Automated MCQ generation using domain specialized SLM on Current Affairs

Final Presentation

Name: ABDUL WAZED
Mentor’s name: Dr. Vivek Kumar Dwivedi

Executive Summary

- **Challenge**: Manual MCQ creation in current affairs is time-consuming and inconsistent
- **Solution**: Domain-specialized Small Language Model (SLM) for automated MCQ generation
- **Approach**: Curated news datasets → SLM fine-tuning → Evaluation (BLEU, ROUGE, BERTScore + expert review: Relevance, Clarity, Correctness, Distractor quaity, Cognitive Level)
- **Results**: Comparison of SLM vs. general LLMs on accuracy, domain relevance & efficiency
- **Usage**: Scalable, unbiased, and timely assessments for educators and e-learning platforms

Problem Statement

With the rise of generative AI, language models are increasingly used for educational tasks such as generating Multiple Choice Questions (MCQs). Large Language Models (LLMs) like Gemini offer broad language capabilities but often lack up-to-date knowledge of rapidly changing domains like current affairs. In contrast, Small Language Models (SLMs), when fine-tuned on domain-specific news content and retrained regularly, may generate more relevant and up to date content despite limited generalization.

This capstone examines whether domain-specific SLMs can outperform LLMs in generating MCQs from unseen current affairs content, specifically in terms of factual accuracy, temporal relevance, and distractor quality, addressing a key research gap in evaluating model effectiveness in time-sensitive, real-world educational applications.

Gap Analysis

- **Lack of current-affairs–specific MCQ resources.** Most QG work targets education or general reading-comprehension; news/current-affairs datasets are mostly QA rather than MCQ with high-quality distractors. Notable exceptions like **NewsQuizQA** exist, but broader, up-to-date MCQ corpora for rapidly evolving news are scarce. [1]

- **Under-explored potential of domain-specialized SLMs:** Small Language Models (SLMs) promise efficiency and competitive quality when specialized, but there's little empirical evidence on SLMs fine-tuned for current-affairs MCQ generation (vs. larger general LLMs). Systematic comparisons of SLM + retrieval/GPT2 fine-tuning vs. LLM baselines are largely missing. [2]

Research Questions

- **RQ1**: What is the most effective approach for developing a Small Language Model (SLM) for MCQ generation on current affairs—building a custom model from scratch or fine-tuning a pre-trained GPT-based model such as **GPT2** or DistilGPT2 or TinyLLaMA?

  - **Null Hypothesis (H_0):** There is **no significant difference** in the effectiveness of MCQ generation on current affairs between a custom-built SLM trained from scratch and an SLM obtained by fine-tuning a pre-trained GPT-based model.

  - **Alternative Hypothesis (H_1):** An SLM obtained by **fine-tuning a pre-trained GPT-based model** (e.g., GPT-2, DistilGPT-2, TinyLLaMA) demonstrates **significantly higher effectiveness** in MCQ generation on current affairs compared to building a custom model from scratch.

Research Questions

- **RQ2:** While fine-tuning is expected to enhance an SLM's ability to generate relevant MCQs, does domain-specific fine-tuning make it more effective (in terms of overall score) than a general-purpose Large Language Model (LLM) such as **Gemini** in generating high-quality MCQs?

  - **Null Hypothesis (H_0):** There is **no significant difference** in the effectiveness of MCQ generation quality between a domain-specific fine-tuned Small Language Model (SLM) and a general-purpose Large Language Model (LLM) such as Gemini.

  - **Alternative Hypothesis (H_1):** A domain-specific fine-tuned Small Language Model (SLM) demonstrates **significantly higher effectiveness** in generating high-quality MCQs compared to a general-purpose Large Language Model (LLM) such as Gemini.

Research Questions

- **RQ3:** Is a Small Language Model (SLM) capable of generating accurate answers for MCQs, and how does its accuracy compare to that of MCQs generated by a Large Language Model (LLM) using the same prompts?

  - **Null Hypothesis (H_0):** A Small Language Model (SLM) is **not significantly different** from a Large Language Model (LLM) in terms of answer accuracy for MCQs generated with the same prompts.

  - **Alternative Hypothesis (H_1):** There **is a significant difference** in answer accuracy between MCQs generated by a Small Language Model (SLM) and those generated by a Large Language Model (LLM) using the same prompts.

Research Questions

- **RQ4:** Measuring distractor quality is a crucial aspect of evaluating MCQs, especially when generated by AI models such as Small Language Models (SLMs). How does the distractor quality of MCQs generated by a fine-tuned SLM compares to those produced by a general-purpose Large Language Model (LLM) like Gemini?

  - **Null Hypothesis (H_0):** There is **no significant difference** in distractor quality between MCQs generated by a fine-tuned Small Language Model (SLM) and those produced by a general-purpose Large Language Model (LLM) such as Gemini.

  - **Alternative Hypothesis (H_1):** There is a **significant difference** in distractor quality between MCQs generated by a fine-tuned Small Language Model (SLM) and those produced by a general-purpose Large Language Model (LLM) such as Gemini.

Data Description


RSS_Feed.txt

- Times of India RSS Feeds

This news portal provided latest news for multiple topics (India, Politics, World, Environment etc.) in RSS format. The content from this source have been used for following purpose

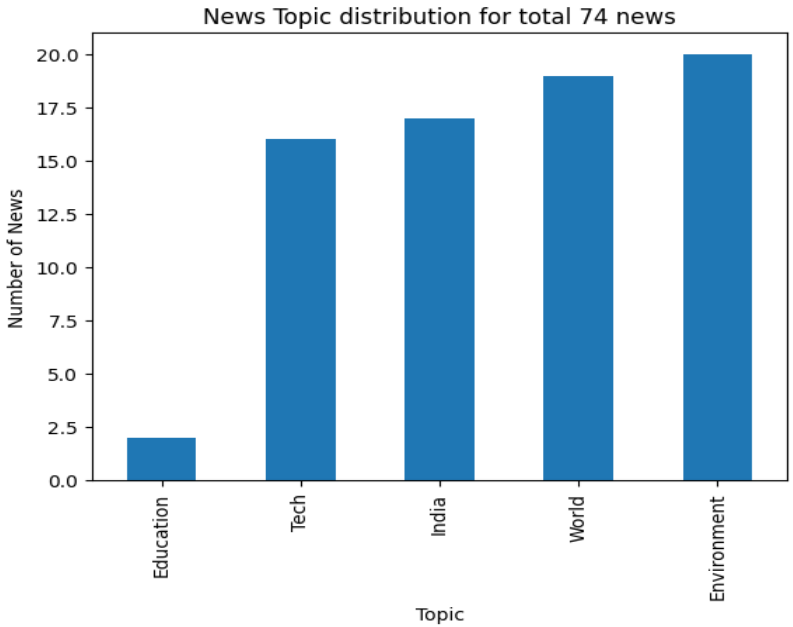
  - Extract, curate, archive current affairs news content
  - Train SLM from scratch on small subset of news content
  - Use as input context for generating current affairs MCQ using finetuned SLM and LLM

- GKToday - Current Affairs, GK (General Knowledge), General Studies for UPSC, IAS, Banking

GK TODAY is well-known portal on educational domain, provides monthly MCQs (multiple choice question) on current affairs for its Indian subscribers. This data source have been used for following purpose

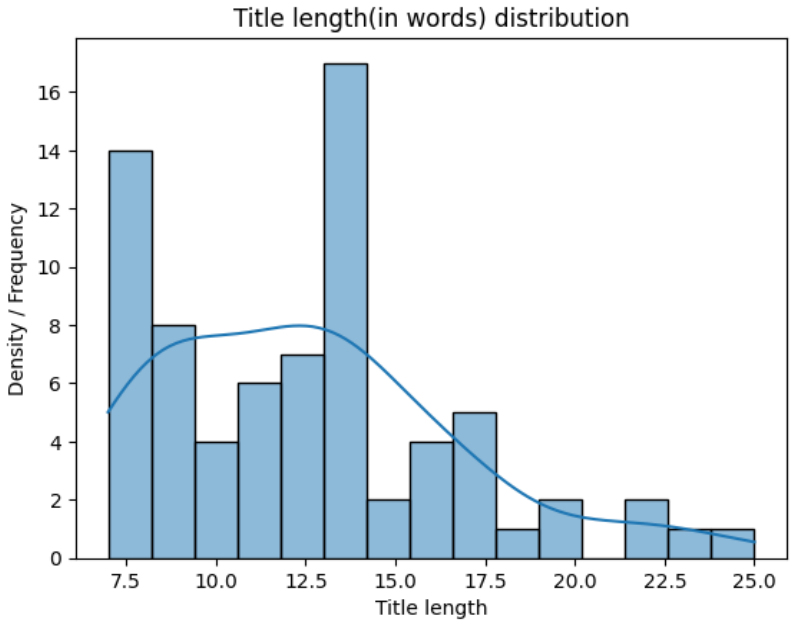
  - Extract, curate, archive current affairs MCQs for SLM fine tuning
  - Prepare language model fine tuning instruction
  - Fine tune GPT2 based SLM

EDA – Topic Distribution



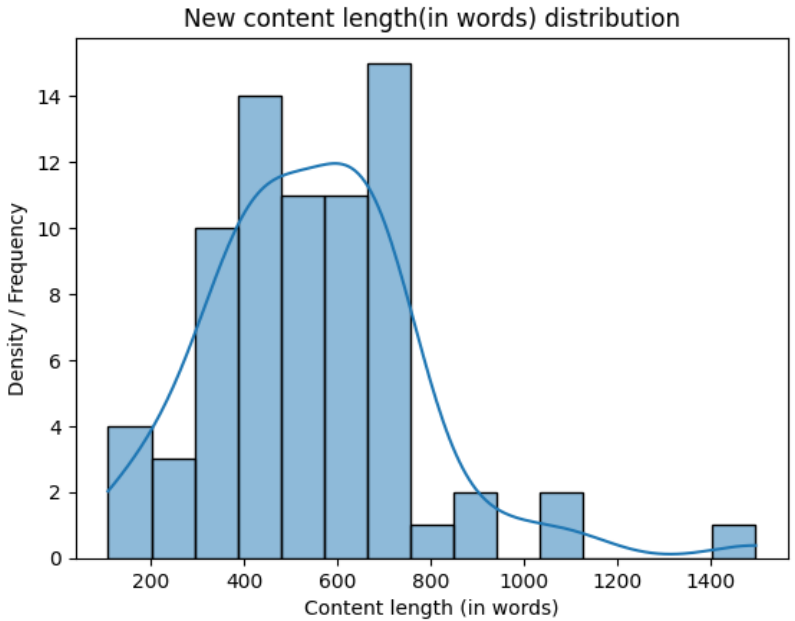
- The dataset spans 5 unique topics.
- Environment is the most frequent topic, followed by World, and India.
- This indicates a topic imbalance (Environment is over-represented) than Education topic.

EDA – News Title Distribution



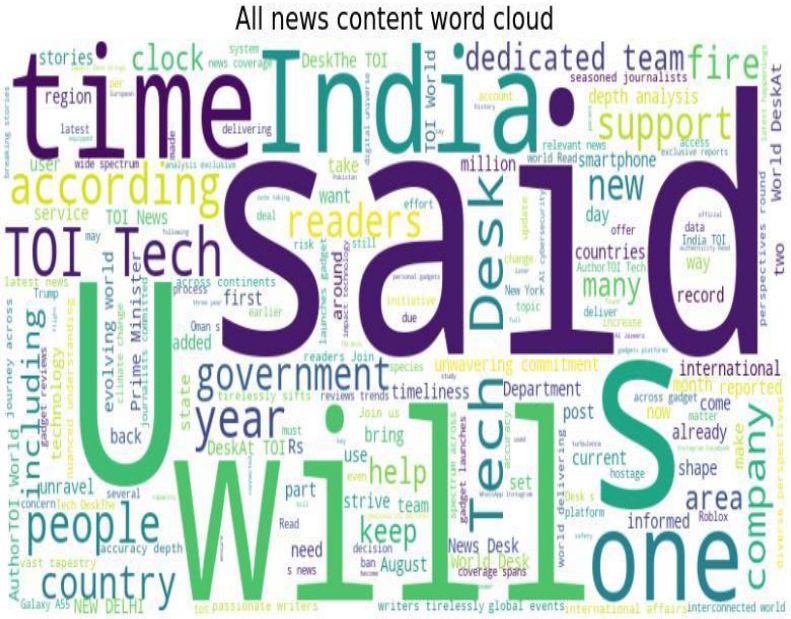
- Average title length ≈ **12.5 words**.
- Most titles fall within **8–15 words**, showing concise and headline-like phrasing, suitable for MCQ stems.

EDA – News content distribution



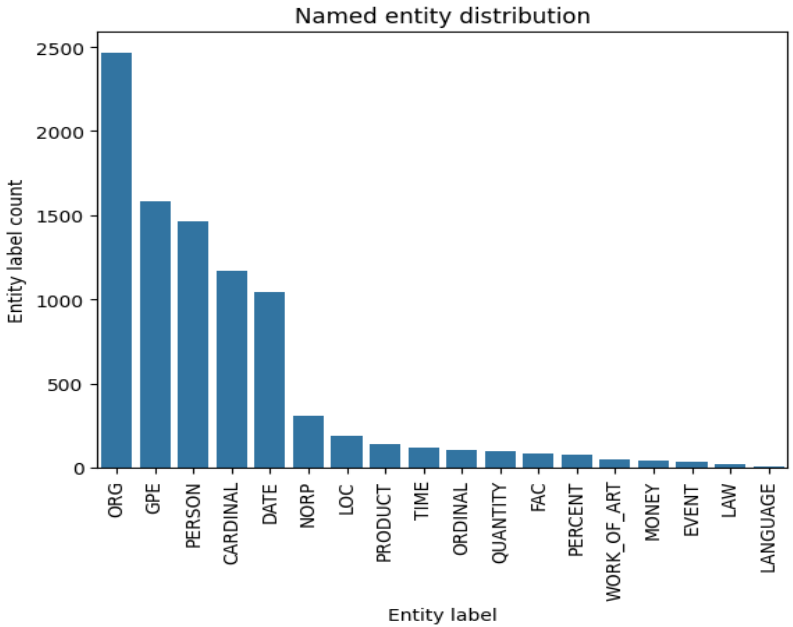
- Average input length ≈ **548 words** per article.
- Inputs are fairly detailed, ensuring enough context for generating fact-based MCQs.
- Distribution shows a majority of articles between **400–700 words**.

EDA – News content word cloud



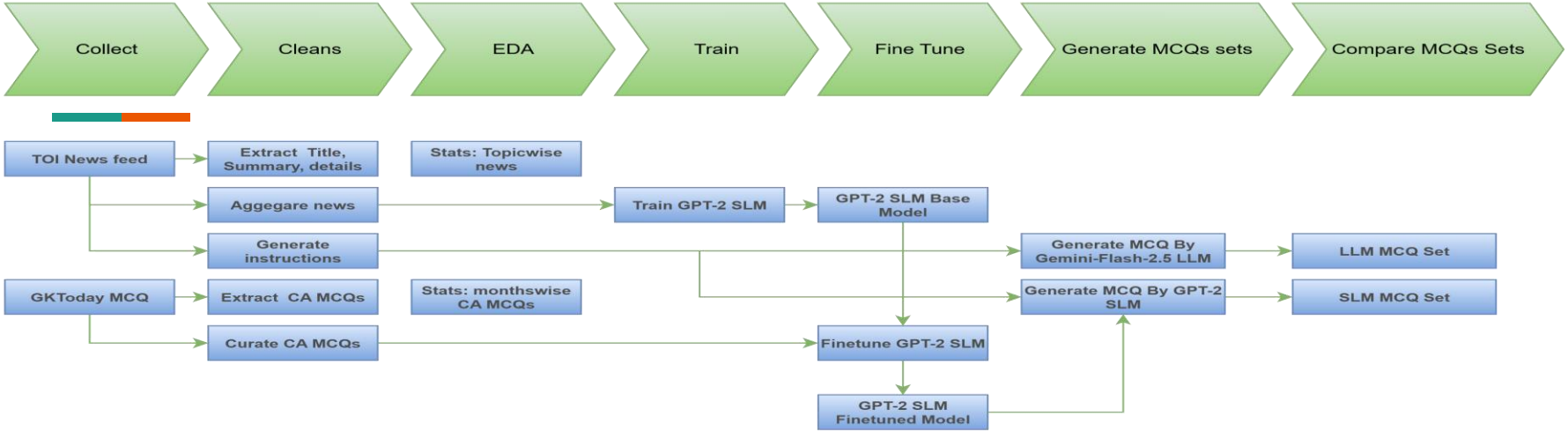
- **Dominant Terms**: Words like “said”, “India”, “will”, and “time” appear most frequently. **This indicates that much of the dataset centers around statements, future actions, and temporal references.**
- **Thematic Signals**: Strong presence of “**government**”, “**country**”, “**people**”, and “**support**” shows the dataset emphasizes political and social current affairs. Words such as “tech”, “TOI Tech”, and “company” suggest substantial coverage of technology and business-related news.
- **Reporting Style**: Frequent words like “**according**”, “**said**”, “**keep**”, and “**help**” show a reporting and explanatory narrative style, consistent with journalistic writing.

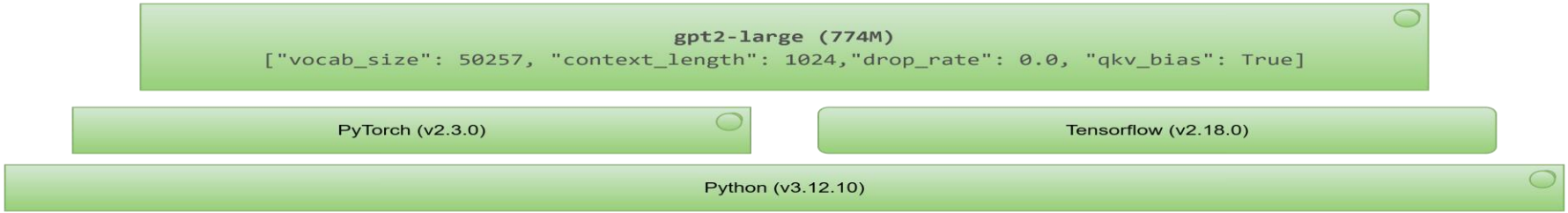
EDA – News content named entity distribution



- **Organizations (ORG: 2,466**) are the most frequently occurring entities, reflecting the heavy emphasis on institutions such as governments, political parties, corporations, and agencies in Current Affairs.
- **Geopolitical Entities (GPE: 1,586)** and Persons (PERSON: 1,466) also appear prominently, which is expected in news reporting since Current Affairs often highlight leaders, officials, and countries.
- **Cardinal numbers (CARDINAL: 1,170)** and Dates (DATE: 1,044) are highly frequent, showing that news stories often cite numerical facts, counts, and specific timelines
- **NORP (310) entities** (Nationalities, Religious or Political groups) highlight identity-based references, relevant for questions on international relations and politics.
- **Locations (LOC: 190)** and Facilities (FAC: 84) add geographical and infrastructural details that can enrich contextual MCQs.

Architecture diagram/Workflow





Laptop
AMD RYZEN 1.70 GhZ, 32 GB RAM (No GPU)

Model building – gpt2-large (774M) baseline

gpt2-large (774M) chosen as baseline Small Language Model for training on unlabeled latest news content as well as fine-tuning for Automated MCQ Generation on Current Affairs. Why GPT-2 is a Good Choice:

- **Model Size vs. Efficiency:** GPT-2 comes in multiple sizes (124M, 355M, 774M, 1.5B parameters). The smaller variants (124M / 355M / 774M) are lightweight, fast to train, and require modest GPU/CPU resources — practical for a student capstone project
- **Proven Baseline for Text Generation**: GPT-2 is one of the earliest autoregressive transformers widely adopted for text generation.
- **Ease of Fine-Tuning:** For domain specialization (e.g., Current Affairs), GPT-2 can be easily fine-tuned on curated datasets without complex infrastructure.
- **Controllability**: With careful prompt design and dataset curation, GPT-2 can be steered to generate structured MCQs rather than long, open-ended outputs.

Results – Score Comparision

SLM (GPT2-large(774M)

- Generated MCQ Count: 74
- Avg BLEU Score: 0.0001985
- Average ROUGE Score: 0.03463242
- Average BERT Score: 0.6892379
- Average Relevance Score: 1.662162162
- Average Clarity Score: 1.541
- Average Correctness score: 0.6756757
- Average Distractor quality score: 1.689189189
- Average Cognitive level score: 0.621621622

LLM (Gemini-Flash-2.5)

- Generated MCQ Count: 74
- Avg BLEU Score: 0.002533919
- Average ROUGE Score: 0.11142004
- Average BERT Score: 0.811804149
- Average Relevance: 4.72972973
- Average Clarity Score: 4.243243243
- Average Correctness score: 4.2567568
- Average Distractor quality score: 4.216216216
- Average Cognitive level score: 3.851351351

Microsoft Excel Worksheet

In the current experiment, I found LLM perform better than SLM in all scores

Microsoft Excel Worksheet

Results – Hypothesis Test - Research Question 1

- μ1 = Baseline SLM MCQ Avg score, μ2 = Finetuned SLM MCQ Avg score
- H0 : μ1 = μ2
- H1: μ1 < μ2 (Finetuned SLM performs better tan Baseline SLM)

Sample Size: 74

t-statistic: -6.539400390204156

p-value: 7.354164560212714e-09

Result: The difference is statistically significant.

- **t-statistic = -6.54**: The negative sign indicates that the **mean score of the baseline SLM group is significantly lower** than the finetuned SLM group.
- **p-value = $7.35 \times 10^{-9} < 0.05$**: The difference in overall average scores between the baseline SLM and the finetuned SLM is statistically significant.
- **This strongly suggests that finetuning the SLM led to a significant improvement in the quality of the generated MCQs.**

Results – Hypothesis Test - Research Question 2

- μ1 = Finetuned SLM MCQ Avg score, μ2 = LLM(Gemini Flash-2.5) MCQ Avg score
- H0 : μ1 = μ2
- H1: μ1 ≠ μ2 (Finetuned SLM or LLM can performs better)

Sample Size: 74

t-statistic: -11.108328368133103

p-value: 4.645716456754879e-21

Result: The difference is statistically significant.

- **t-statistic =** -11.11: The negative value shows that the **mean score for SLM-generated MCQs is much lower than that for LLM-generated MCQs**.
- **p-value =** 4.65×10^{-21} **< 0.05**: There is a statistically significant difference between the SLM and LLM MCQ average scores, with LLM-generated MCQs performing significantly better than SLM-generated ones.
- **This indicates that the LLM approach provides a superior quality of generated questions compared to the SLM approach.**

Results – Hypothesis Test - Research Question 3

- μ1 = Finetuned SLM MCQ **correctness** score, μ2 = LLM(Gemini Flash-2.5) MCQ **correctness** score
- H0 : μ1 = μ2
- H1: μ1 ≠ μ2 (Finetuned SLM or LLM can generate more accurate answers for MCQs)

Sample Size: 74

t-statistic: -12.905943602917617

p-value: 7.461259196657257e-26

Result: The difference is statistically significant.

- **t-statistic =** -12.91: The negative value means that the **correctness score for the SLM-generated MCQs is significantly lower** than that for the LLM-generated MCQs.
- **p-value =** 7.46×10^{-26} **< 0.05: The difference in correctness scores between SLM and LLM MCQs is statistically significant**, with **LLM-generated questions being much more correct** than those from the SLM.
- **This indicates that the LLM approach provides more correct MCQ answers for generated questions compared to the SLM approach.**

Results – Hypothesis Test - Research Question 4

- μ1 = Finetuned SLM MCQ **distractor** score, μ2 = LLM(Gemini Flash-2.5) MCQ **distractor** score
- H0 : μ1 = μ2
- H1: μ1 ≠ μ2 (Finetuned SLM or LLM, either can generate better distractors)

Sample Size: 74

t-statistic: -11.108328368133103

p-value: 4.645716456754879e-21

Result: The difference is statistically significant.

- **t-statistic =** -7.29: The negative sign shows that the **distractor quality score for SLM-generated MCQs is significantly lower** than that for LLM-generated MCQs..
- **p-value =** 2.27×10^{-11}**< 0.05: The distractor quality in LLM-generated MCQs is significantly better than that in SLM-generated MCQs.**
- This suggests that the LLM produces more effective distractors, which likely improves the overall question quality and difficulty balance..

Implementation and User Benefit

- **Data Preparation:** Collect and preprocess domain-specific current affairs datasets (e.g., news articles, online current affairs quizzes from the defined source) for fine-tuning the SLM.
- **Model Fine-Tuning:** Adapt a pre-trained Small Language Model (e.g., GPT-2) using supervised fine-tuning on curated current affairs MCQs.
- **MCQ Generation Pipeline**: Build a pipeline that generates questions, options, and correct answers, with quality control mechanisms (e.g., filtering, ranking, or post-processing).
- **Evaluation Framework:** Assess generated MCQs using both automatic metrics (BLEU, ROUGE, BERTScore) and expert human evaluation (relevance, correctness, distractor quality, cognitive level).
- **Deployment**: Integrate the system into an educational or training platform where educators can generate and review MCQs in real-time.

Conclusion - Automated MCQ Generation using SLM on Current Affairs

- **In the current context and as per the hypothesis test LLM performed better than Domain specialized SLM**
- Fine tuned in 200MCQs, Domain-specialized SLMs shows promising alternative to LLMs in MCQ generation for current affairs, and it can outperform general purpose LLM if fine tuned with large dataset and fed more temporal information.
- Data collection, Training/Retraining, Fine-tuning pipeline is very effective with better controllability and deployment

Limitations:

- **Availability of large MCQ dataset for fine tuning**
- **Restricted context length 1024 for the SLM**
- **Availability of large GPU based compute resource for quick fine tuning the SLM**

Future work:

- **Experimenting with larger context length**
- **Experimenting with large(10k) finetune dataset**
- **Experimenting with modern leading baseline SLM such as** TinyLlama , Microsoft Phi Series , Mistral Small 3.1

Bibliography

1. Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., & Suleman, K. (2017, August). *NewsQA: A machine comprehension dataset.* In P. Blunsom, A. Bordes, K. M. Cho, S. Cohen, C. Dyer, E. Grefenstette, K. M. Hermann, L. Rimell, J. Weston, & S. Yih (Eds.), *Proceedings of the 2nd Workshop on Representation Learning for NLP* (pp. 191–200). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2623

2. Chen, W., Wang, X., & Wang, W. Y. (2021). A dataset for answering time-sensitive questions. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) – Track on Datasets and Benchmarks. NeurIPS. Retrieved from https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/1f0e3dad99908345f7439f8ffabdffc4-Paper-round2.pdf

3. Biancini, G., Ferrato, A., & Limongelli, C. (2024). Multiple-choice question generation using large language models: Methodology and educator insights. In Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization(pp. 584–590). ACM. https://doi.org/10.1145/3631700.3665233

Bibliography

4. Nazarov, B., Frolova, D., Lubarsky, Y., Gaissinski, A., & Kisilev, P. (2025). Rethinking data: Towards better performing domain-specific small language models. In Proceedings of the IEEE Global Communications Conference (GLOBECOM) 2024 Workshop IMMLLM6G. arXiv. https://doi.org/10.48550/arXiv.2503.01464

5. A. M. Olney, “Generating multiple choice questions from a textbook: LLMs match human performance on most metrics,” in Empowering Education with LLMs—The Next-Gen Interface and Content Generation, Workshop Proc., co-located with AIED ’23, Tokyo, Japan, Jul. 2023, pp. 111–128.

6. Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kültür, C., Savelka, J., & Sakr, M. (2024, January 29–February 2). A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. In Proceedings of the Australian Computing Education Conference (ACE 2024), Sydney, NSW, Australia (pp.114–123). ACM. https://doi.org/10.1145/3636243.3636256

Bibliography

7. C. Grévisse, M. A. S. Pavlou, and J. G. Schneider, “Docimologicalquality analysis of LLM-generated multiple choice questions in computer science and medicine,” SN Computer Science, vol.5, no.5, p.636, Jun.2024, doi:10.1007/s42979-024-02963-6

8. Artsi, Y., Sorin, V., Konen, E., Glicksberg, B. S., Nadkarni, G., & Klang, E. (2024). Large language models for generating medical examinations: systematic review.BMC Medical Education, 24(1), 354. https://doi.org/10.1186/s12909-024-05239-y

9. Nwafor, C.A., & Onyenwe, I.E. (2021). An automated multiple-choice question generation using natural language processing techniques. International Journal on Natural Language Computing, 10(2), 1–10. https://doi.org/10.5121/ijnlc.2021.10201

Bibliography

10. Nwafor, C.A., & Onyenwe, I.E. (2021). An automated multiple-choice question generation using natural language processing techniques. International Journal on Natural Language Computing, 10(2), 1–10. https://doi.org/10.5121/ijnlc.2021.10201

11. Robinson, J., Rytting, C.M., & Wingate, D. (2023). Leveraging large language models for multiple choice question answering. In Proceedings of the International Conference on Learning Representations (ICLR 2023). arXiv. https://doi.org/10.48550/arXiv.2210.12353

Appendix - A

- Fine Tuned SLM generated MCQs

Microsoft Excel Worksheet

- LLM (Gemini Flash-2.5) generated MCQs:

Microsoft Excel Worksheet

- Score sheet

Microsoft Excel Worksheet

THANK YOU