MCQ score	Gemini No MCQ score	3B + RAG score
Undergraduate students taking an introductory	0.91	0.93	0.91	0.92	0.93	0.92
A study of reaction times was conducted with 1	0.93	0.94	0.9	0.92	0.93	0.94
A survey of 55 Duke University students asked	0.91	0.89	0.92	0.94	0.93	0.93
Subjects from Central Prison in Raleigh, NC, v	0.92	0.91	0.91	0.91	0.94	0.91
Microhabitat factors associated with forage and	0.94	0.95	0.95	0.92	0.95	0.93
Raina and Luke survey a random sample of 40	0.93	0.95	0.94	0.92	0.94	0.92
Real estate researchers aim to understand hov	0.91	0.92	0.92	0.92	0.93	0.92
A professor who teaches a large introductory s	0.88	0.89	0.89	0.89	0.9	0.89
Participation in physical activity often declines	0.94	0.93	0.95	0.95	0.93	0.92
A university wellness survey asked students w	0.92	0.93	0.94	0.91	0.94	0.94

Table 5: Cosine similarity scores for all models. All scores are around 0.9, signifying that all justifications have close alignment with gold answers. Gemini has the highest average score (0.93), but all other models' average scores are 0.91-0.2, which is a very small difference.

Automated scores (BERTScore)						
Scenario	3B MCQ score	3B No MCQ score	8B MCQ score	8B No MCQ score	Gemini No MCQ score	3B + RAG score
Undergraduate students taking an introductory	0.7758	0.778	0.7947	0.797	0.7691	0.7963
A study of reaction times was conducted with 1	0.7814	0.7664	0.7902	0.7924	0.7202	0.8006
A survey of 55 Duke University students asked	0.8071	0.8092	0.7933	0.8095	0.756	0.8408
Subjects from Central Prison in Raleigh, NC, v	0.7876	0.7874	0.7865	0.8006	0.7423	0.7768
Microhabitat factors associated with forage and	0.7907	0.7617	0.7864	0.7608	0.7376	0.7962
Raina and Luke survey a random sample of 40	0.7875	0.7811	0.7927	0.8008	0.7484	0.7945
Real estate researchers aim to understand hov	0.7875	0.8093	0.801	0.8076	0.759	0.8115
A professor who teaches a large introductory s	0.7795	0.777	0.7685	0.7627	0.7573	0.7827
Participation in physical activity often declines	0.7997	0.8029	0.8108	0.8067	0.7743	0.7802
A university wellness survey asked students w	0.7742	0.8036	0.8109	0.7877	0.7818	0.8394

Table 6: BERTScore scores for all models. All scores are around 0.79, suggesting relatively great alignment with gold answers. 3B+RAG has the highest average score (0.8). All other models have average scores between 0.75 and 0.79.

All 3 evaluation metrics agree that all justifications for correct answers are relatively close to the rubric or golden answers. All scores hover around 5, 0.9, and 0.79 for the 3 metrics respectively. This shows that if a model suggests the correct test, they have the ability to provide most of the information required. Cosine similarity and manual evaluation agree in their rankings of models' justification quality. For both metrics, Gemini provides the most complete and accurate justifications in general.

However, surprisingly, with BERTScore, Gemini actually receives the lowest scores on average and Llama 3B+RAG performs best. This inconsistency highlights the difficulty of text evaluation in comparison to classification accuracy. Different metrics of evaluating semantic alignment can lead to different rankings and results. Additionally, even though manual scoring is more time consuming, the rubric and integer scores make results much more interpretable compared to more scalable automated approaches.

VI. Discussion

The results demonstrate a clear relationship between model scale, prompting strategy, and performance in statistical test selection. Larger models such as Gemini Flash 2.5 consistently outperforms smaller locally hosted models, both in classification accuracy and in the clarity of their reasoning. The smaller Llama 3B and Llama 8B models show more limited ability to distinguish between tests, provide clear explanations, and provide all required assumptions, particularly when not given explicit answer constraints.

The evaluation also highlights the importance of prompt structure. Multiple-choice prompting substantially improves performance for all models, especially the smaller ones, by constraining the list of valid choices. This effect was most dramatic for Llama 8B, whose accuracy drops sharply from MCQ to open-ended prompting. These findings reinforce that carefully designed prompts can meaningfully shape LLM outputs in classification-like tasks.

The RAG-enhanced Llama 3B model provides insight into how retrieval can compensate for limited model capacity. While RAG does not fully close the performance gap with larger models, it reduces errors and improves the quality of justifications when the retrieved text directly matches the scenario. These results confirm that retrieval is most beneficial when the model lacks internal domain knowledge and when relevant material appears in the knowledge base. However, retrieval can also introduce noise and mislead the model in cases where the retrieved excerpts are not entirely relevant, leading to unrelated reasoning and missclassifications.

Finally, the combination of rubric-based and automated evaluation provides a detailed view of explanation quality. Gemini's consistently strong scores reflect an ability not only to select the correct test but also to articulate justifications well. Smaller models are more likely to produce incomplete or inconsistent explanations, highlighting a limitation of current lightweight LLMs for complex tasks. The improvements observed with RAG indicate a potential for more resource-efficient models to support statistical reasoning tasks.

VII. Conclusion

This study examines how large language models of different scales perform on the task of selecting appropriate statistical tests and generating clear justifications from short descriptions of research scenarios. Across all evaluations, the state-of-the-art model Gemini Flash 2.5 performs best, demonstrating high classification accuracy and consistently strong explanations. Smaller locally hosted models such as Llama 3B and 8B show more frequent classification errors and weaker reasoning, especially under open-ended, non-MCQ prompting. Incorporating retrieval improves performance for the smallest model, reducing specific misclassifications and strengthening some explanations. However, it does not fully close the gap with larger models.

Overall, the findings show that while model scale is a very strong determinant of performance, retrieval-augmented generation has the potential to meaningfully support smaller models in Q&A tasks. These results highlight the potential for LLM-based tools to assist learners and researchers in understanding statistical test selection specifically, but also in Mathematics/Statistics education in general.

The learning reflection for the project can be found in Appendix A.

References:

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Diez, D. M., Barr, C. D., & Çetinkaya-RundelM. (2019). *OpenIntro statistics* (4th ed.). Openintro, Inc.

Es, S., James, J., Anke, L. E., & Schockaert, S. (2024, March). Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150-158).

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, November). Retrieval augmented language model pre-training. In *International conference on machine learning* (pp. 3929-3938). PMLR.

Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., & Kedzie, C. (2024). LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. *arXiv preprint arXiv:2501.00274*.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., & Lu, X. (2019, November). Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 2567-2577).

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, *7*, 453–466. https://doi.org/10.1162/tacl_a_00276

Levonian, Z., Li, C., Zhu, W., Gade, A., Henkel, O., Postle, M. E., & Xing, W. (2023). Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. *arXiv preprint arXiv:2310.03184*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459-9474.

Mondal, H., Mondal, S., & Mittal, P. (2024). Evaluating large language models for selection of statistical test for research: A pilot study. *Perspectives in clinical research*, *15*(4), 178–182. https://doi.org/10.4103/picr.picr_275_23

Najmi, A., Sadasivam, B., & Ray, A. (2021). How to choose and interpret a statistical test? An update for budding researchers. *Journal of family medicine and primary care*, *10*(8), 2763–2767.

Song, X., Lee, L., Xie, K., Liu, X., Deng, X., & Hong, Y. (2025). StatLLM: A Dataset for Evaluating the Performance of Large Language Models in Statistical Analysis. *arXiv preprint arXiv:2502.17657*.

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ digital medicine*, *7*(1), 258. https://doi.org/10.1038/s41746-024-01258-7

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhu, Y., Du, S., Li, B., Luo, Y., & Tang, N. (2024). Are large language models good statisticians?. *Advances in Neural Information Processing Systems*, *37*, 62697-62731.

Appendix A
Learning Reflection

This project has been both an incredibly rewarding and challenging learning experience. At the beginning of the semester, I had a very clear plan and goals I wanted to accomplish through this independent project. I knew I wanted to extend my last summer work, have a research project that somehow combines Statistics, Education, and LLMs, compare the performance of different models, and learn more about RAG. I was very excited about the plan. However, as with any worthwhile Computer Science project, things don't always go smoothly, and it was the challenges that really helped me expand my skillset and confidence.

Firstly, this is my first experience with creating my own dataset for a research question, and the process was very interesting. I needed to make sure that each research scenario makes sense, that the labels are correct, and that the dataset is balanced. Even though this part of the project is quite time consuming, I learned a lot about dataset creation, such as how to look for good sources and how to best tailor and organize compiled data to a specific task.

The second, and probably most significant, challenge of this project was definitely learning how to implement RAG. I learned that there are many variables in a RAG workflow, such as different strategies of chunking the information source, different embedding methods, and different approaches to assessing semantic similarity between texts. The modest results reflect my limited experience with RAG, but the process itself was incredibly instructive. Experimenting with all of the variables and critically thinking about how they can impact results not only expanded my technical skillset but also gave me a deeper understanding of limitations and what is required to use RAG effectively.

Finally, the stage of analyzing and writing up the results and discussion introduced its own set of challenges. Working with four models, two prompting strategies, and both classification and text evaluation tasks created an abundance of possible observations and avenues for investigation. Selecting the most significant findings, determining how to organize them coherently, and trying to make sure that the analysis is accurate and representative of the project required a lot of time and effort.

Despite these difficulties throughout the project, the experience was overwhelmingly positive. I learned a great deal, strengthened my scientific writing skills, and gained a deeper appreciation for the research process as a whole. Ultimately, this was exactly the kind of challenge and growth I was seeking by doing an independent study.