Automated MCQ generation using domain specialized SLM on Current Affairs

Interim Report

ABDUL WAZED

Walsh College

QM640 V1: Data Analytics Capstone

Dr. Vivek Kumar Dwivedi

Winter 2025 Term

GitHub Code Repository: abdulwazed/Automated_MCQ_By_SLM_On_CA

1 Introduction

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.

So specifically, 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 in 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

2 Scope and Objective

This research aims to explore below four key aspects of small language models (SLMs) in the context of MCQ generation.

2.1 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?

2.2 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 than a general-purpose Large Language Model (LLM) such as Gemini in generating high-quality MCQs?

2.3 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?

2.4 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?

3 Literature Survey

Research on Automated Question Generation (AQG) has evolved from rule-based pipelines to neural approaches that model question formation as a conditional generation task. Early systems relied on syntactic transformations, semantic role labeling, and answer selection heuristics to produce wh-questions from declarative text, establishing core steps still reused today: content selection, question formulation, and (for MCQs) distractor generation (Heilman & Smith, 2010).

Neural methods reframed AQG as sequence-to-sequence learning with attention, showing strong gains in fluency and diversity (Du et al., 2017; Zhou et al., 2017). Pretrained encoder–decoder models such as BART and T5 further improved quality through large-scale pretraining and task-specific fine-tuning, enabling controllable question styles and better handling of paraphrase and context (Lewis et al., 2020; Raffel et al., 2020).

MCQ generation adds the challenge of **distractor synthesis**. Classical approaches use semantic similarity, lexical resources, and part-of-speech constraints to create plausible but incorrect options (Agarwal & Mannem, 2011). Recent work uses embedding space proximity, learning-to-rank, or adversarial sampling to balance plausibility with discriminability while avoiding clueing and triviality (Susanti et al., 2018). Surveys consistently note that distractor quality is the main bottleneck for automatic MCQs (Kurdi et al., 2020).

For **domain specialization**, small language models (SLMs) such as GPT-2, DistilBERT, and ALBERT can be fine-tuned on targeted corpora to achieve competitive task performance with modest compute, provided careful data curation and regularization (Radford et al., 2019; Sanh et al., 2019; Lan et al., 2019). Domain adaptation strategies—continued pretraining on in-domain text followed by task fine-tuning—help SLMs internalize domain vocabulary and discourse patterns critical for current-affairs reasoning (Gururangan et al., 2020).

Current Affairs introduces temporal drift and verification concerns. Pipelines commonly harvest articles via RSS/News APIs, normalize them, and maintain date-stamped corpora to support time-bounded question generation and evaluation. Content selection often prioritizes named entities, events, and facts with strong source agreement to reduce hallucinations. Ethical and quality safeguards include de-duplication, bias checks, and provenance tracking.

Evaluation remains multifaceted: automatic metrics (BLEU/ROUGE) correlate weakly with human judgments for AQG, so studies emphasize rubric-based human evaluation (fluency, answerability, relevance, and distractor plausibility), item statistics (difficulty, discrimination), and external validity via learner performance (Soni & Sahu, 2021). When comparing SLMs to

LLMs, recent findings suggest LLMs excel in fluency and coverage, while domain-specialized SLMs can match relevance and controllability at lower cost—especially when prompts/constraints and curated distractor banks are used (Brown et al., 2020; Lewis et al., 2020).

In sum, the state of the art for automated **MCQ generation on time-sensitive news** combines: (i) in-domain pretraining/fine-tuning of compact models, (ii) entity/event-centric content selection, (iii) hybrid distractor generation (neural + constraints), and (iv) human-aligned evaluation. Your capstone's focus—domain-specialized SLMs for current affairs—sits squarely in a practical sweet spot of accuracy, cost, and operational control.

4 Data Description

Following data sources has be used extract data which has been used in training and multiple-choice question generation using gpt2 bases SLM – Small Language Model.

4.1 Times of India RSS Feeds

RSS is a way of providing content to the user's browser or desktop in an efficient way. By using RSS feeds, the user can stay updated on the news from TOI and other news sources with little extra effort.

RSS (Really Simple Syndication) feeds are normally provided in three ways: headlines only,

headlines with excerpts and full text feeds. TOI provides you headlines with excerpts, for free.

TOI grants you permission to only access and make personal use of its RSS feeds and you agree not to, directly or indirectly, download, modify, alter, change, amend, vary, transform, revise, translate, copy, publish, distribute or otherwise disseminate these RSS feeds, or any portion of these — except with the express consent of Times Internet Ltd (TIL). TIL forbids you from any attempts at displaying, hosting, aggregating, reselling or putting to commercial use either directly or indirectly TIL's RSS feeds or any part of the same.

You must not retain any copies of these pages saved to disk or to any other storage medium except for the purposes of using the same for your individual/personal use.

From Times of India RSS Feeds, following main feeds has been used to extract latest news daily and archive locally.

- India
- World
- NRI
- Business
- US
- Cricket
- Sports
- Science
- Environment
- Tech
- Education
- Entertainment

4.1.1 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. It has some **core elements** that define the structure. Here's a breakdown:


RSS_Feed.txt

Detailed definition of RSS Feed schema is available here RSS 2.0 Specification (Current)

4.2 GK TODAY

GK TODAY provides monthly MCQs (multiple choice question) on current affairs which has been used to finetune SLM. Specifically from following sections MCQ has been collected and curated

- Current Affairs Quiz – August 2025: 50 MCQs
- Current Affairs Quiz – July 2025: 50 MCQs
- Current Affairs Quiz – June 2025: 50 MCQs
- Current Affairs Quiz - May 2025: 50 MCQs
- Current Affairs Quiz - April 2025: 50 MCQs

Here is one sample MCQs:

- **Question**: What is the base year for the Reserve Bank of India- Digital Payments Index (RBI-DPI)?
- **Options**:
    - [A] 2016

- [B] 2017
- [C] 2018
- [D] 2019

Correct Answer: C [2018]

Notes/Input context:

Recently, the Reserve Bank of India (RBI) announced that its Digital Payments Index (RBI-DPI) rose to 493.22 in March 2025. This is a sharp increase from 465.33 recorded in September 2024, showing rapid growth in digital transactions. The Digital Payments Index is a tool created by RBI to measure the extent of digitisation of payments across India. It was first launched in January 2021 to track and encourage the adoption of digital payment systems. It is the first-of-its-kind index to map the nationwide spread of digital transactions. The base year for the index is March 2018, with the score for that year set at 100.

5 Analysis

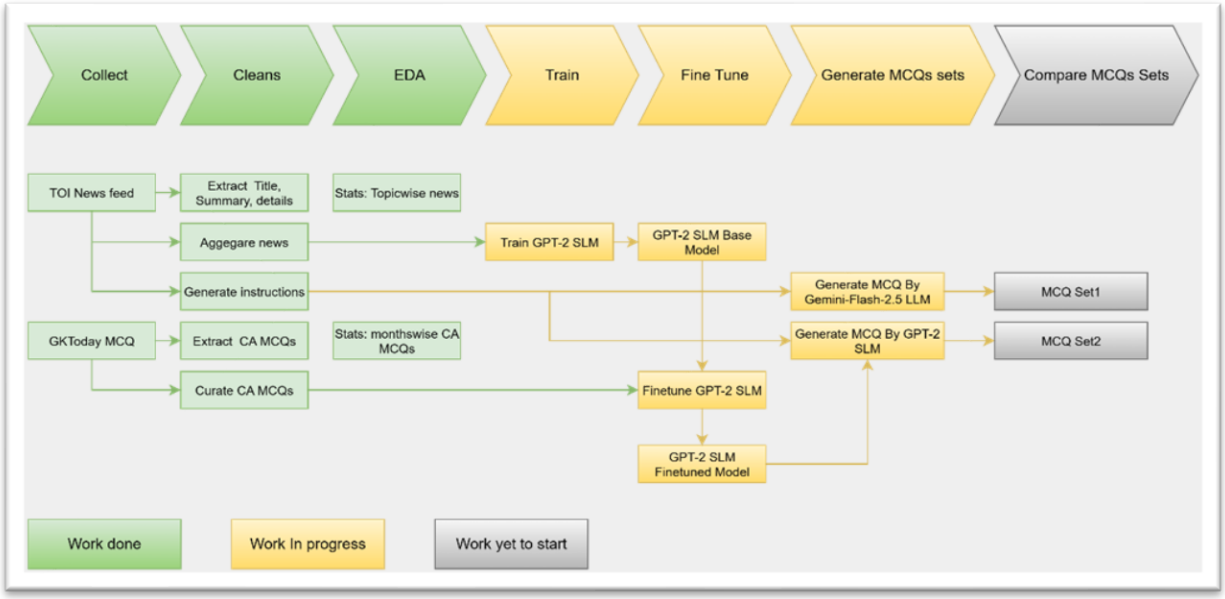


Fig1: Hierarchical Flow Diagram Showing End-to-end flow for Automated MCQ generation using GPT-2 based SLM and Gemini-Flas-2.5 based LLM

5.1 Data cleansing

Data cleansing ensures that the input fed into GPT-2 is **reliable, noise-free, and consistent**, which directly impacts the quality of generated MCQs. Since the project relies on **news articles and curated question datasets**, the cleansing process involves both **textual preprocessing** and **domain-specific filtering**.

- **Removing Noise and Redundancy:** Eliminated HTML tags, special characters, emojis, and non-UTF8 symbols from RSS news feeds. Normalized whitespace, punctuation, and line breaks for clean text input. Remove **boilerplate text** such as “Read more at…” or ad banners embedded in scraped content.

- **Deduplication:** Duplicate articles or repeated headlines are removed to avoid biased training and over-representation of specific events.
- **Sentence-Level Cleaning:** Split text into well-formed sentences for easier candidate question extraction. Removed extremely short (e.g., headlines like *"Breaking News!"*) or very long complex sentences that confuse the model.
- **Language and Domain Filtering:** Retained only English-language content (or the chosen training language). Filtered news to **Current Affairs domains** relevant for MCQs (politics, economy, environment, sports, science).
- **MCQ Dataset Cleansing:** Ensured existing MCQs (used for fine-tuning) follow a consistent format: Question Stem + 4 Options + Correct Answer. Remove incomplete, ambiguous, or factually outdated questions. Normalize labels (e.g., "A", "B", "C", "D") for uniformity.
- **Temporal Relevance:** Since Current Affairs evolve daily, tag articles with dates and discard outdated ones beyond the evaluation scope. Maintain **time-stamped datasets** to track model performance across different news cycles.
-

5.2 EDA results

1. Dataset Size:
    a. Total records: 74
2. Columns Present
    a. link → News article URL

b. topic → Category/domain of news (e.g., India, World, Economy)

c. title → Headline of the article

d. instruction → Prompt instruction for MCQ generation

e. input → News content/context text

f. output → Expected MCQ format (question + options + answer)

3. Missing Values

    a. No missing values across any column (all complete).

4. Content Observations

    a. Each record links to a Times of India news article.

    b. topic is dominated by India-related news (political, policy, security).

    c. input fields are long-form article texts, typically 200–400 words.

    d. output is in a structured MCQ template, though many entries appear as placeholders (blank options/answers).

5. Data Quality

    a. Data is clean and consistent (no nulls, no structural errors).

    b. Titles are concise (~10–15 words), while inputs are longer narrative content.

    c. Outputs require completion/filling during MCQ generation (ideal for fine-tuning GPT-2).

5.3 Minimum sample size computation per RQ

To evaluate the effectiveness of the proposed **domain-specialized SLM (GPT-2) for automated MCQ generation on Current Affairs**, a 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).

Using these inputs, the **minimum required sample size per group** was calculated as:

- 64 samples per group
- **128 samples in total** (for two groups being compared)

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

Below python api has been used to calculate minimum sample size as mentioned above for the given parameters.

minimum_sample_size_calculation.py

The above minimum sample size holds for all 4 research questions

5.3.1 Research Question 1 minimum Sample Size

- Minimum sample (MCQs) size for GPT2-large based SLM: 64

- Minimum sample (MCQs) size for Gamini Flash 2.5 based SLM: 64

5.3.2 Research Question 2 minimum Sample Size

- Minimum sample (MCQs) size for GPT2-large based SLM: 64
- Minimum sample (MCQs) size for Gamini Flash 2.5 based SLM: 64

5.3.3 Research Question 3 minimum Sample Size

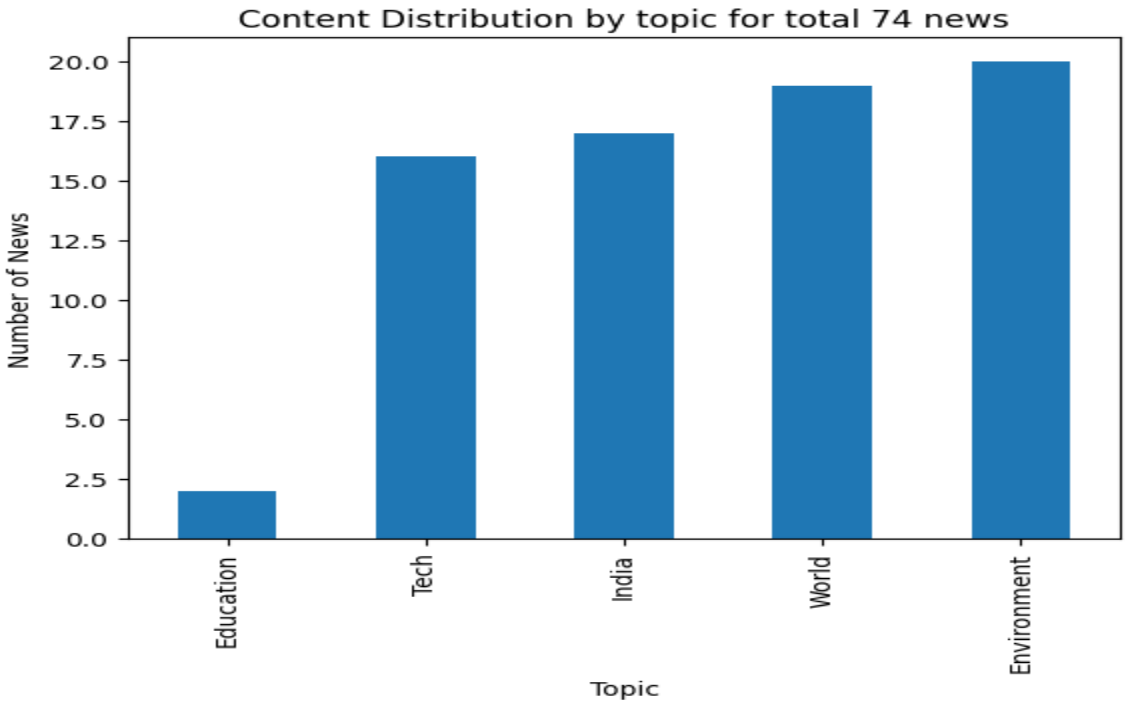
- Minimum sample (MCQs) size for GPT2-large based SLM: 64
- Minimum sample (MCQs) size for Gamini Flash 2.5 based SLM: 64

5.3.4 Research Question 4 minimum Sample Size

- Minimum sample (MCQs) size for GPT2-large based SLM: 64
- Minimum sample (MCQs) size for Gamini Flash 2.5 based SLM: 64

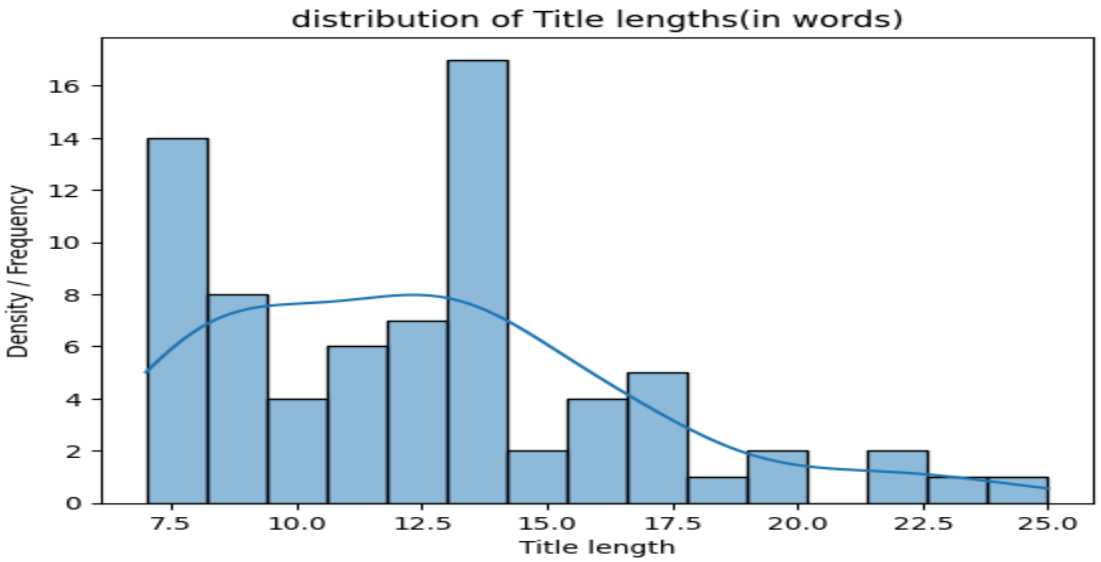
5.4 Visualizations and respective insights

5.4.1 Topic Distributions



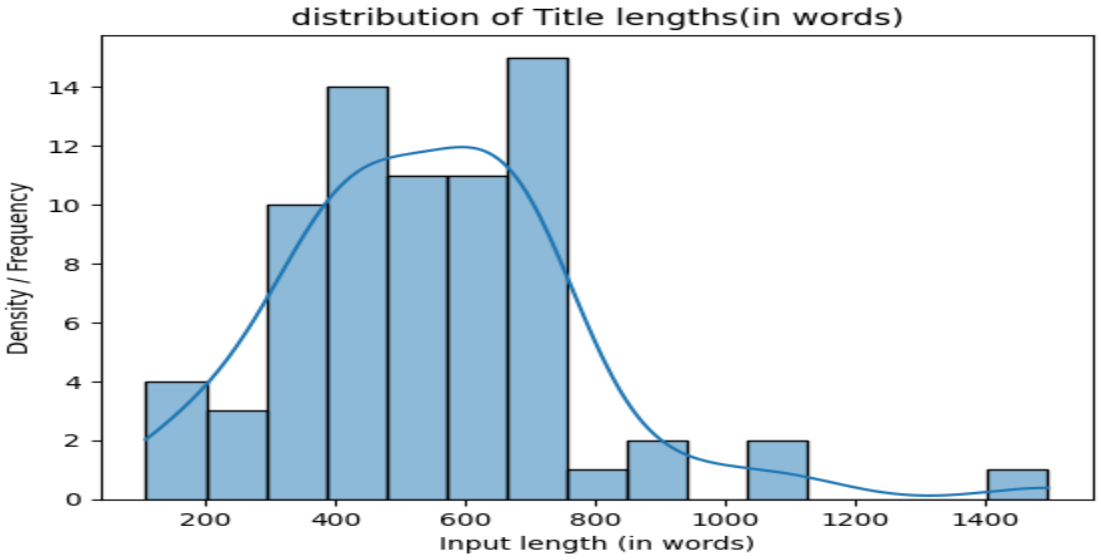
- The dataset spans **5 unique topics**.
- **Environment** is the most frequent topic, followed by Politics, Economy, World, and India.
- This indicates a **topic imbalance** (Environment is over-represented).

5.4.2 Title Lengths



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

5.4.3 Input Text Lengths



- 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**.

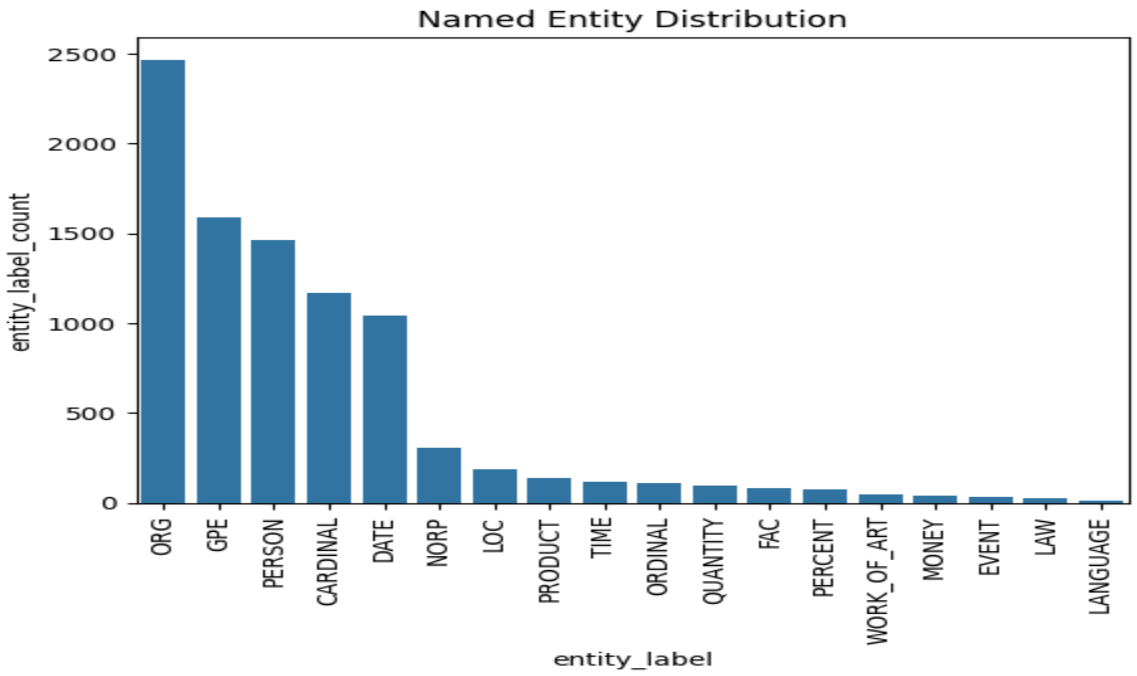
5.4.4 Word cloud

- **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.

The dataset provides rich factual entities (India, government, company, tech) that are suitable as correct answers in MCQs. Generic verbs like “said” or “will” are high-frequency but not useful for answer candidates — filtering such stopwords is important in preprocessing. The prominence of time-related words (“time”, “year”, “month”) indicates opportunities to generate time-specific questions, a key feature of Current Affairs MCQs.

5.4.5 Named Entity



- Dominant Entity Types:

  - **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.

- Quantitative References:
  - **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 — crucial for generating fact-based MCQs.
- Socio-Political Context:
  - **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.

The dataset is **entity-rich**, especially in **organizations, locations, people, and dates**, aligning well with the needs of Current Affairs MCQ generation. This distribution ensures that the fine-tuned GPT-2 model will have ample factual grounding to produce **relevant, diverse, and exam-quality questions**, while also highlighting areas (like Events and Laws) where more data collection could improve coverage.

5.4.6 Insights

- Dataset is **rich and well-structured** for MCQ generation.
- Slight **topic imbalance** suggests the need for sampling or weighting during training to avoid bias.
- **Concise titles + long input content** provide a good balance: headlines for stem framing, body text for answer/distractor generation.
- With ~550 words per input, the dataset can support **multiple MCQs per article**.

6 Modeling

6.1 Choice of models with justification

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 Feature selection with justification

When fine-tuning **GPT-2** for *automated MCQ generation in fast-changing domains like Current Affairs*, **feature selection** refers to identifying **what input signals (features) to expose to the model** to improve generation quality. Here is the list of features identified and curated for fine-tuning gpt-2 SLM

- **Textual Features from News Content:**
    - **Headline / Title** – captures the core event or fact.
    - **Lead Paragraph / Summary** – provides background context for generating stems.
    - **Named Entities** – persons, organizations, locations, and dates identified using NER help form answer keys and distractors.
    - **Keywords / Keyphrases** – extracted via TF-IDF or RAKE to highlight central themes.
- Temporal & Domain-Specific Features

  - **Publication Date** – ensures questions are time-relevant (important for Current Affairs).
  - **Source Metadata** – reliability or topic category (e.g., politics, economics, sports).
  - **Recency Score** – a derived feature giving weight to newer articles (helps avoid outdated MCQs).
- Question-Oriented Features:
  - **Answer Candidate (Key Fact)** – explicit marking of the correct answer (e.g., person/event extracted from text).
  - **Distractor Pool** – semantically similar entities (e.g., other politicians, countries, or dates) chosen as incorrect options.
  - **Context Window** – surrounding sentences that clarify the answer but can be used to shape the stem.
- Linguistic Features
  - **Part-of-Speech Tags** – ensure grammatical well-formedness of stems and distractors.
  - **Dependency Relations** – help identify subject–predicate–object triplets for fact-based MCQs.
  - **Sentence Complexity** – filtering out overly long or ambiguous sentences before feeding into GPT-2.
- Model Control Features (Auxiliary Inputs)
  - **Question Type Label** – e.g., Fact-based / Event-based / Person-based; guides the style of MCQ.

- **Difficulty Level** – derived from vocabulary complexity or entity familiarity, useful for tailoring assessments.
- **Topic Domain Tag** – labels like “Politics”, “Economy”, “Science” to steer the model toward domain-appropriate phrasing.

So, Feature selection for GPT-2 involves augmenting raw news text with structured signals (entities, dates, metadata, linguistic cues) that help the model focus on what’s answer-worthy, distractor-worthy, and time-relevant.

7 Preliminary Results

As per calculated sample size for each research question, minimum 64 MCQs need to be generated by GPT-2 based SLM and Gemini-Flash-2.5 LLM. Below section have attachment of MCQ generated by each model.

7.1 MCQs generated by fine-tuned GPT-2 SLM

The below file contains 74 MCQs generated on the given instruction with input context on current affairs news content of various topics such as India, Education, Environment etc.

__20250823_ca_mcqs_by_slm_gpt2-large.j

Here is one sample instruction, input context and output MCQ – multiple choice question.

```
```

{

"**link**": "https://timesofindia.indiatimes.com/india/we-have-reached-moon-and-mars-pm-modi-greets-the-nation-on-national-space-day-recalls-meeting-with-shubhanshu-shukla/articleshow/123466876.cms",

"**topic**": "India",

"**title**": "'We have reached Moon and Mars': PM Modi greets the nation on National Space Day; recalls meeting with Shubhanshu Shukla",

"**instruction**": "Generate a Multiple Choice Question with answer on the below India input context.",

"**input**": "Prime Minister Modi addressed the nation on National Space Day celebrating Indias space achievements and the accomplishments of its scientists and astronauts….

"**output**": "Question: Which country has become the first in the world to land on the South Pole of the Moon with Chandrayaan 3?

Options:

[A] India

[B] China

[C] Japan

[D] Australia

Correct Answer: A [India]"

}

Sample_mcq_generated_by_slm_gpt2.json

```
```

On high level self-review by human SME (self), the MCQ generation instruction, input context and output MCQ looks good.

7.2 MCQs generated by Gemni-flas-2.5 LLM

The below file contains 74 MCQs generated on the given instruction with input context on current affairs news content of various topics such as India, Education, Environment etc.

__20250823_ca_mcqs_by_gemini-2.5-flash

Here is one sample instruction, input context and output MCQ – multiple choice question.

```
{
        "link": "https://timesofindia.indiatimes.com/india/india-temporarily-suspends-most-postal-services-to-us-effective-august-25-amid-new-customs-order-know-whats-exempt-from-the-ban/articleshow/123469918.cms",
    "topic": "India",
    "title": "India temporarily suspends most postal services to US; effective August 25 amid new customs order \u2013 Know what\u2019s exempted from the ban",
    "instruction": "Generate a Multiple Choice Question with answer on the below India input context.",
    "input": "Effective August 25, 2025, the Department of Posts will temporarily suspend most international postal services to the US due to new customs regulations….”
,
```

"**output**": ### Multiple Choice Question: What is the primary reason cited by India's Department of Posts for the temporary suspension of most international postal services to the United States, effective August 25, 2025?

A) The US government imposed a complete ban on all postal articles from India due to economic sanctions.

B) **Air carriers expressed inability to comply with new US customs regulations due to unresolved implementation details and lack of preparedness.**

C) The Department of Posts decided to retaliate against the US for unrelated visa freezes and immigration issues.

D) India lacked the necessary infrastructure to handle an unexpected surge in international mail volume to the US.

Correct Answer: B

"

}

```

On high level self-review by human SME (self), the above MCQ generation instruction, input context and  output MCQ looks good.
```

8 Bibliography

[1] Agarwal, M., & Mannem, P. (2011). Automatic gap-fill question generation from text books. *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 56–64.

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

[3] Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. *Proceedings of ACL*, 1342–1352.

[4] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of ACL*, 8342–8360.

[5] Heilman, M., & Smith, N. A. (2010). Good question! Statistical ranking for question generation. *Proceedings of NAACL-HLT*, 609–617.

[6] Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *IEEE Access, 8*, 173222–173249.

[7] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv:1909.11942*.

[8] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., … Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of ACL*, 7871–7880.

[9] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67. (T5)

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

[11] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv:1910.01108*.

[12] Soni, S., & Sahu, S. (2021). Automatic question generation for educational applications: A survey. *International Journal of Artificial Intelligence in Education, 31*(2), 250–278.

[13] Susanti, Y., Tokunaga, T., Nishikawa, H., & Obari, H. (2018). Automatic distractor generation for multiple-choice English vocabulary questions. *Research and Practice in Technology Enhanced Learning, 13*(1), 1–16.

[14] 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.

9 Appendix