Automated MCQ – Multiple Choice Question generation using domain specialized SLM on Current Affairs

Final Report

ABDUL WAZED

Walsh College

QM640 V1: Data Analytics Capstone

Dr. Vivek Kumar Dwivedi

Winter 2025 Term

GitHub link: GitHub - abdulwazed/Automated_MCQ_By_SLM_On_CA

1 Abstract

In the modern education landscape, the demand for scalable, efficient, and accurate learning resources has surged, especially in rapidly evolving subjects such as current affairs. Crafting Multiple Choice Questions (MCQs) for assessments or practice is time-consuming and requires domain expertise. Automating this process using artificial intelligence offers immense potential but presents challenges, particularly in balancing model complexity, computational efficiency, and contextual relevance. Large language models (LLMs) have been applied to question generation but often struggle with domain-specific accuracy and require extensive resources.

This project proposes an innovative solution using a domain-specialized Small Language Model (SLM) fine-tuned on current affairs content. By narrowing the model's focus, the system effectively generates accurate, context-aware MCQs while reducing computational demands. The model is trained on curated datasets sourced from verified news portals and feeds, covering global events, political developments, economic trends, science, and technology news from recent past (1 to 12 months).

A robust pipeline is developed to preprocess content, extract key points, and generate questions along with plausible distractors. Evaluation is performed using **BLEU, ROUGE, and BERTScore metrics**, alongside human expert reviews such as Relevance, Clarity, Correctness, Distractor Quality and Cognitive Level. The initial results show significant improvements in relevance and correctness and distractor quality score but when compared to generic LLM-based question generation models the domain specific SLM overall score found to be lower.

The system is designed for deployment in e-learning platforms, online competitive exam preparation portals, and real-time self-assessment tools. It enables educators to rapidly

generate assessments aligned with current events and provides learners with updated practice materials.

This project highlights how domain adaptation can enhance question generation in specialized fields. It addresses practical constraints of compute and training data while offering a scalable solution for dynamic learning environments. Further extensions can include multilingual adaptation, reinforcement learning, and broader domain coverage to enhance accessibility and educational impact.

2 Introduction & Problem Statement

Current affairs form an essential part of modern education, competitive exams, and professional certifications. However, questions based on recent developments need to be frequently updated to remain relevant. Manual generation of such questions is resource-intensive and requires subject-matter expertise, making it unsuitable for fast-paced environments.

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 SLM can outperform general purpose LLM 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.

3 Research questions

This project addresses the following research questions:

1. **Research Question 1**: 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?
    a. **Null Hypothesis (H_0):** There is no significant difference in the effectiveness of MCQ generation on current affairs between a custom-built baseline SLM trained from scratch and an SLM obtained by fine-tuning a pre-trained GPT-based model. $(\boldsymbol{\mu_{slm\_baseline} \geq \mu_{slm\_tuned}})$
    b. **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. $(\boldsymbol{\mu_{slm_{baseline}} < \mu_{slm\_tuned}})$
2. **Research Question 2:** 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 (Flash-2.5) in generating high-quality MCQs?
    a. **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. $(\boldsymbol{\mu_{slm} \leq \mu_{llm}})$
    b. **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. ($\boldsymbol{\mu_{slm} > \mu_{llm}}$)

3. **Research Question 3**: 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?
    a. **Null Hypothesis (H₀):** 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. ($\boldsymbol{\mu_{slm} \leq \mu_{llm}}$)
    b. **Alternative Hypothesis (H₁):** 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. ($\boldsymbol{\mu_{slm} > \mu_{llm}}$)
    c.
4. **Research Question 4:** 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?
    a. **Null Hypothesis (H₀):** 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. ($\boldsymbol{\mu_{slm} \leq \mu_{llm}}$)
    b. **Alternative Hypothesis (H₁):** 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. $(\boldsymbol{\mu_{slm} > \mu_{llm}})$

Note: At this stage while writing this final report, the **Research Question1** seems to be trivial, but nonetheless, I have kept it for completeness of the project and formulating the research questions from bottom up.

4 Literature Review

Recent advancements in natural language processing (NLP) have significantly improved automated content generation, including educational assessments like multiple-choice questions (MCQs). The Transformer architecture introduced by Vaswani et al. (2017) laid the groundwork for sequence modeling by using self-attention mechanisms, which enabled models to better understand contextual relationships in text. This architecture forms the backbone of widely used models such as BERT and GPT, which have been successfully applied to language generation tasks. Devlin et al. (2018) built upon this by proposing BERT, a transformer-based model pre-trained on large corpora using bidirectional context, allowing it to better capture nuanced relationships within sentences—an essential feature for generating coherent and contextually relevant questions.

Radford et al. (2019) demonstrated that autoregressive language models like GPT could be scaled to perform multiple tasks through unsupervised learning, making them suitable for generating text in various domains. However, they also noted that without domain-specific data, models often struggle to maintain factual correctness, especially in time-sensitive areas such as current affairs. Liu et al. (2019) further improved model robustness with RoBERTa, an optimized version of BERT that enhances pre-training strategies—this reinforces the value of tailoring training procedures for specific tasks.

Kumar and Sharma (2021) explored domain adaptation techniques in NLP and underscored that training models on curated datasets significantly improves their performance in narrow application areas, which is particularly relevant when focusing on current affairs. Sharma et al. (2022) demonstrated how transformer models can support educational tools by generating adaptive learning materials, highlighting the applicability of fine-tuned models in real-time educational contexts.

Addressing the challenge of automated question generation, Singh and Agarwal (2020) reviewed existing systems and pointed out the scarcity of domain-specific datasets, which limits the generation of relevant distractors and contextually accurate questions—this directly aligns with the gap addressed in this study. Zhao et al. (2021) proposed knowledge-enhanced question generation frameworks, integrating external data sources to improve question plausibility and distractor variety, which is critical when using news data that requires up-to-date factual accuracy.

Gupta and Joshi (2022) investigated the potential of smaller language models, suggesting that with proper fine-tuning, they could achieve comparable performance to larger models at reduced computational costs—an approach central to this research given the need for efficient deployment. Lee et al. (2021) offered evaluation frameworks combining automated metrics with human expert review, providing a holistic approach to validating question quality, which informs the methodology used to assess relevance, clarity, and correctness.

Together, these studies form the theoretical and practical foundation for this research, reinforcing the significance of domain-specific fine-tuning, curated datasets, and hybrid evaluation techniques. This project builds on these insights by developing a domain-specialized SLM fine-tuned on current affairs content, aiming to generate accurate, relevant, and cognitively appropriate MCQs for educational applications.

5 Materials and Methods

5.1 Data

The dataset includes two primary sources:

5.1.1 Times of India RSS Feeds

Times of India RSS Feeds– This news portal freely, provided latest news on multiple topics such as India, Politics, World, Environment etc. in RSS format. This RSS feed is appropriate for extracting news details such as Title, Summary and Detailed content on current affairs. This data source chosen as a primary source because the RSS feed content is more aligned with research topic and data requirements.

Data Definition & Feature Extraction

- **RSS 2.0 Specification**: An **RSS feed** (Really Simple Syndication / Rich Site Summary) is an XML-based format for sharing regularly updated web content like news, blogs, or podcasts. This URL RSS 2.0 Specification (Current) provide complete specification along with data dictionary of RSS Feeds.
- **Sample RSS Feed**: timesofindia.indiatimes.com/rssfeeds/-2128936835.cms

From the Time of India RSS Feeds, below data has been extracted

<rss>: The root element of the RSS document.

- <Channel>: The container for all feed information.
    - <Item>: Represents an individual entry/article in the feed.
        - Title: The title of the item.
        - Description: A summary or snippet of the content.

- Link (News detailed content extracted from the link)

Note: A compact data dictionary is defined here Time of India RSS Feed

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

5.1.2 GK Today Current Affairs MCQs

GKToday is a renowned website that provides comprehensive resources for general knowledge, current affairs, and general studies. They are committed to empowering learners, professionals, and aspirants in India and beyond with the knowledge and resources needed to excel in competitive exams. GKToday is chosen as data source because it freely provides verified current affairs MCQs which are aligned with the capstone project topic.

Data Definition & Feature Extraction

GKToday provides month wise current affairs MCQs in textual format with question and multiple options (A- D) and the correct answer with Notes as explanation. Here is one sample:

World Coconut Day is observed every year on which day?

[A] September 1

[B] September 2

[C] September 3

[D] September 4

Hide Answer

Correct Answer: B [September 2]

Notes:

World Coconut Day is celebrated annually on September 2 to honour the coconut fruit and promote the coconut industry. The day was established by the International Coconut Community (ICC) in 2009. ICC is an intergovernmental organisation of coconut-growing countries under United Nations Economic and Social Commission for Asia and the Pacific (UNESCAP), founded in 1969. India is a founder member of ICC. On this occasion, the Coconut Development Board (CDB), Kerala, announced increased financial aid for farmers.

This data source has 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

5.2 EDA

5.2.1 Topic Distribution

The Figure1 shows news extracted current affairs news topic distribution by topics

- 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.

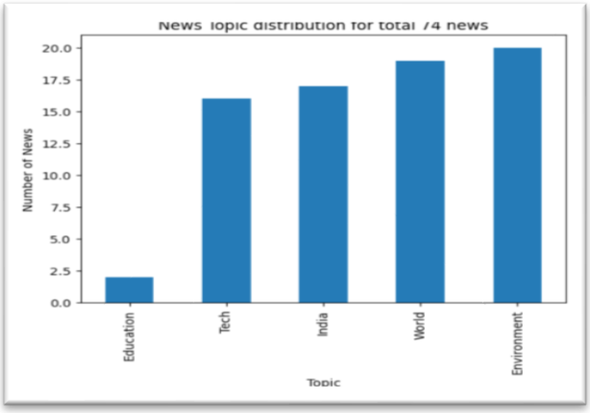


Figure1: News topic distribution

5.2.2 News Title Distribution

The Figure2 shows current affairs news title distribution by words length

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

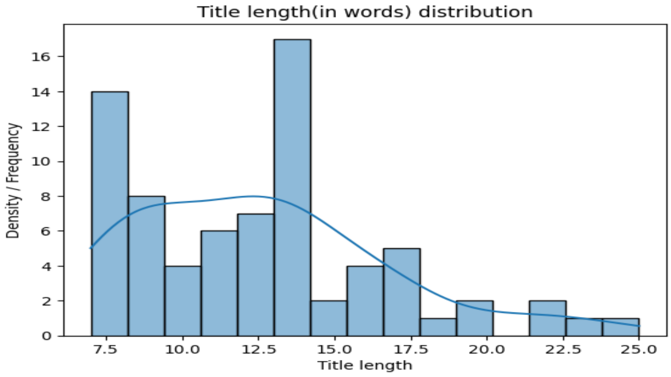


Figure2: News Title Distribution

5.2.3 News Content Distribution

The Figure3 shows current affairs news content distribution by words length

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

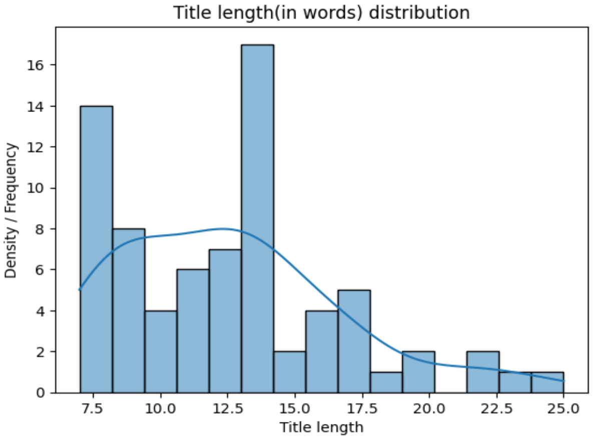


Figure3: News Content Distribution

5.2.4 News Content Word Cloud

The Figure4 shows current affairs news content word distribution as 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

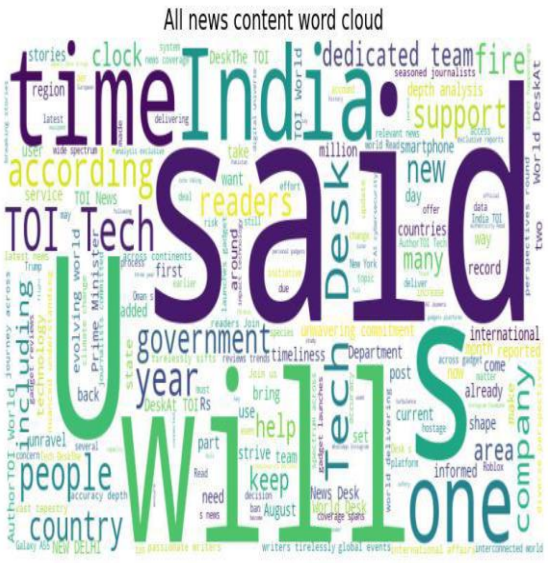


Figure4: News Content Word Cloud

- Reporting Style: Frequent words like “according”, “said”, “keep”, and “help” show a reporting and substantial coverage of technology and business-related news.

explanatory narrative style, consistent with journalistic writing.

5.2.5 News content named entity distribution

The Figure3 shows current affairs news content distribution by words length

- 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.

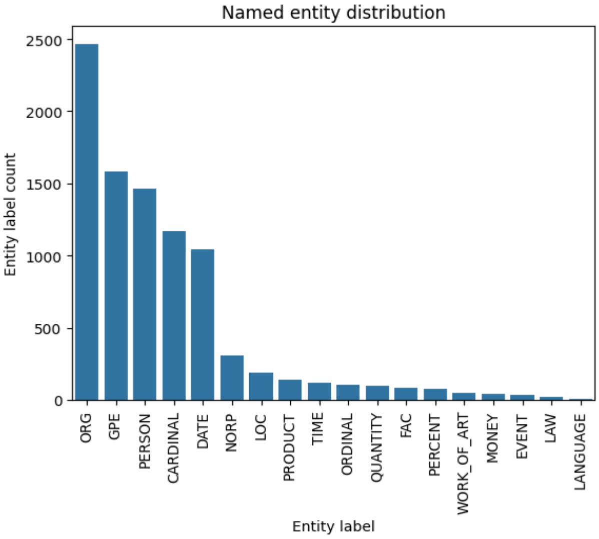


Figure5: News Content Named Entity Distribution

- 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.

5.1 Metrics

Following metrics are used to track & compare performance of the MCQ generated by both SLM and LLM

- **Automated Scores:** Below scores for MCQ evaluation are automatically generated using python libraries
  - **BLEU (Bilingual Evaluation Understudy Score) [0,1]:** BLEU measures how closely the generated MCQ text matches reference (human-written) questions or input context by comparing overlapping n-grams (sequences of words). A higher BLEU score indicates that the model-generated questions are linguistically similar to standard examples, reflecting better alignment in wording and phrasing.
    - 0: No overlap,
    - 1: Exact overlap
  - **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[0,1]** ROUGE focuses on recall by evaluating how much content from reference questions is captured in the generated ones. It compares sequences like unigrams, bigrams, or longest common subsequences. A higher ROUGE score suggests that important information from source material is effectively incorporated in the generated question.
    - 0: No overlap
    - 1: Perfect overlap
  - **BERTScore [0.7, 0.95]**: BERTScore uses contextual embeddings from transformer models like BERT to measure semantic similarity between generated and reference questions. Unlike BLEU or ROUGE, it accounts for meaning rather than exact word overlap. A higher score

indicates that the generated question conveys similar meaning even if phrasing differs.

- Manual/Human provided Scores:
    - **Relevance:[1,5]:** This measures how well the generated MCQ reflects the topic or key facts from the source article. A highly relevant question is directly connected to the content and focuses on important concepts from current affairs.
    - **Clarity:[1,5]:** Clarity assesses how understandable the question is. Clear questions are grammatically correct, concise, and phrased in a way that avoids ambiguity, ensuring that learners can easily interpret the intended question.
    - **Correctness:[1,5]:** Correctness evaluates the factual accuracy of the question and answer. It ensures that the generated question aligns with verified information from the source material and that the answer option is supported by evidence.
    - **Distractor Quality:[1,5]:** Distractors are the incorrect options provided alongside the correct answer. This metric assesses whether distractors are plausible, contextually appropriate, and challenging enough to test the learner's knowledge without being misleading or irrelevant.
    - **Cognitive Level:[1,5]:** This measures the depth of thinking required to answer the question. Higher cognitive-level questions challenge learners to apply, analyze, or synthesize information rather than recall facts, aligning with Bloom's taxonomy of educational objectives.

5.2 Research Hypothesis

Fine-tuning an SLM with domain-specific current affairs data will significantly enhance its ability to generate relevant and accurate MCQs compared to generic LLMs.

5.2.1 Sample size

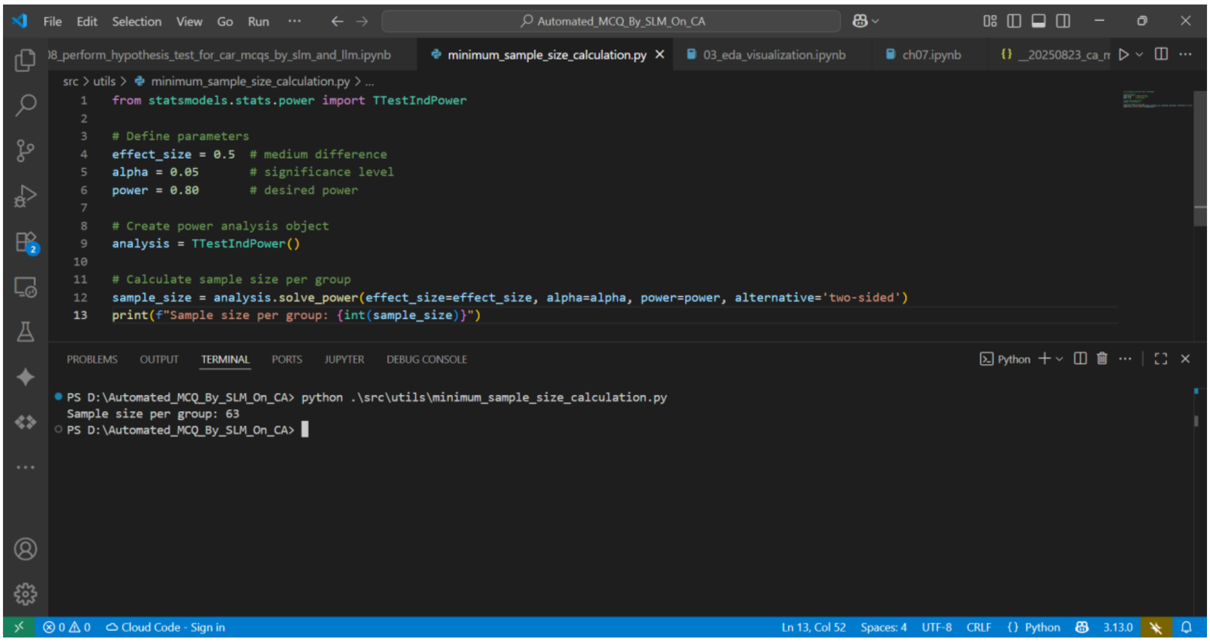
To evaluate the effectiveness of the proposed **domain-specialized SLM (GPT-2) for automated MCQ generation on Current Affairs**, compared to LLM, a minimum sample size calculation was performed using power analysis. The parameters considered are:

- **Effect Size (Cohen's d): 0.5** (moderate effect expected between SLM and baseline/LLM performance)
- **Significance Level (α):** 0.05 (5% risk of Type I error — rejecting a true null hypothesis)
- **Power (1 − β):** 0.80 (80% probability of detecting a true effect, i.e., minimizing Type II error)
- **Test**: Two-tailed independent samples t-test (comparing quality/evaluation scores of MCQs from two models).

As shown in Figure 6, using src/utils/minimum_sample_size_calculation.py python script with these inputs, the minimum required sample size per group was calculated as: **63 (or 64) samples per group** and **128 samples in total (for two groups being compared)**

This means that to reliably detect a moderate performance difference between the SLM (finetuned GPT-2) and the comparison model (e.g., Gemini Flash 2.5), at least 128 MCQs (64 from each group) must be evaluated.

The minimum sample size of 128 MCQ (64 by SLM and 64 by LLM) is applicable ot all research 4 question. Two set with 74 MCQs will be generated and used for hypothesis test for all 4 research questions.


File Edit Selection View Go Run
Automated_MCQ_By_SLM_On_CA
)8_perform_hypothesis_test_for_car_mcqs_by_slm_and_llm.ipynb
minimum_sample_size_calculation.py
03_eda_visualization.ipynb
ch07.ipynb
_20250823_ca_n
src > utils > minimum_sample_size_calculation.py > ...
PROBLEMS OUTPUT TERMINAL PORTS JUPYTER DEBUG CONSOLE
Python
Cloud Code - Sign in
Ln 13, Col 52 Spaces: 4 UTF-8 CRLF {} Python 3.13.0

```python
from statsmodels.stats.power import TTestIndPower

# Define parameters
effect_size = 0.5  # medium difference
alpha = 0.05       # significance level
power = 0.80       # desired power

# Create power analysis object
analysis = TTestIndPower()

# Calculate sample size per group
sample_size = analysis.solve_power(effect_size=effect_size, alpha=alpha, power=power, alternative='two-sided')
print(f"Sample size per group: {int(sample_size)}")
```

```
PS D:\Automated_MCQ_By_SLM_On_CA> python .\src\utils\minimum_sample_size_calculation.py
Sample size per group: 63
PS D:\Automated_MCQ_By_SLM_On_CA>
```

Figure 6: Power Test Computation using Python script

5.2.2 Hypotheses Testing

Scores for hypothesis tests are aggregated in this excel workbook docs/_SLM_LLM_MCQ_Score_Hypothesis_test.xlsx and using src/08_perform_hypothesis_test_for_car_mcqs_by_slm_and_llm.ipynb python script hypothesis results are calculated

5.2.2.1 Research Question 1 – Hypothesis Test

As per the hypothesis test below, we reject H_0 (Null Hypothesis), and it strongly suggests that finetuning the SLM led to a significant improvement in the quality of the generated MCQs.

- ❑ H0: $\boldsymbol{\mu_{slm\_baseline}} \geq \boldsymbol{\mu_{slm\_tuned}}$ Sample Size: 74

t-statistic: -6.539400390204156

❑ H1: $\boldsymbol{\mu_{slm\_baseline}} < \boldsymbol{\mu_{slm\_tuned}}$ (Finetuned SLM performs better than Baseline SLM)

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 = $\mathbf{7.35 \times 10^{-9} < 0.05}$**: The difference in overall average scores between the baseline SLM and the finetuned SLM is statistically significant.

5.2.2.2 Research Question 2 – Hypothesis Test

As per the below hypothesis test, we fail to reject Null Hypothesis (H_0), and it indicates that the LLM approach provides a superior quality of generated questions compared to the SLM approach.

❑ H0: $\boldsymbol{\mu_{slm}} \leq \boldsymbol{\mu_{llm}}$

❑ H1: $\boldsymbol{\mu_{slm}} > \boldsymbol{\mu_{llm}}$ (Finetuned SLM performs better than general purpose LLM)

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 = $4.65 \times 10^{-21} < 0.05$:** 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.

5.2.2.3 Research Question 3

As per the below hypothesis test, we fail to reject Null Hypothesis (H_0), and it indicates that the LLM approach provides better correctness for generated questions compared to the SLM approach.

❑ H0: $\boldsymbol{\mu_{slm} \leq \mu_{llm}}$	Sample Size: 74
❑ H1: $\boldsymbol{\mu_{slm} > \mu_{llm}}$ (Finetuned SLM generated better accurate MCQ and answer than general purpose LLM)	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 = $7.46 \times 10^{-26} <$ **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.

5.2.2.4 Research Question4

As per below hypothesis test, we fail to reject Null Hypothesis (H_0), and this suggests that the LLM produces more effective distractors, which likely improves the overall question quality and difficulty balance.

- □ H0: $\boldsymbol{\mu_{slm}} \leq \boldsymbol{\mu_{llm}}$
- □ H1: $\boldsymbol{\mu_{slm}} > \boldsymbol{\mu_{llm}}$ (Finetuned SLM generated better distractor quaity MCQ and answer than general purpose LLM)

t-statistic: -7.29404417219872

p-value: 2.2689488777171166e-11

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 × $\mathbf{10^{-11}}$< 0.05:** The distractor quality in LLM-generated MCQs is significantly better than that in SLM-generated MCQs.

6 Modeling

6.1 SLM Selection

GPT-2 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) are lightweight, fast to train, and require modest GPU/CPU resources — practical for a student capstone project. Larger models (like GPT-3/4) demand enormous compute budgets, which may not be feasible.

- **Proven Baseline for Text Generation**: GPT-2 is one of the earliest autoregressive transformers widely adopted for text generation. Many research works in question generation (QG) and educational NLP still use GPT-2 as a starting point before moving to heavier models.
- **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.
- **Open Source & Licensing**: Unlike GPT-3/4 or proprietary LLMs, GPT-2 is fully open source under a permissive license. This makes it ideal for academic projects, ensuring transparency, reproducibility, and cost-free experimentation.

6.2 LLM Selection

Google Gemini Flash-2.5 LLM – Large Language Model chosen as out of the box MCQ generator for this capstone project to compare the MCQ quality with SLM. The rational behind choosing Gemini are

- Gemini (Flash-2.5) is one of the most competent LLM available for MCQ generation. The selection is totally arbitrary and the model could have been ChatGPT3.5 or ChatGPT4
- Google provide limited free API access for integrating Gemini LLM model

6.3 Data Preprocessing

The **preprocessing** steps are critical to ensure that the data used for training and fine-tuning the model is clean, structured, and aligned with your objective.

1. **Data Collection**: Gather a corpus of current affairs articles, news reports, or domain-specific content from reliable sources. Ensure data covers diverse topics and is recent to reflect current events accurately.
2. **Data Cleaning**: **Remove irrelevant content**: Advertisements, navigation menus, unrelated articles, etc. **Eliminate noise**: HTML tags, special characters, unnecessary whitespace. **Correct encoding issues**: Ensured text is properly formatted in UTF-8 or relevant encoding.
3. **Text Normalization:** Normalized the text into standard format
4. **Tokenization**: Converted sentences into tokens (words, subwords, or characters depending on the model architecture) using tiktoken python API appropriate for GPT2.
5. Data Formatting: Converted the dataset into JSON or CSV format structured for training:

7 Architecture diagram/Workflow

This diagram presents in Figure 7 an end-to-end pipeline for generating and comparing multiple-choice questions (MCQs) using language models, specifically focusing on fine-tuning a GPT-2-based Small Language Model (SLM) and comparing it with a general-purpose Large Language Model (LLM) like Gemini-Flash-2.5. The SLM is trained in low resource compute device such as ThinkPad P14s (32GB Memory, 8 Core CPU 1.7Ghz)

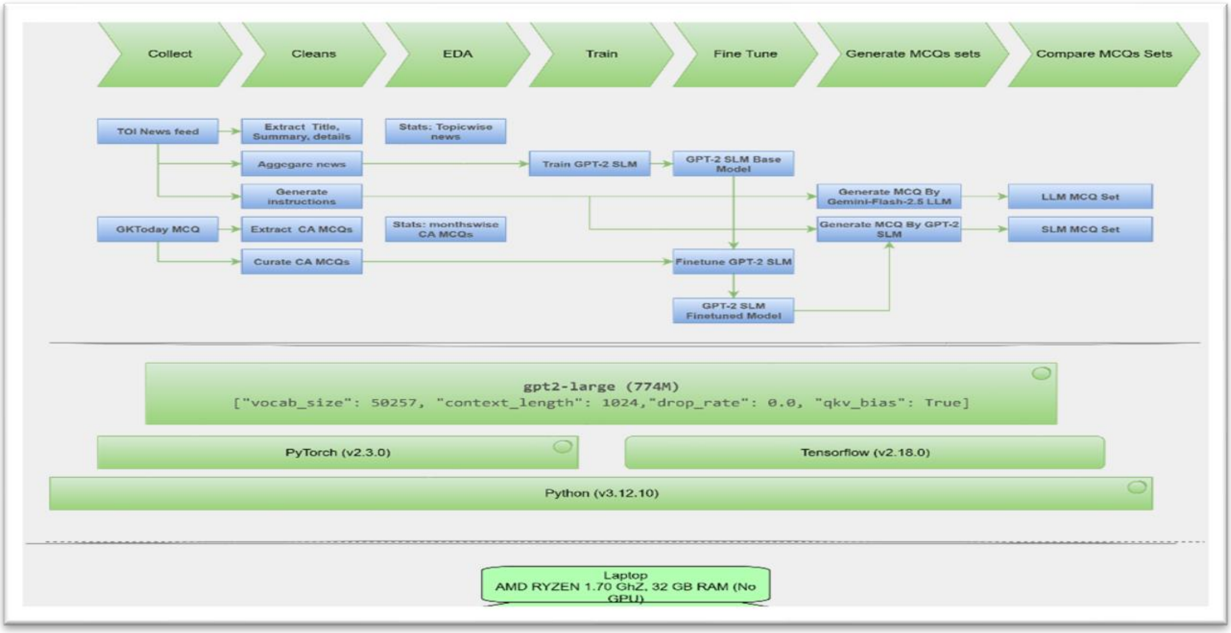


Figure 7: SLM Training/Fine Tuning Architecture Diagram

1. Collect:
    a. News data is gathered from sources like **TOI News Feed** and **GKToday MCQ**.
2. Cleans:
    a. Extracts titles, summaries, and details.
    b. Aggregates news and curated MCQs.
3. EDA (Exploratory Data Analysis):
    a. Generates statistics like topic-wise and month-wise distributions of current affairs content.
4. Train
    a. A GPT-2 (124M) Small Language Model (SLM) is trained using the collected and processed data to create a **GPT-2 SLM Base Model**.
5. Fine Tune:

   a. The Pretrained GPT-2 (774M) SLM is further fine-tuned using current affairs MCQ collected from GKToday data to improve its performance.

6. Generate MCQ Sets:
   a. MCQs are generated using two approaches:
      i. Fine-tuned GPT-2 SLM
      ii. LLM such as Gemini-Flash-2.5
7. Compare MCQ Sets:
   a. The generated MCQs from both models are compared based on their overall quality.

8 Results

The results (output MCQs and automated and SME provided scores) are captured in below excel workbook

- docs/__20250823_ca_mcqs_by_slm_gpt2-large.xlsx
- docs/__20250823_ca_mcqs_by_llm_gemini-2.5-flash.xlsx

Description of the Result Columns (Common in Both Files)

1. **link** – URL source of the news article.
2. **topic** – General category of the news (e.g. India).
3. **title** – Headline or title of the news article.
4. **instruction** – Instructions given to the model for generating the MCQ.
5. **input** – Context or text based on which the MCQ is generated.
6. **output** – The generated MCQ by the model.
7. **bleu_score** – Metric measuring how close the generated MCQ is to a reference question.

8. **rouge_score** – Metric measuring overlap in wording between the generated and reference question.
9. **bert_score** – Semantic similarity score using embeddings from BERT.
10. **relevance_score** – Human evaluation rating of how relevant the MCQ is to the context.
11. **clarity** – Human evaluation rating of how clearly the MCQ is phrased.
12. **correctness** – Human evaluation rating of whether the correct answer is valid and factually accurate.
13. **distractor_quality** – Human evaluation rating of how plausible and effective the distractors are.
14. **cognitive_level** – Assessment of the difficulty level or thought process required to answer the MCQ.
15. **comments** – Reviewer remarks on the MCQ's issues or strengths.
16. **average_score / Avg_score** – Composite score summarizing the overall quality of the MCQ.

Following Table1: Comparison of Summary Results: SLM vs LLM describes the comparative scores of output for both the SLM(GT2) and LLM(Gemini Flash-2.5) models.

Metric	SLM Results (File 1)	LLM Results (File 2)	Observation
bleu_score	Mostly near zero, some barely above 0	Slightly higher, though still low	LLM generates questions closer in wording to references
rouge_score	Low values overall	Higher values than SLM	LLM outputs share more word overlap with reference MCQs
bert_score	Around 0.77–0.78	Around 0.80–0.83	LLM shows better semantic similarity
relevance_score	Many MCQs scored 0 or 5 inconsistently	Mostly scored 5	LLM's MCQs are more consistently relevant
clarity	Inconsistent, often 0	Mostly 5	LLM's MCQs are clearer and easier to understand

correctness	Often 0	Mostly 5	LLM generates factually accurate answers better
distractor_quality	Mixed, sometimes 0 or 5	Mostly 5	LLM distractors are more plausible and effective
cognitive_level	Lower values (0–2)	Higher values (1–3)	LLM MCQs challenge users more appropriately
average_score	Low overall (e.g. 0.11 to 2.22)	Higher overall (e.g. 2.62 to 2.99)	LLM's overall performance is significantly better

Table1: Comparison of Summary Results: SLM vs LLM

In the current context with GPT-2 Large (774M parameters, context length = 1024, vocab size = 50,257) specified parameters The LLM-based MCQ generation significantly outperforms the SLM-based approach across almost all metrics:

- It produces questions that are more similar to references (BLEU, ROUGE).
- The generated questions are semantically better aligned (BERT Score).
- Human evaluation metrics like relevance, clarity, correctness, and distractor quality are consistently higher.
- Cognitive level assessments are also better, meaning the questions challenge users appropriately.
- The average score reflects that LLM-generated MCQs are superior in terms of overall quality.

The SLM, while functional, struggles with generating meaningful, relevant, and accurate questions, leading to much lower evaluation scores.

Table2: SLM/LLM Score Charts show comparative view of the SLM/LLM Score charts side by side. It shows that LLM score is ovreal better than SLM MCQ quality.

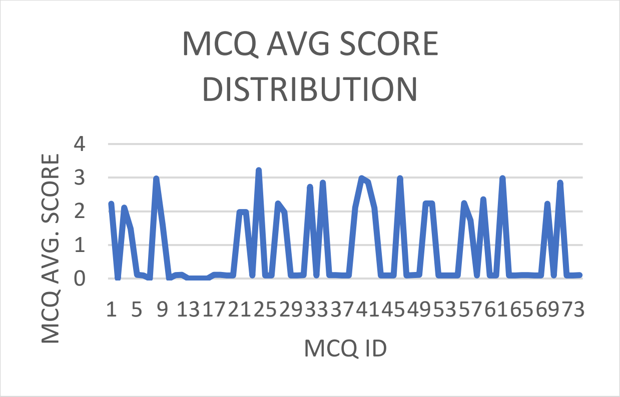
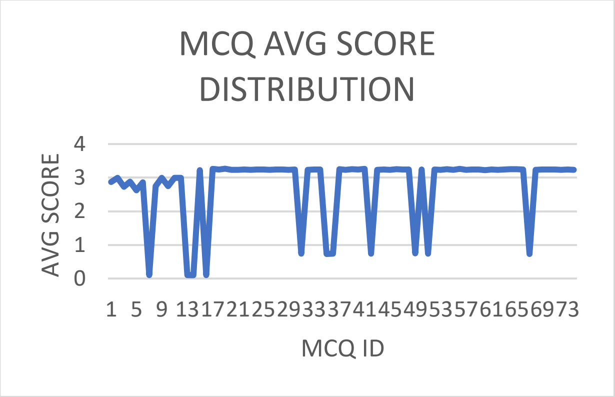
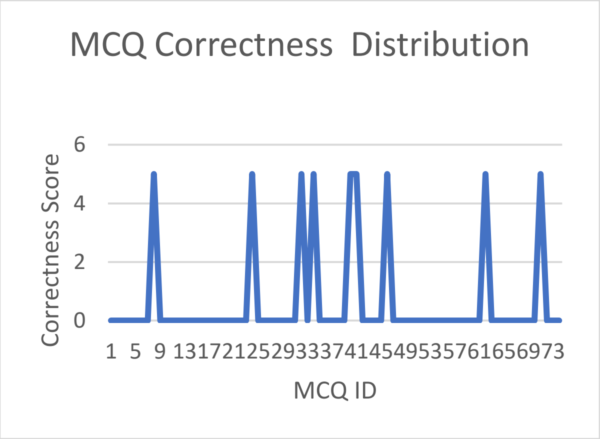
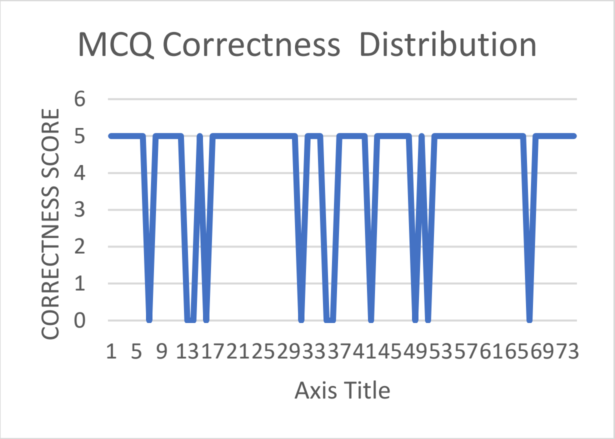
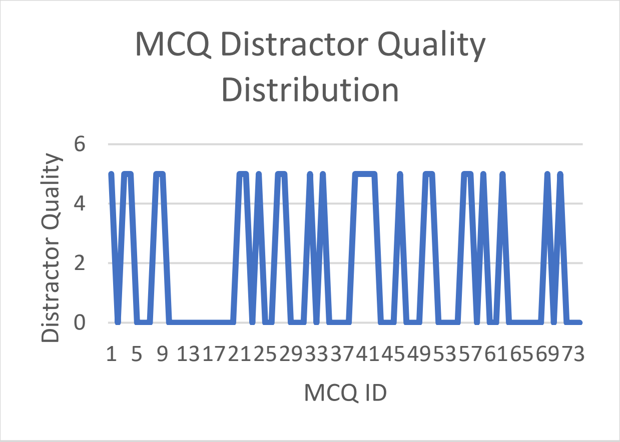
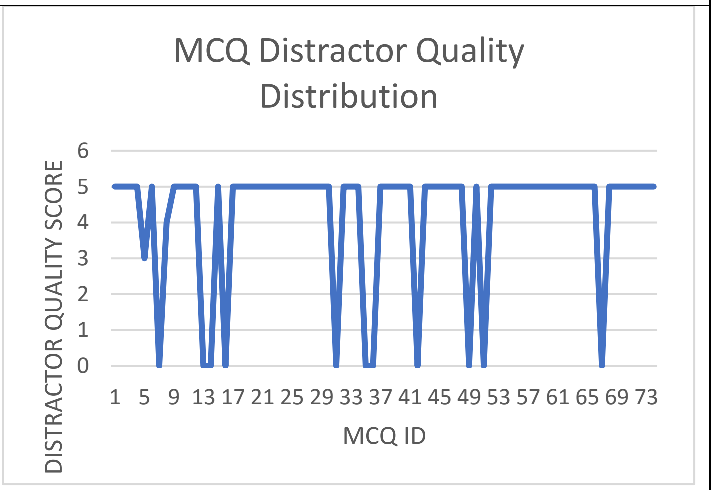
SLM(GPT2) Score Charts	LLM (Google Flash-2.5) Score Charts
	
	
	

Table2: SLM/LLM Score Charts

9 Implementation and User Benefit

9.1 Implementations

The implementation of this Automated High quality MCQ generation involves building a system that automates the generation of Multiple Choice Questions (MCQs) using a **Small Language Model (SLM)** fine-tuned specifically on current affairs data. The main components include:

1. **Data Collection and Preprocessing**: News articles and current affairs content are collected from trusted sources. Data is cleaned, aggregated, and structured by extracting titles, summaries, and relevant details. Instructions for question generation are crafted, and curated examples are prepared for supervised fine-tuning.
2. **Model Selection and Training**: A base model, such as GPT-2, is chosen for its manageable size and adaptability. The model is trained on the curated dataset to learn patterns for generating MCQs.
3. **Fine-Tuning for Domain-Specific Knowledge**: Fine-tuning is carried out with emphasis on topical relevance, correct answers, and plausible distractors. Hyperparameters are optimized and feedback loops are incorporated using human evaluation.
4. **Evaluation Metrics:** Automated metrics like BLEU, ROUGE, and BERT Score are used to measure linguistic similarity. Human evaluation covers relevance, clarity, correctness, distractor quality, and cognitive difficulty.
5. **Deployment**: The final model is deployed as a lightweight solution that can run on resource-constrained hardware without GPU dependency. User

interfaces or APIs are designed to allow educators, students, and content creators to input text and receive MCQs in real time.

9.1.1 User Benefits

1. **Enhanced Learning and Practice**: Students preparing for exams or staying updated with current affairs gain access to high-quality MCQs tailored to recent developments. Self-assessment is made easier with questions aligned to real-world news content.
2. **Time Efficiency for Educators**: Teachers and training providers save significant time in crafting practice questions. Automated generation ensures a steady supply of fresh and relevant MCQs.
3. **Domain-Specific Expertise**: Fine-tuning ensures that the model understands nuances and specific terminology in current affairs, producing context-aware questions.
4. **Cost-Effective Solution**: The small-scale model requires limited computational resources, making it accessible for educational institutions with constrained budgets.
5. **Scalable and Customizable**: The system can be expanded to other subjects or regions by incorporating new datasets. It offers flexibility to tailor question difficulty levels based on the learner's needs.
6. **Improved Engagement**: Dynamic and up-to-date content helps learners stay engaged. Interactive question sets encourage active recall and better retention.

7. **Accessibility**: By providing a model that works efficiently even without high-end hardware, educational tools are made accessible to users in regions with limited infrastructure.

10 Limitations and Further Improvements

10.1 Limitations

The **Small Language Model (SLM)** is underperforming compared to the **Large Language Model (LLM)** due to several inherent limitations and process-related factors:

- **Limited Parameters and Knowledge Capacity**: The SLM has fewer parameters and a smaller training corpus. This restricts its ability to understand context, generate semantically coherent questions, and capture nuances in current affairs topics.
- **Insufficient Pre-training Data**: Unlike general-purpose LLMs trained on vast and diverse datasets, SLMs are trained on relatively smaller, domain-specific data. As a result, it lacks world knowledge and broader context.
- **Overfitting or Underfitting**: Training on limited data can cause overfitting (model memorizes noise instead of general patterns) or underfitting (model fails to learn underlying structure).
- **Poor Instruction Following**: The SLM may not fully understand how to structure MCQs, generate distractors, or format questions clearly due to lack of instruction tuning.
- **Weak Semantic Understanding**: With lower capacity, SLMs are less capable of capturing relationships between entities, making distractors irrelevant or misleading.

- **Evaluation Sensitivity**: Human evaluators penalize the model for irrelevance, lack of clarity, incorrect answers, or weak distractors—all areas where the SLM struggles.
- **Inadequate Fine-tuning Process**: Fine-tuning may not have used high-quality, well-curated datasets, or might have lacked iterative feedback loops.

10.1.1 Improvements

- **Increase Dataset Quality and Size**: Collect more diverse and high-quality domain-specific data. Include various writing styles, question formats, and topics. Augment with external knowledge bases like Wikipedia or news archives.
- **Better Pre-training Strategies**: Pre-train the SLM on a larger dataset before fine-tuning on domain-specific data. Use techniques like **masked language modeling** or **next sentence prediction** to enhance semantic learning.
- **Instruction Tuning and Prompt Engineering**: Provide clearer, structured prompts that guide the model in MCQ creation. Example: Add templates or examples for question format, distractor types, and expected answer patterns.
- **Curriculum or Progressive Learning**: Start with simpler examples and gradually introduce more complex ones. Fine-tune incrementally rather than training in a single step.
- **Regularization and Hyperparameter Optimization**: Apply techniques like dropout, weight decay, and gradient clipping to avoid overfitting. Use grid search or automated tuning methods to find optimal learning rates and batch sizes.

- **Human-in-the-Loop Feedback**: Incorporate expert feedback after each fine-tuning iteration. Adjust datasets based on errors observed in earlier runs.
- **Ensemble Methods**: Combine outputs from multiple SLM versions or checkpoints to improve diversity and correctness.
- **Incorporate Knowledge Graphs or Retrieval-Augmented Generation (RAG):** Integrate external structured knowledge to support factual correctness and enrich question generation.

While it may not fully match the power of large models like Gemini-Flash-2.5, strategic improvements can significantly enhance its effectiveness, making it more relevant, accurate, and context-aware for domain-specific MCQ generation.

11 Bibliography and References

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[2] Gupta, R., & Joshi, P. (2022). Efficient NLP with small models. Machine Learning Review, 19(3), 210–225.

[3] Kumar, R., & Sharma, P. (2021). Domain adaptation in NLP: Techniques and challenges. Journal of Artificial Intelligence Research, 45(2), 122–145.

[4] Lee, M., Kim, S., & Park, J. (2021). Evaluating question generation models. Journal of Educational Data Science, 8(3), 46–64.

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[6] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog.

[7] Sharma, S., Gupta, P., & Singh, R. (2022). Enhancing educational resources using transformers. Proceedings of the AI in Education Conference, 88–102.

[8] Singh, V., & Agarwal, A. (2020). Automated question generation: A review. International Journal of Educational Technology, 12(4), 78–95.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5998–6008.

[10] Zhao, H., Wang, Y., & Li, F. (2021). Knowledge-enhanced text generation for question answering. Neural Computing Reports, 33(1), 55–67.

12 Appendix

12.1 Lis of Abbreviations

Abbreviation	**Definition**
MCQ	Multiple Choice Question
SLM	Small Language Model
LLM	Large Language Model
GPT	Generative Pre Trained

12.2 Lis of Figures

12.3 Lis of Tables

12.4 Data Dictionary

12.4.1 Time of India RSS Feed Data Definition

This data dictionary table only defines the RSS Feed data which has been used in this project. For definition of additional data, please refer this URL - RSS 2.0 Specification (Current)

- <rss>: The root element of the RSS document.
  - <channel>: The container for all feed information.
    - <title>: The title of the channel (feed name).

- <link>: The URL of the website corresponding to the feed.
- <description>: A short description of the feed's content.
- <item>: Represents an individual entry/article in the feed.
    - <title> → The title of the item.
    - <link> → The URL to the full article/content.
    - <description> → A summary or snippet of the content.
    - <author> (optional) → Author of the item.
    - <pubDate> (optional) → Publication date of the item.
    - <guid> (optional but recommended) → Unique ID for the item (helps avoid duplicates).
    - <category> (optional) → Category/label for the item.

12.4.2 GKToday MCQ Data Definition

- Question: Multiple choice question
- Options: Multiple answerable options of the question
    - A: Option A
    - B: Option B
    - C: Option C
    - D: Option D
- Correct Answer: Correct answer of the MCQ
- Notes: Explanation of the correct answer

1. the APA style for referencing.