Automated MCQ generation using domain specialized SLM on Current Affairs

Name: Abdul Wazed
Mentor: Vivek Kumar Dwivedi

Introduction

This research aims to investigate the comparative language generation capabilities of different language models in the context of MCQs (Multiple Choice Question) - generation based on current affairs (news content). Specifically, it compares:

- **Small Language Models (SLMs)** trained on domain-specific knowledge
- **Large Language Models (LLMs such as Gemini)**

The study also seeks to identify can domain specific SLMs outperform other LLMs on current affairs content which is unseen to Large Language Models.

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 timely 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.

Research Questions

This research aims to explore below four key aspects of small language models (SLMs) in the context of MCQ generation. The questions are mentioned in next four slides

Research Questions #1

Question: 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 DistilGPT2 or TinyLLaMA?

ML Solution: Two parallel SLMs will be developed—one from scratch and the other by fine-tuning a pre-trained model on generic publicly MCQs dataset. These models will be evaluated based on the standard evaluation metrices for the generated MCQs. The findings will guide the selection of the optimal approach for SLM-based MCQ generation..

Research Questions #2

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

Statistical Solution: The most promising SLM identified in Question 1 will be fine-tuned on domain-specific content. Two sets of MCQs will then be generated using the same prompts—one from the fine-tuned SLM and the other from the LLM (e.g., Gemini). The MCQ's qualitative score captured based on standard evaluation metrics from both models will be compared using statistical analysis such as t-test to evaluate performance differences.

Research Questions #3

Question: 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?

Statistical Solution: Two sets of MCQs will be generated—one by the SLM and one by the LLM—using the same input contexts and prompts. The correctness of the answers will be evaluated, and the average accuracy scores will be statistically compared using a two-sample t-test to determine which model performs better in generating accurate answers.

Research Questions #4

Question: 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?

Statistical Solution: Two sets of MCQs will be generated using the same input prompts—one from the fine-tuned SLM and another from the LLM. The distractors from both sets will be evaluated using standardized metrics such as plausibility, relevance, and grammatical consistency. The results will then be statistically compared to determine which model produces higher-quality distractors.

Data Description

a. For this study some or all of Main Feeds category and news archive content to be used

- Times of India RSS Feeds
- The Times of India: Archives

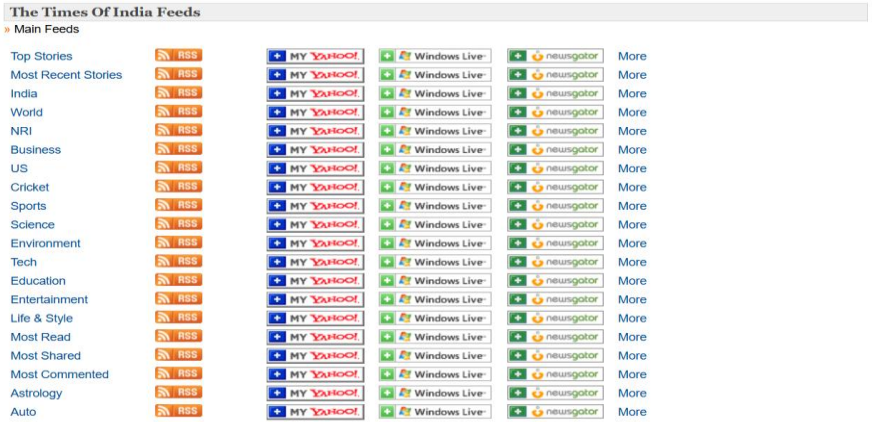


Fig 1: TOI Main Feeds Screenshot

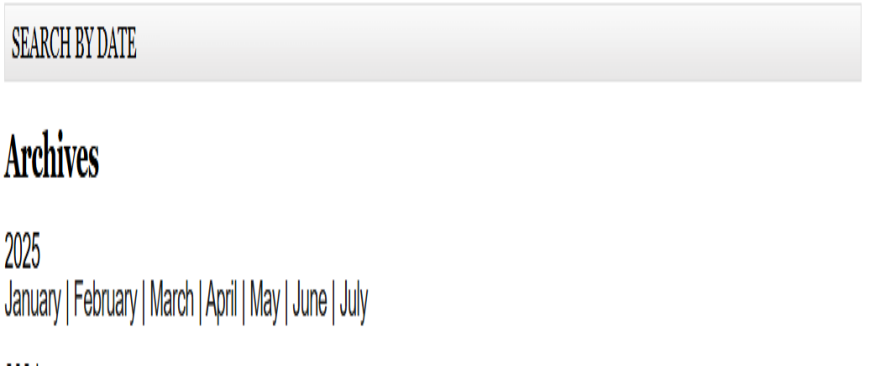


Fig 2: TOI ePaper Archive Screenshot

Data Description

b. Data dictionary mandatory

The **RSS 2.0 (Really Simple Syndication)** feed format is an XML-based schema used for publishing frequently updated content such as news headlines, Here is the schema definition link:

- RSS 2.0 Specification (Current)

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<rss version="2.0">

<channel>
  <title>W3Schools Home Page</title>
  <link>https://www.w3schools.com</link>
  <description>Free web building tutorials</description>
  <item>
    <title>RSS Tutorial</title>
    <link>https://www.w3schools.com/xml/xml_rss.asp</link>
    <description>New RSS tutorial on W3Schools</description>
  </item>
  <item>
    <title>XML Tutorial</title>
    <link>https://www.w3schools.com/xml</link>
    <description>New XML tutorial on W3Schools</description>
  </item>
</channel>

</rss>
```

Fig 3: RSS Document Example (Source: XML RSS)

Analytic approach

This study adopts a comparative analytic framework to evaluate the effectiveness of small language models (SLMs) and Large Language Models (LLMs) in generating multiple-choice questions (MCQs) from current affairs data. The approach consists of the following stages:

Analytic approach – Implementation Stages

a. **Dataset Preparation**: Collect domain-specific text data from curated current affairs sources

b. **Model Training/Fine Tuning**: Fine-tune a SLM model on the prepared domain-specific dataset.

c. **MCQ Generation:** Generate parallel sets of MCQs from SLM and LLM using the same prompts and data

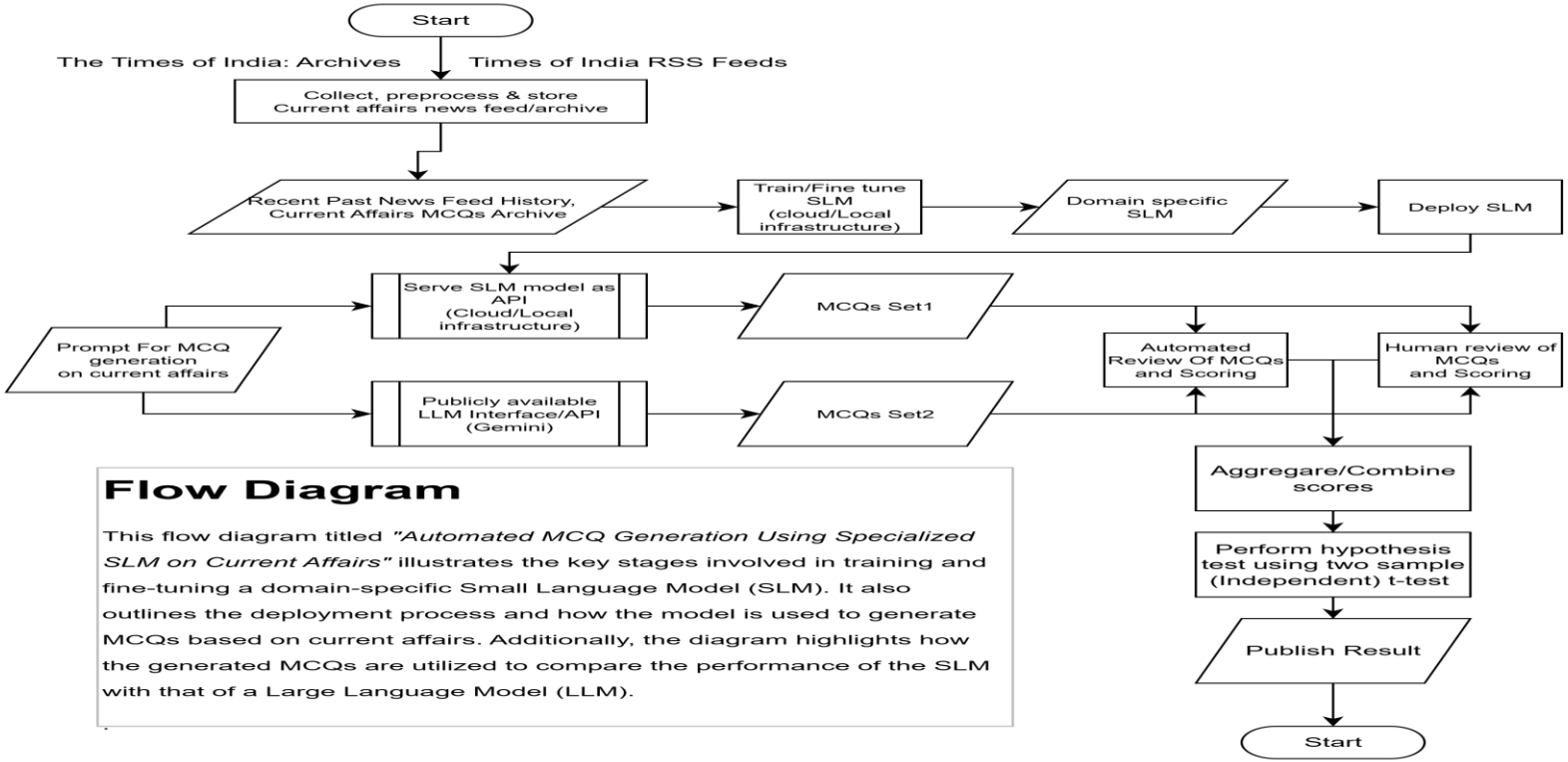
d. **Evaluation Criteria: Use** Automated Metrics & Human Evaluation rubric as evaluation criteria

e. **Statistical Analysis:** Paired t-test to be used to compare MCQ quality between models

f. **Interpretation:** Analyze which model performs better in domain-specific MCQ generation

g. **Recommendations:** Identify gaps and propose techniques (e.g., prompt tuning) to improve SLM output.

Analytic approach – Implementation Stages Flow



Flow Diagram

This flow diagram titled *"Automated MCQ Generation Using Specialized SLM on Current Affairs"* illustrates the key stages involved in training and fine-tuning a domain-specific Small Language Model (SLM). It also outlines the deployment process and how the model is used to generate MCQs based on current affairs. Additionally, the diagram highlights how the generated MCQs are utilized to compare the performance of the SLM with that of a Large Language Model (LLM).

Analytic Approach - Metrics

To measure the quality of MCQs generated by a micro language model (SLM), we need a multi Automated Evaluation Metrics-dimensional evaluation framework combining both **automated metrics** and **human assessment**.

Here's a concise and structured explanation of **evaluation metrics for MCQ generation**, including **formulas and calculation**

Analytic Approach - Automated Evaluation Metrics

The generated MCQs will be automatically reviewed to assess on following metrics.

- **BLEU (Bilingual Evaluation Understudy) Score:** Measures n-gram overlap between generated and reference questions. (BLEU Score calculation)
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Measures recall of n-gram overlap. (ROUGE Score Calculation)
- **BERTScore**: Uses contextual embeddings (from BERT) to compare semantic similarity. (BERTScore Calculation)

Analytic Approach - Human Evaluation Rubric

The generated MCQs will be independently reviewed by subject matter experts (SMEs) to assess on following metrics.

- **Relevance (1-5)**: Is the question related to the source content?
- **Clarity (1-5)**: Is the question easy to understand?
- **Correctness (1-5)**: Is the correct answer accurate and fact-based?
- **Distractor Quality (1-5)**: Are incorrect options plausible and non-obvious?
- **Cognitive Level (1-5)**: Based on Bloom's taxonomy (e.g., recall, apply)

Recommendation and applications

Educators, curriculum designers, e-learning platform developers, and assessment specialists who focus on generating domain-specific content, especially current affairs-based multiple-choice questions (MCQs).

- Faster MCQ generation tailored to current affairs content
- Cost-effective deployment of SLMs on low-resource environments
- Improved control and transparency over question generation

References and bibliography

1. 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

2. 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

References and bibliography

3. Mucciaccia, S. S., Paixão, T. M., Mutz, F. W., Badue, C. S., de Souza, A. F., & Oliveira-Santos, T. (2025). *Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions*. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 2246–2260). Association for Computational Linguistics.

4. C. N. Hang, C. W. Tan, and P. D. Yu, “MCQGen: A large language model-driven MCQ generator for personalized learning,” *IEEE Access*, vol. 12, pp. 102261–102273, Jun. 2024, doi: 10.1109/ACCESS.2024.3420709.

References and bibliography

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

References and bibliography

7. C. Grévisse, M. A. S. Pavlou, and J. G. Schneider, “Docimological quality 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

References and bibliography

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

10. 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