Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024

Use of Fine-Tuned LLMs in Engineering Education: Evaluating Answer Quality using Metric-Based Scoring

Siddhesh More
SIES Graduate School of Technology
Navi Mumbai, India

Saniya Nande
SIES Graduate School of Technology
Navi Mumbai, India

***Abstract*— Using a variety of large language models (LLMs)—such as Llama 3.1, Llama 3, Llama 2, Mistral and Phi 3.5—this study aims to provide chapter-by-chapter academic content for engineering courses. The objective is to assess these models by providing answers to exam questions from the past. Several quantitative metrics, such as ROUGE, Cosine Similarity, METEOR, Coherence, and Accuracy scores, are used to compare the generated replies with the original answers. Teaching academics evaluate the responses provided by the top three performing models manually, assigning a grade based on the academic quality and relevance of each response. The goal of this comparative study is to determine which model performs the best when it comes to algorithmic and human-based evaluation.**

***Keywords*— Large language models, Llama 3.1, Llama 3, Llama 2, Mistral, Phi 3.5, engineering courses, chapter-wise content, past exam questions, ROUGE, Cosine Similarity, METEOR, Coherence, Accuracy score, teaching academics, human evaluation, academic quality, relevance, algorithmic performance, educational content delivery.**

I. INTRODUCTION

In the field of natural language processing (NLP), large language models (LLMs) have become revolutionary tools due to their exceptional ability to produce text that is human-like in a variety of applications. The ways in which we engage with and handle textual data have been completely transformed by these models, from chatbots and content generation to translation and summarization. LLMs are incredibly promising in the field of education, especially when it comes to creating academic content. Their ability to analyze and synthesize large volumes of information and deliver well-reasoned, contextually appropriate responses makes them ideal for supporting teaching, learning, and evaluation processes.

LLMs like Llama 3.1, Llama 2, Mistral, and Phi 3.5 have become more potent because of the constant improvements in model topologies, such as transformers, and the volume of training data. These models may now specialize in domain-specific tasks through fine-tuning. Optimizing these models for scholarly fields like engineering offers a special chance to automate the creation of excellent, topic-appropriate material. When responding to intricate exam questions, where both breadth and depth of topic knowledge are essential, this can be especially helpful.

Exam questions typically call on in-depth topic knowledge, and the capacity to clearly communicate difficult ideas. So, the question is: Is it possible to fine-tune LLMs to mimic these kinds of cognitive processes? If yes, what level of accuracy can they achieve in producing responses that demonstrate coherence, accuracy, and academic quality?

This study aims to evaluate the ability of multiple cutting-edge LLMs to respond to engineering exam questions chapter-by-chapter. We concentrate on optimizing these models to get proficiency in engineering content by assessing their capacity to produce precise, comprehensive, and rationally sound responses. To achieve this, we use a range of quantitative criteria to assess how well the generated answers correspond with the original, human-provided solutions, including ROUGE, Cosine Similarity, METEOR, Coherence, and Accuracy. These measures enable us to compare each model's performance in a methodical way across a variety of characteristics, including factual correctness, linguistic quality, textual overlap, and semantic relevance.

We also provide a qualitative component through human evaluation by academic experts, in addition to computational examination. Engineering educators will grade the generated answers from the best-performing models according to academic relevance, accuracy, and clarity. This human evaluation is essential to assess how well these models fulfil academic expectations in the actual world and guarantees that the output is accurate, useful, and compliant with rigor standards for education.

The aim of this research is to find an LLM that can help teachers and students create excellent academic content with little to no human involvement. This concept could be used to create study materials for exams, automate processes for answering questions, and assist teachers in developing extra resources. By optimizing LLMs to focus on academic fields, we can improve student learning, lighten the burden on teachers, and give students quick access to trustworthy, organized information.

II. MOTIVATION

The number and complexity of educational content is always increasing; hence it is necessary to develop automated tools to help instructors and students create, assess, and improve academic content. The creation of academic content might be completely transformed by the application of LLMs in this field, which would guarantee quality while minimizing manual labor. It is possible to automate test preparation, question answering, and information delivery in a way that is contextually appropriate and adaptive by fine-tuning LLMs for subject-specific content. The goal of this research is to find an

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024

optimal model that not only fits the academic requirements set by educators but also performs well algorithmically, with the goal of bridging the gap between machine-generated material and human review.

III. OBJECTIVES

Optimize LLMs on Engineering Content: Develop chapter-by-chapter expertise for precise and context-specific responses by fine-tuning models like Llama 3.1, Llama 2 and Mistral etc. for engineering subjects.
Evaluate Exam Question Responses: Test the models by producing responses to actual previous exam questions and contrasting them with responses supplied by humans.
Make Use of Quantitative Measures: Examine model results using:
ROUGE (overlap of text)
- Cosine Similarity, or semantic correspondence
- METEOR (recall and precision)
- Coherence (flow of reasoning)
- Accuracy (correctness of facts)
Select Best Model for Academic Use: Decide which model will improve student learning and automate exam preparation the best.
Aimed at AI-Assisted Learning: Promote the application of LLMs in the classroom, enhancing the effectiveness and accessibility of scholarly materials.

IV. PROPOSED METHODOLOGY

The methodology that is being suggested provides a step-by-step guide for optimizing large language models (LLMs) on engineering subjects and assessing the quantitative and qualitative aspects of their performance.

1. Gathering and Preparing Data:
- Data Source: Gather engineering course materials, online resources, and textbooks' chapter-by-chapter content. Collect previous test questions and the answers from reliable academic resources.
- Preparation: Clean up the gathered data by deleting unnecessary information, tokenizing the content, structuring the text, and dividing it into chapters or themes for further subject-specific analysis.

2. Adjusting LLMs:
- Model Selection: For fine-tuning, choose cutting-edge LLMs (Llama 3.1, Llama 2, Mistral etc.). These models need subject-specific fine-tuning to adapt to engineering content because they have already been pre-trained on vast corpora.
- Fine-tuning Process: Apply a chapter-by-chapter method to each LLM to refine them on the prepared engineering subject data and get specialized knowledge in particular areas.

3. Exam Questions with Model Inference:
Enter Test Questions: Give the refined LLMs a set of past exam questions so they can draw conclusions. Every model will produce responses to these queries, modelling its ability to react in accordance with the insights it gained throughout refinement.
- Assembled Responses: Save the responses that each model produces so that they can be assessed further.

4. Analytical Assessment:
- Comparative Measures: Compare the model-generated responses to the initial human-provided responses based on a variety of metrics:

- ROUGE (Understudy for Gisting Evaluation with an emphasis on recall): Measures the degree of overlap between the reference and generated answers.

- Cosine Similarity: Determines how similar the two sets of answers are semantically to make sure the text produced by the model makes sense.

- Metric for Evaluation of Translation with Explicit Ordering, or METEOR: evaluates recall and precision using linguistic factors such as stemming and synonymy.

- Coherence: Indicates how well-organized and logically coherent the generated responses are.

- Accuracy Score: Compares particular points or concepts in the answers to determine factual correctness.

- Ranking Models: These scores should be used to rank the performance of each LLM, with the top 3 models being determined by their capacity to produce responses that are accurate, pertinent, and logical.

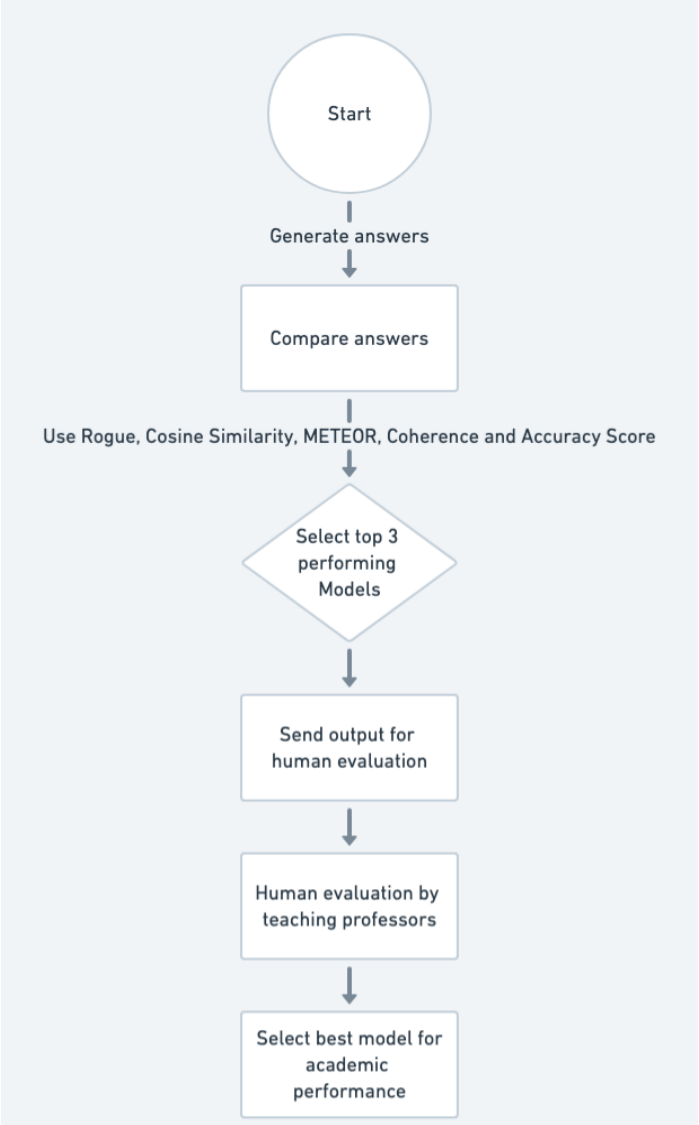
5. Professors' Human Evaluation:
- Selection of Top 3 Models: Choose the top 3 LLMs with the best overall performance based on quantitative parameters.
- Professor Evaluation Process: Show teaching professors who are experts in the relevant engineering areas the generated solutions from these three models. Instructors will grade each response according to:
- Academic Relevance: The degree to which the responses conform to the anticipated academic requirements.
- Correctness: The responses' technical content's accuracy.
- Ranking: For each response, professors will provide a score (out of 10) that will be used in the qualitative analysis.

6. Optimal Model Final Selection:
- Comparison of Human and Quantitative Scores: Compile the findings of the human assessment along with the quantitative scores to determine which model works optimally both algorithmically and subjectively.
- Best Model Identification: Select the LLM that produces the most accurate and academically relevant content, giving it the most suitable option for usage in educational environments.

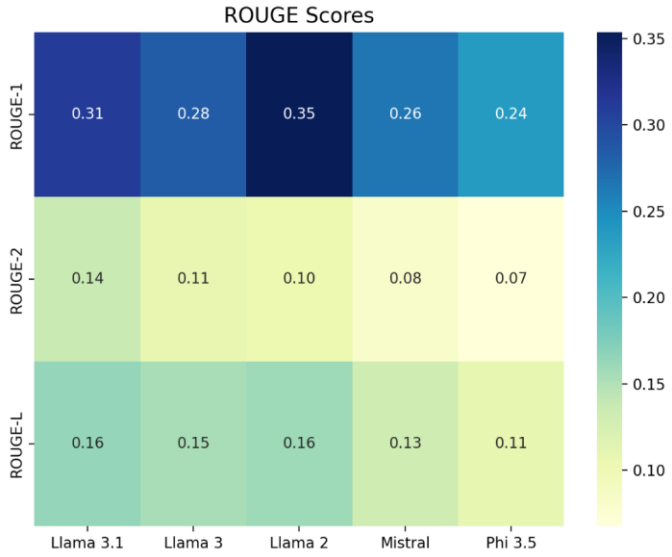
7. Application of Best Model in Academic Settings:
- AI-Assisted Learning: Explain how the chosen LLM can improve educational results by providing accurate and well-organized academic content to AI-assisted learning tools.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024



V. RESULTS AND DISCUSSION



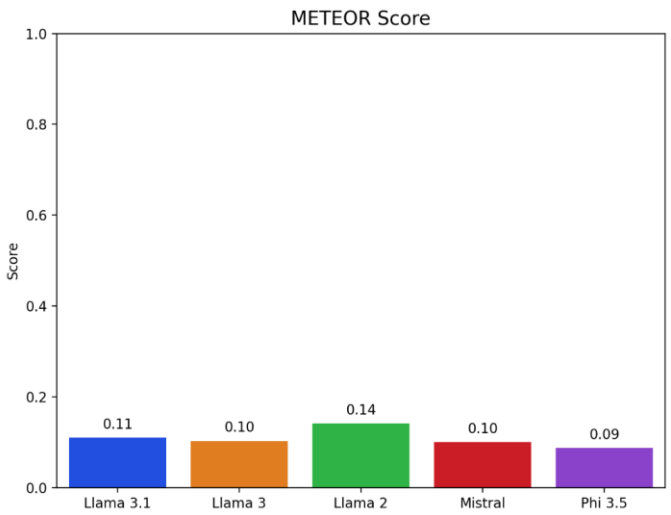
5.1 Rouge score

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used to evaluate automatic summarization and machine-generated text by comparing it to a reference text. It primarily measures the overlap of n-grams (e.g., unigrams, bigrams) between the generated answer and the reference answer. ROUGE places more emphasis on recall, making it particularly useful for evaluating whether all relevant information has been captured in the generated text.

ROUGE is employed to determine how much of the original human-provided answer is captured in the model-generated response. Given the academic context, this is crucial, as students' answers are expected to cover key points from reference material. A high ROUGE score suggests that the model has successfully identified and incorporated critical information from the reference answer.

ROUGE is calculated by comparing the n-gram overlap between the model's output and the reference answer. The more overlap, the higher the ROUGE score.

Top 3 Rouge model scores are: Llama 2, Llama 3.1 and Llama 3



5.2 METEOR Score

METEOR (Metric for Evaluation of Translation with Explicit Ordering) is a metric that measures the precision and recall of generated text by considering synonymy, stemming, and word order. Unlike ROUGE, it accounts for linguistic variations like synonyms and reordering's, making it more flexible in assessing the quality of the output.

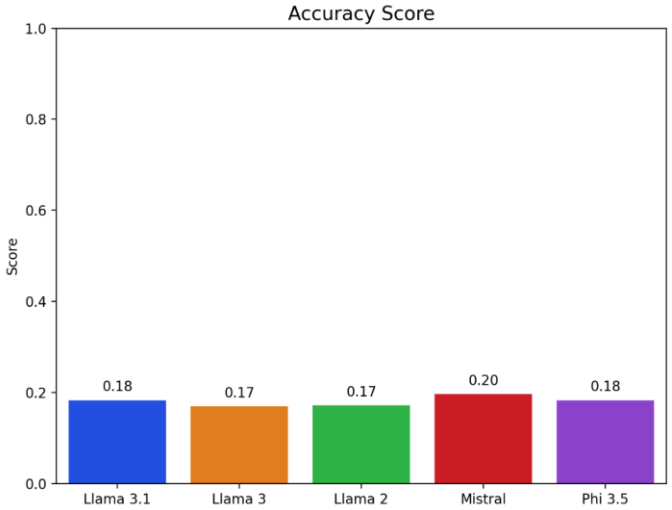
METEOR is especially useful in evaluating the flexibility and linguistic accuracy of model-generated text, as academic answers might use different wording or phrasing while still conveying the same information. It also penalizes errors such as disordered answers.

METEOR is calculated by matching the generated text with the reference answer using stems, synonyms, and word order. Higher METEOR scores indicate that the model-generated answer is both precise and linguistically well-formed.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024

Top 3 METEOR model scores are: Llama2, Llama 3.1 and Llama 3



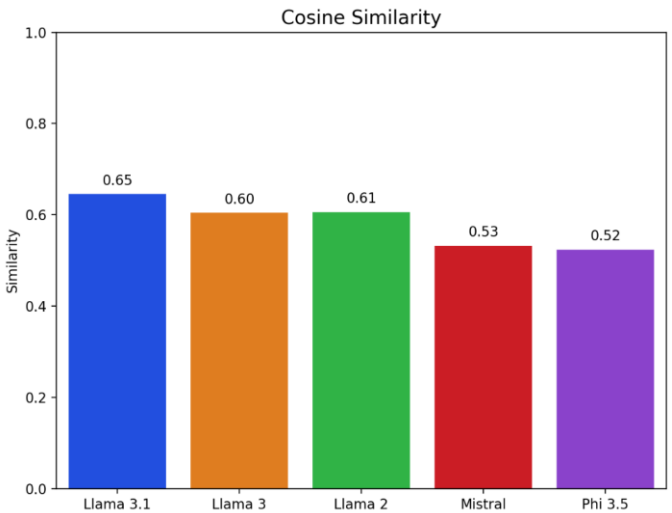
5.3 Accuracy Score

Accuracy score measures the factual correctness of the generated text by comparing it to the reference answer. It checks whether the key facts, concepts, formulas, and terminologies are correctly presented in the model-generated answers.

Since engineering exam answers often require precise facts, equations, and logical reasoning, the accuracy score ensures that the generated answers are not only coherent but also factually correct.

The accuracy score is determined by comparing factual elements in the generated text with those in the reference answer. Any factual errors or omissions reduce the accuracy score.

Top 3 Meteor model scores are: Mistral, Llama 3.1 and Phi 3.5



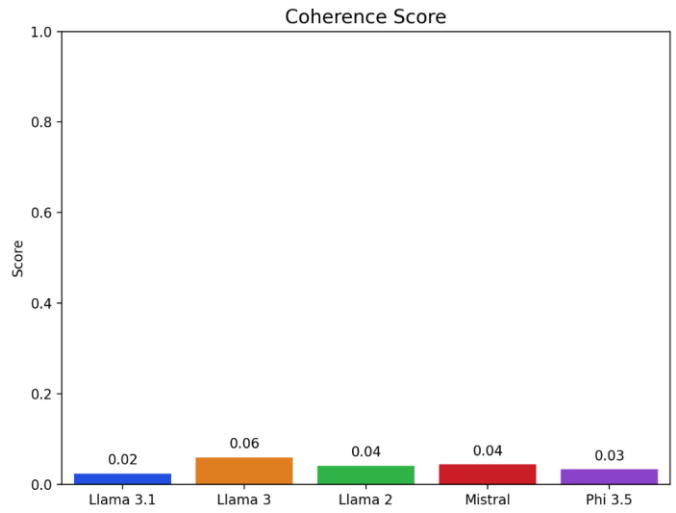
5.4 Cosine Similarity

Cosine similarity is a measure of the semantic similarity between two texts by calculating the cosine of the angle between their vector representations in a multi-dimensional space. It ranges from -1 to 1, with 1 indicating identical content in terms of meaning.

This metric evaluates how closely the generated answers align with the meaning and intent of the reference answers. It ensures that even if the text is not a word-for-word match, the overall meaning is preserved.

Cosine similarity is computed by vectorizing the text from both the model-generated and reference answers, then calculating the cosine of the angle between them. A higher score indicates greater semantic similarity.

Top 3 Cosine Similarity scores are: Llama 3.1, Llama 2 and Llama 3



5.5 Coherence Score

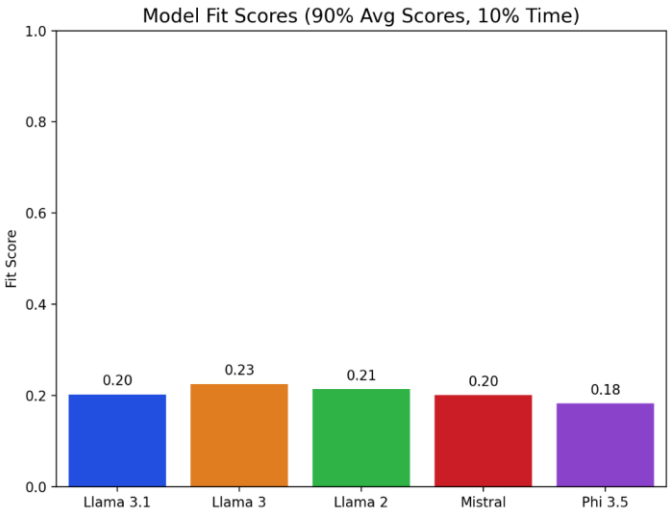
The coherence score measures the logical flow and consistency of the generated text. It assesses whether the answer follows a logical structure, whether ideas are presented in a clear and understandable manner, and whether the answer remains on topic.

In academic answers, coherence is essential for presenting concepts and reasoning in a logical, structured way. A high coherence score indicates that the generated answers are easy to follow and logically constructed, important for engineering subjects that require structured problem-solving.

Coherence is scored by analyzing the logical transitions between sentences and paragraphs in the generated text. Inconsistent or poorly structured responses result in a lower coherence score.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024

Top 3 Coherence scores are: Llama 3, Llama 2 and Mistral



5.6 Model Fit score

After collecting the scores from each of the metrics, a composite score—referred to as the Model Fit Score—was calculated for each of the top models. This score represents a weighted aggregate of the various evaluation metrics, with 90% of the score derived from the average of the ROUGE, Accuracy, Coherence, Cosine Similarity, and METEOR scores, and 10% based on the time it took for each model to generate its answers. The inclusion of time efficiency reflects the practical importance of having models that are not only accurate and coherent but also computationally efficient, making them more viable for real-world applications.

Calculation of Model Fit Score:

Formula:
$$\text{Model Fit Score} = (0.9 \times \text{Average of Evaluation Metrics}) + (0.1 \times \text{Time Factor})$$

The Model Fit Score is designed to assess both the quality of responses generated by the models and their computational efficiency. This score is calculated with 90% weightage given to linguistic and semantic performance metrics (ROUGE, Accuracy, Coherence, Cosine Similarity, and METEOR) and 10% weightage given to the time taken by each model to generate answers. By considering time, the evaluation accounts for not just the accuracy and coherence of the model-generated answers, but also how quickly the model can produce these results, which is crucial in real-world applications.
The time taken by each model to produce answers varied significantly. Llama 3 was the fastest, with a generation time of 49.52 seconds, while Llama 3.1 was the slowest, taking 75.56 seconds. Llama 2 took 67.18 seconds, placing it between the two. Other models, like Mistral and Phi 3.5, had times of 55.28 seconds and 58.44 seconds respectively, but did not perform as well in the linguistic metrics.

When incorporating both the performance and time into the Model Fit Score, Llama 3 emerged as the best-performing model, achieving a score of 0.23 due to its superior linguistic performance and faster processing time. Llama 2, while delivering high-quality answers, was slower and had a slightly lower score of 0.21. Llama 3.1, although balanced in terms of quality, was slower and less accurate than Llama 3, resulting in a score of 0.20. This evaluation shows that Llama 3 is the optimal choice for generating high-quality, coherent, and factually accurate academic content while maintaining efficient processing speeds.

Human Evaluation

1) Selection of Best Models: From the total five models that were fine-tuned and evaluated, the top 3 performing models were selected based on their scores in metrics such as ROUGE, Cosine Similarity, METEOR, Coherence, and Accuracy.
2) Anonymous Answer Labelling: The answers generated by the selected models were anonymized and labeled as Answer 1, Answer 2, and Answer 3. This ensured that the evaluators were unbiased and unaware of which model generated each answer.

3) Professor Involvement: The labeled answers were then sent to two domain experts:

- Dr. Varsha Patil (Associate Professor | Computer Engineering): Evaluated the answers for the Natural Language Processing (NLP) subject.

- Prof. Rohini Gaikwad (Assistant Professor | AI-DS): Evaluated the answers for the Computer Networks (CN) subject.

These professors rated the answers based on factors such as correctness, completeness and overall relevance to the input questions.

4) Alignment of Human Evaluation with Model Fit Scores: The responses from both professors were in alignment with the Model Fit Scores. Specifically, the rankings assigned by the evaluators matched the performance scores of the models, with:

- Llama 3 ranking 1st,
- Llama 2 ranking 2nd,
- Llama 3.1 ranking 3rd.

This alignment between the human evaluations and the quantitative metrics reinforced the reliability of the proposed benchmarking approach, confirming that the combined use of automated scoring metrics and expert evaluation can effectively identify the best-performing model for generating high-quality academic content.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024

VI. CONCLUSION

In this research, we fine-tuned multiple large language models (LLMs), including Llama 3.1, Llama 2, Mistral etc on engineering educational data and compared their performances in generating academic content. Using a combination of quantitative metrics—such as ROUGE, Cosine Similarity, METEOR, Coherence, and Accuracy—and human evaluation by subject experts, we identified Llama 3 as the best-performing model for academic content generation, followed by Llama 2 and Llama 3.1.

The alignment between human assessments and automated scoring validated the robustness of our evaluation framework, indicating that LLMs like Llama 3 can be effectively utilized for academic purposes, providing high-quality answers to engineering questions. Future research can expand on this by exploring further fine-tuning techniques and incorporating more diverse evaluation criteria to continue improving model accuracy and reliability in the educational domain.

VII. FUTURE SCOPE

There is a chance that this project will see major advancements in the future. By extending the fine-tuning methodology to additional academic fields including the humanities, law, and medicine, large language models (LLMs) can provide a wider variety of educational content. Better human-AI cooperation may also develop, with teachers utilizing LLMs to create customized learning materials, automate grading, and give students immediate feedback. The research's top-performing model might be included into online learning environments to improve material delivery, test-taking strategies, and student-specific adaptive learning environments. AI's place in contemporary education is further cemented by the model's ability to adapt over time through ongoing upgrades and fine-tuning that guarantee it stays in line with curricular innovations and current academic standards.

VIII. ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude to those who contributed to the success of this research. Our sincere thanks go to Dr. Varsha Patil for her invaluable insights and expertise in evaluating the outputs related to Natural Language Processing (NLP), and to Prof. Rohini Gaikwad for her thorough assessment of the Computer Networks (CN) content. Their feedback and guidance were instrumental in refining the evaluation process and enhancing the quality of this study.

IX. REFERENCES

[1] Y. Chang et al., “A Survey on Evaluation of Large Language Models,” ACM Transactions on Intelligent Systems and Technology, vol. 15, no. 3, pp. 1–45, Jan. 2024, doi: 10.1145/3641289.

[2] Li, Q. et al. (2024) Adapting large language models for education: Foundational capabilities, potentials, and challenges, arXiv.org. Available at: https://arxiv.org/abs/2401.08664 (Accessed: 24 September 2024).

[3] F. Umar, “Enhancing Document Accessibility and User Interaction through Large Language Model: A Comparative Study for Educational Content : A Comparative Analysis of LLM and Traditional Site Search,” DIVA, 2024. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1873179&dswid=-6888

[4] M. R. J, K. Vm, H. Warrier, and Y. Gupta, “Fine Tuning LLM for Enterprise: Practical guidelines and recommendations,” arXiv.org, Mar. 23, 2024. https://arxiv.org/abs/2404.10779

[5] M. M. Rashid et al., “Humanizing AI in Education: A Readability Comparison of LLM and Human-Created Educational Content,” Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Aug. 2024, doi: 10.1177/10711813241261689.

[6] O. Mañas, B. Krojer, and A. Agrawal, “Improving automatic VQA evaluation using large language models,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 5, pp. 4171–4179, Mar. 2024, doi: 10.1609/aaai.v38i5.28212.

[7] A. Aynetdinov and A. Akbik, “SemScore: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity,” arXiv.org, Jan. 30, 2024. https://arxiv.org/abs/2401.17072

[8] Y. Xia et al., “Understanding the performance and estimating the cost of LLM Fine-Tuning,” arXiv.org, Aug. 08, 2024. https://arxiv.org/abs/2408.04693