# How Good are Modern LLMs in Generating Relevant and High-Quality Questions at Different Bloom's Skill Levels for Indian High School Social Science Curriculum?

**Nicy Scaria[1], Suma Dharani Chenna[1,2], Deepak Subramani[1]**
[1]Computational and Data Sciences, Indian Institute of Science, India
[2]School of Computer Science and Engineering, VIT-AP University, India
{nicyscaria,deepakns}@iisc.ac.in
sumadharanichenna@gmail.com

## Abstract

The creation of pedagogically effective questions is a challenge for teachers and requires significant time and meticulous planning, especially in resource-constrained economies. For example, in India, assessments for social science in high schools are characterized by rote memorization without regard to higher-order skill levels. Automated educational question generation (AEQG) using large language models (LLMs) has the potential to help teachers develop assessments at scale. However, it is important to evaluate the quality and relevance of these questions. In this study, we examine the ability of different LLMs (Falcon 40B, Llama2 70B, Palm 2, GPT 3.5, and GPT 4) to generate relevant and high-quality questions of different cognitive levels, as defined by Bloom's taxonomy. We prompt each model with the same instructions and different contexts to generate 510 questions in the social science curriculum of a state educational board in India. Two human experts used a nine-item rubric to assess linguistic correctness, pedagogical relevance and quality, and adherence to Bloom's skill levels. Our results showed that $91.56\%$ of the LLM-generated questions were relevant and of high quality. This suggests that LLMs can generate relevant and high-quality questions at different cognitive levels, making them useful for creating assessments for scaling education in resource-constrained economies.

## 1 Introduction

In recent years, large language models (LLMs) have seen significant advances. They undergo training on extensive text datasets sourced from the internet and are utilized for a variety of natural language processing tasks. The introduction of OpenAI's ChatGPT and Google's Bard has made LLMs more accessible to a wider audience, enabling individuals without expertise in natural language processing (NLP) to leverage them for their everyday needs. These models are characterized by their substantial size and their ability to comprehend and produce intricate text. Through instruction fine-tuning, language models are calibrated to adhere to user directives (Zhang et al., 2022). In contrast to conventional language models, these LLMs possess zero-shot capabilities, allowing them to handle various tasks without specific training by simply interpreting the given instructions (Kojima et al., 2022). The educational applications of LLMs are varied and promising, covering personalized content generation, assessments, and feedback (Kasneci et al., 2023).

According to World Bank data, the teacher-pupil ratio in India's high schools is 1:29[1], compared to middle and high-income countries with an average of 1:18 and 1:13, respectively. This increases the workload on teachers and the quality of the instruction and assessment decreases. In India, subjects such as history are taught and evaluated, focusing on rote memorization (Sreekanth, 2007) with minimal emphasis on higher-order thinking skills or inquiry. Inquiry-based learning with high-quality questions fosters deep engagement and real-world connections for learners (Grant et al., 2022). Assessments aligned with Bloom's taxonomy levels (Anderson and Krathwohl, 2001), as detailed in Table 1, help educators identify learning gaps and personalize instruction, but require significant time and effort to create (Kurdi et al., 2020). Automated Educational Question Generation Systems (AEQG) have the potential to reduce this burden (Mulla and Gharpure, 2023), allowing teachers to personalize instruction and enhance student participation. This study investigates the capabilities of open source and proprietary LLMs to generate high-quality, context-aligned questions with different cognitive skills for effective assessments.

Although LLMs are capable of Natural Lan-

[1]https://data.worldbank.org



*Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10

Table 1: Revised Bloom's taxonomy (Anderson and Krathwohl, 2001) in ascending order in the cognitive dimension

| Bloom's level | Description |
|---|---|
| Remember | Retrieve relevant knowledge from long-term memory. |
| Understand | Construct meaning from instructional messages, including oral, written, and graphic communication. |
| Apply | Carry out or use a procedure in a given situation. |
| Analyze | Break material into foundational parts and determine how parts relate to one another and the overall structure or purpose. |
| Evaluate | Make judgments based on criteria and standards. |
| Create | Put elements together to form a coherent whole; reorganize into a new pattern or structure. |

guage Generation (NLG) tasks, their output can have errors and inconsistencies for specific contexts. These models are also prone to hallucinations (Ji et al., 2023). These issues directly impact the quality of educational questions generated, which can vary significantly across LLMs. For this reason, evaluating the quality of these questions is important. Despite the existence of automated techniques focusing on readability and linguistic aspects, these methods do not address pedagogical aspects and question appropriateness for the given context (Amidei et al., 2018a). Therefore, expert evaluation remains essential to guarantee the quality of LLM-generated questions.

In this study, we followed a zero-shot prompting approach for question generation. We prompted LLMs to generate questions at different cognitive levels, as defined in Bloom's taxonomy, on topics covering events of the Indian independence struggle from 1857 to 1947. Using five different LLMs, we generated 510 questions in total. Two subject matter experts evaluated the generated questions based on a nine-item rubric to consider both the linguistic and pedagogical aspects of the questions (Horbach et al., 2020).

This work investigates the following research questions. *(i)* Can modern LLMs generate relevant and high-quality educational questions of different cognitive levels and follow the instructions provided in the prompt?; *(ii)* Which LLM performs the best in question generation?

Our experiments and evaluations demonstrate that the questions generated by LLMs are relevant and of good quality. These LLMs can be used for AEQG with minimal effort of the educator. Our dataset 'HistoryQ'[2] containing 510 questions evaluated by two experts and annotated with Bloom's taxonomy levels will be made available for research in the development and evaluation of AEQG systems.

[2] https://github.com/nicyscaria/AEQG-SocialSciences-BloomsSkills

## 2 Related Work

Traditional automated question generation (AQG) systems mainly relied on question-answering datasets before the widespread adoption of LLMs. The primary reading comprehension datasets used for question generation tasks included SQUAD (Rajpurkar et al., 2016), SQuAD 2.0 (Rajpurkar et al., 2018) and NQ (Kwiatkowski et al., 2019). One of the crowd-sourced educational datasets used for question generation tasks is SciQ (Welbl et al., 2017). LearningQ(Chen et al., 2018) and EduQG(Hadifar et al., 2023) are the other two popular datasets available for AEQG. The lack of availability of these datasets for all subjects and the human expert labor associated with creating high-quality datasets restricted the ability to develop effective AQG systems (Zhang et al., 2021). With the advent of large transformer-based pre-trained large language models, NLG tasks in recent years have improved rapidly (Zhang et al., 2022). Pre-trained and fine-tuned models such as the Text-to-Text Transfer Transformer (T5) and GPT3 were used for question generation (Nguyen et al., 2022). Leaf (Vachev et al., 2022) is a question generation developed using a pre-trained T5 model. A pre-trained T5 model (EduQG) was developed in educational text to improve the quality of the generated question (Bulathwela et al., 2023). Most AEQG systems are generic with a focus on reading comprehension or science and mathematics. AEQG research for social sciences is minimal (Bechet et al., 2022; Antoine et al., 2023). Subjects like science and mathematics tend to seek precise, quantifiable,

and objective answers. But for subjects like social sciences, the questions can be more subjective, often do not have a single correct answer, and can be interpreted differently by different people.

Many AQG systems, built by fine-tuning LLMs on specific datasets such as the ones mentioned above, often generate questions that focus on lower-order cognitive skills or simply retrieve answers directly from the context information provided (Ushio et al., 2022; Bulathwela et al., 2023). Most of the questions in EduQG(Hadifar et al., 2023) are within the first three levels of Bloom's taxonomy. These questions do not assess students' higher-order thinking abilities. Bloom's taxonomy guides educators in generating learning objectives and questions to teach and test different cognitive skills. A recent work (Sridhar et al., 2023) uses GPT4 to create course content based on Bloom's taxonomy. Although automated metrics exist to evaluate machine-generated questions, they primarily analyze linguistic aspects. In the case of educational question generation, pedagogical elements play a crucial role. Expert evaluation is necessary to understand the pedagogical aspects of machine-generated questions (Horbach et al., 2020; Steuer et al., 2021). Such evaluations are also used in student-generated questions (Moore et al., 2022).

## 3 Methodology

### 3.1 Language models and content

We chose five recent open-source and proprietary LLMs for the study. LLMs used in this study were Falcon 40B (falcon-40b-instruct), Llama 2 70B (Llama-2-7b-chat-hf), Palm 2 (chat-bison-001), GPT-3.5 (gpt-3.5-turbo-0613), and GPT-4 (gpt-4-0613). Among these, Falcon 40B is the smallest LLM with 40 billion parameters and GPT 4 is the largest (rumored, as the exact number of parameters is unknown). The questions were generated for the subject "History", covering events of the Indian independence struggle from 1857 to 1947. We used content from two chapters of the tenth grade social science textbook called *Samacheer Kalvi* (Tamil Nadu Textbook and Educational Services Corporation. State Council of Educational Research and Training, 2022) used in schools under the Indian state of Tamil Nadu's educational board. The text is in English. This content served as the context for LLMs based on the questions generated. The average length of the context was around 450 words, making it equivalent to around 600 tokens. The LLMs used had a sequence length of more than 1024 tokens to accommodate this context length and instructions. We consider 17 such contexts, so that overall nearly 500 (510, to be exact) questions are generated.

### 3.2 Prompt design and question generation

Each prompt had a context and instructions associated with it. The prompts were designed using techniques of pattern reframing, itemizing reframing, and assertions (instead of negations) (Mishra et al., 2022). Most Indian students, even at the tertiary level of education, are only within level B2 of the Common European Reference Framework (CEFR) for English (Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, 2001; Ravindra Babu and Shiela Mani, 2018). Therefore, additional instruction was provided in the prompt to use words within the CEFR B2 level. This approach would help students better understand the questions, thus decreasing the chances of confusion or misunderstanding arising from difficulties in comprehending the language.

We gave the same prompt to all LLMs. Each LLM had to generate six questions, one for each level in Bloom's taxonomy corresponding to the 17 contexts. Each model generated 102 questions, resulting in a total of 510 questions. The sampling temperature of an LLM typically varies between 0 and 1 in most implementations. A lower temperature results in a more deterministic output from the LLM, giving preference to the most probable predictions, while a higher temperature increases the randomness in the LLM output, resulting in less probable predictions (Hinton et al., 2015; Wang et al., 2020, 2023). A temperature value of 0.9 was used for AEQG with the LLMs to maximize the variety and diversity of the generated questions. The example of generation prompts is given in the Appendix A.1.

### 3.3 Human evaluation

Two experts evaluated the relevance and quality of the 510 questions based on a nine-item rubric (Table 2), a modified version of the nine-item rubric in Horbach et al.'s (2020). The two experts had subject knowledge and experience in teaching the subject social sciences and worked on question-generation tasks for multiple organizations. The experts were presented with the LLM questions in random order with only context information.

Table 2: Hierarchical nine-item rubric used to evaluate questions generated by LLMs along with the percentage agreement and Cohen's $\kappa$ for each item

| Rubric item | Definition |
|---|---|
| **Understandable** (100.00%, $\kappa$ = 1.00) | Could you understand what the question is asking? |
| ContextRelated (100.00%, $\kappa$ = 1.00) | Is the question related to the context given? |
| Grammatical (100.00%, $\kappa$ = 1.00) | Is the question grammatically well-formed? |
| **Clear** (99.61%, $\kappa$ = 0.79) | Is it clear what the question asks for? |
| **Answerable** (99.60%, $\kappa$ = 0.88) | Can students answer the question? |
| InformationNeeded (86.80%, $\kappa$ = 0.73) | What kind of information is needed to answer the question?<br>• Information presented directly and in one place only in the text<br>• Information presented in different parts of the text<br>• A combination of information from the text with external knowledge<br>• General knowledge about the topic, not from the text<br>• The reader's feelings /judgements /... about the text<br>• The reader's feelings/judgements/... about the text with external knowledge |
| Central (100.00%, $\kappa$ = 1.00) | Do you think being able to answer the question is important to work on the topic covered in the context? |
| WouldYouUseIt (90.87%, $\kappa$ = 0.84) | If you were a teacher working with that text in class, do you think you would use this question? |
| **Bloom'sLevel** (89.41%, $\kappa$ = 0.95) | What is the Bloom's skill associated with the question? |

They were asked to respond to each question on the rubric hierarchically from top to bottom. Seven items in the rubric were a 'yes' or 'no' response. The *InformationNeeded* item comprises six unique options that indicate what information is needed to answer the question. The questions in social sciences can be subjective and sometimes do not have a single correct answer. They can be open to interpretation. Due to this, the *InformationNeeded* contains options like 'The reader's feelings /judgements /... about the text' in addition to information derived from both the text itself and external sources. The *Bloom'sLevel* item consists of the different skills defined in Bloom's taxonomy cognitive dimension, viz., remember, understand, apply, analyze, evaluate, and create. The specifics regarding the meaning of each level of Bloom's Skill are provided in Table 1. Along with 'yes' or 'no', the option 'maybe' is also added in the *WouldYouUseIt* rubric item. In the evaluation metrics, *WouldYouUseIt* is the most subjective one.

The rubric items are structured hierarchically (Table 2), which means that if a criterion in bold font is answered with a 'no', the subsequent items in the rubric would not be considered for evaluation. For instance, if *Understandable*, *Clear*, or *Answerable* is marked 'no', the following items are not evaluated for that question and are marked as 'not applicable'. This simplifies the evaluation process.

A question is relevant and of high quality if experts say 'yes' for *Understandable*, *ContextRelated*, *Grammatical*, *Clear*, *Answerable*, and *Central* and mark 'yes' or 'maybe' for *WouldYouUseIt*. Furthermore, we utilized the *Bloom'sSkill* and *CEFRLevel* to understand whether the LLM adheres to the instructions provided in the prompt. Evaluators had to select the Bloom's level for *Bloom'sSkill* metric. We used 'Text Inspector'[3] developed by Cambridge as part of their English Profile Research (Alexopoulou, 2008) to understand the CEFR level of vocabulary used in the question. The LLM adhered to the instructions provided if the *Bloom'sSkill* label given by the evaluators matches the Bloom's

[3] https://www.englishprofile.org/wordlists/text-inspector

skill level in the prompt to the LLM and if the words are within B2 for *CEFRLevel*.

Since experts' opinions on LLM-generated questions are influenced by their writing style preferences, personal beliefs, knowledge base, and focus on detail (Amidei et al., 2018b), two inter-rater reliability measures, namely, percentage agreement and Cohen's Kappa $\kappa$ (Cohen, 1960; McHugh, 2012) were used. The former is the proportion of times experts agreed on a specific rating and the latter is a robust measure that accounts for the chance agreement and provides a more accurate estimate of the true agreement between experts. Cohen's $\kappa$ treats all disagreements as equal, but the disagreements cannot be considered the same for the ordinal metrics, *WillYouUseIt* and *Bloom'sLevel*. In this case, we used the quadratic weighted Cohen's $\kappa$ (Cohen, 1968) instead of the simple Cohen's $\kappa$ to penalize considerable disagreements more than minor disagreements.

## 4 Results and analysis

The percentage agreements and Cohen's $\kappa$ values obtained between the two human evaluators for the nine-item rubric are given in Table 2. The percentage agreements and Cohen's $\kappa$ values are calculated only for questions not labeled 'no' for the preceding rubric items in the hierarchy (marked in bold). These values indicate substantial agreement between experts on most of the metric items. Four items, *Understandable*, *ContextRelated*, *Grammatical*, and *Central* had perfect agreement.

### 4.1 Relevance and quality metrics

Both experts rated 100% of the generated questions as *Understandable*, *ContextRelated*, and *Grammatical*. Of these, 98.82% of the questions were rated as *Clear* and 97.84% as *Answerable*. Among the *Answerable* questions, evaluators chose one option out of the six for *InformationNeeded* item. According to the evaluators, the knowledge needed to answer 19.22% of the questions could be found in one place in the context, 18.24% from a different part of the context, and 23.33% questions needed a combination of information from the context along with external knowledge. Only 0.2% of the questions required general knowledge alone to answer, with no necessary context information. 13.73% and 10.39% of the questions required the reader's judgement about the text and the reader's judgement about the text along with external knowledge, respectively, to provide an answer. Experts rated 95.88% of the questions as *Central* to the topics covered in the respective contexts. The evaluators responded either 'yes' or 'maybe' to *WouldYouUseIt* rubric item for 91.56% of the questions. Thus, we say that the experts rated 91.56% of generated questions as relevant and high quality.

Table 3: Performance of all generated questions on different evaluation metrics

| Metric | Questions (%) |
|---|---|
| Relevant & High quality | 91.56% |
| Adherence | |
| • Bloom'sLevel | 76.53% |
| • CEFRLevel | 87.64% |

It is observed that in the *Bloom'sLevel* metric, there is an adherence of 76.53% between the evaluators and the LLM. In the *CEFRLevel*, the adherence is 87.64% (Table 3). We are releasing our dataset, 'HistoryQ' containing 510 LLM-generated questions annotated with the nine-item metric by experts along with *CEFRLevel* for further study and analysis by the community. Examples of some relevant and high-quality questions based on Bloom's taxonomy that adhered to the instructions in the prompt are given in the Appendix A.2.

### 4.2 Performance of different LLMs

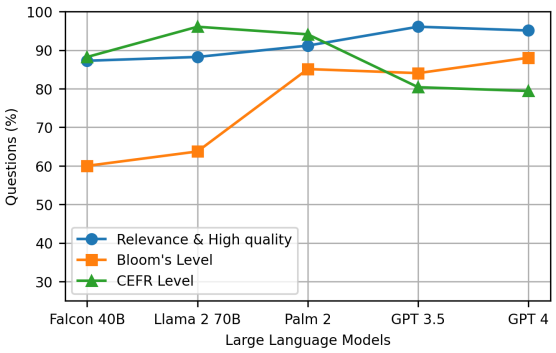


Figure 1: Performance of different LLMs on the different evaluation metrics.

The performance of the five LLMs in the AEQG task according to different evaluation criteria is summarized in Table 4. We observed that proprietary models, Palm 2, GPT 3.5, and GPT 4, which are believed to have 175 billion plus or even trillions of parameters, outperformed open-source models with 40 and 70 billion parameters in all criteria except the CEFR level adherence metric, as

Table 4: Performance of different large language models on different evaluation metrics

| Metric | Falcon 40B | Llama 2 70B | Palm 2 | GPT 3.5 | GPT 4 |
|---|---|---|---|---|---|
| Relevance & High quality | 87.25% | 88.24% | 91.18% | 96.08% | 95.10% |
| Adherence | | | | | |
| • Bloom'sLevel | 60.00% | 63.73% | 85.10% | 84.04% | 88.04% |
| • CEFRLevel | 88.23% | 96.07% | 94.11% | 80.39% | 79.41% |

Table 5: Precision, recall and F1 score of different large language models on Bloom's skill level compared with expert opinion

| Metric | Falcon 40B | Llama 2 70B | Palm 2 | GPT 3.5 | GPT 4 |
|---|---|---|---|---|---|
| Precision | 0.60 | 0.65 | 0.85 | 0.84 | 0.87 |
| Recall | 0.60 | 0.66 | 0.86 | 0.86 | 0.88 |
| F1 score | 0.57 | 0.62 | 0.85 | 0.84 | 0.87 |

indicated in Figure 1.

Aligning with Bloom's taxonomy level was one of the important criteria in this study. The skill levels given by the LLM for the generated questions were compared with the ground-truth skill level labels provided by the human raters. The corresponding precision, recall, and F1 score for this task are shown in Table 5. GPT 4 outperforms other models, while Palm 2 and GPT 3.5 are in the second and third positions.

## 5 Conclusion

We found that 91.56% of the questions generated by different LLMs are relevant and of high quality. This indicates that LLMs can be used for AEQG with minimal effort of the educator. However, the performance varies between different LLMs. GPT 3.5 and GPT 4 generated the highest proportion of relevant and high-quality questions. In the metric of adherence to Bloom's level, GPT 4 outperformed the other models, followed by Palm 2. In contrast, the open source LLMs, Falcon 40B and Llama 2 70B, performed poorly on all metrics, except adherence to CEFR levels. This could be due to the large size of these proprietary models, which results in their ability to capture and represent complex patterns in the text data. Another interesting observation in the study was the inability of most models to generate high-quality questions at the 'Apply' and 'Create' levels of Bloom's taxonomy. GPT 3.5 and GPT 4 showed comparable performance in all criteria. Surprisingly, GPT 4 and GPT 3.5 had poor alignment with the CEFR level requested in the prompt. These models produced complex texts compared to other models.

Our research suggests that educators can leverage Palm 2, GPT 3.5, and GPT 4 to create relevant, high-quality questions of different cognitive levels defined by Bloom's taxonomy for scaling social science research in India. The LLMs must be prompted with the context in English obtained from the relevant curriculum. This approach considerably reduces the workload on teachers, especially in an under-resourced school setting where the teacher-pupil ratio is low. In addition, students can create practice tests for themselves and identify learning gaps. Expert-evaluated 'HistoryQ' could serve as a training and validation dataset for research involving the development and evaluation of AEQG models with a focus on higher-order cognitive skills.

## 6 Limitations

Our study required considerable time and effort from experts. Despite rigorous efforts to ensure objectivity in the evaluation through a detailed rubric and a randomized presentation of LLM-generated questions, it is important to recognize that expert evaluations can still exhibit inherent subjectivity, influenced by individual perspectives and biases. An automated system to assess the quality of machine-generated questions for their pedagogical and linguistic aspects can reduce this time and effort. This paves the way for exploring and creating high-quality automated evaluation systems. Furthermore, our study used the same prompt in different contexts for all LLMs. We did not investigate the performance of models on diverse prompts with additional information or few-shot prompting. This is another potential future direction for exploring the performance of LLMs.

## References


Theodora Alexopoulou. 2008. Building new corpora for english profile. *Research Notes*, 33:15–19.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018a. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018b. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329.

Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

Elie Antoine, Hyun Jung Kang, Ismaël Rousseau, Ghislaine Azémard, Frédéric Bechet, and Géraldine Damnati. 2023. Exploring social sciences archives with explainable document linkage through question generation. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 141–151.

Frédéric Bechet, Elie Antoine, Jeremy Auguste, and Géraldine Damnati. 2022. Question generation and answering for exploring digital humanities collections. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4561–4568.

Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. 2023. Scalable educational question generation with pre-trained language models. In *International Conference on Artificial Intelligence in Education*, pages 327–339. Springer.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: a large-scale dataset for educational question generation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

SG Grant, Kathy Swan, and John Lee. 2022. *Inquiry-based practice in social studies education: Understanding the inquiry design model*. Taylor & Francis.

Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Eduqg: A multi-format multiple-choice dataset for the educational domain. *IEEE Access*, 11:20885–20896.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle, and Montse Maritxalar. 2020. Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1753–1762.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing Instructional Prompts to GPTk's Language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612.

Steven Moore, Huy A Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. 2022. Assessing the quality of student-generated short answer questions using gpt-3. In *European conference on technology enhanced learning*, pages 243–257. Springer.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

Huy A Nguyen, Shravya Bhat, Steven Moore, Norman Bier, and John Stamper. 2022. Towards generalized methods for automatic question generation in educational domains. In *EC-TEL*, pages 272–284. Springer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

KVB Ravindra Babu and K Ratna Shiela Mani. 2018. Where do we stand on cefr? an analytical study on esl learners' language proficiency. *Language in India*, 18(12).

Y Sreekanth. 2007. An analysis of question papers of different boards of examinations in social sciences. *Indian Educational Review*, 43(2):18.

Pragnya Sridhar, Aidan Doyle, Arav Agarwal, Christopher Bogart, Jaromir Savelka, and Majd Sakr. 2023. Harnessing llms in curricular design: Using gpt-4 to support authoring of learning objectives. *arXiv preprint arXiv:2306.17459*.

Tim Steuer, Leonard Bongard, Jan Uhlig, and Gianluca Zimmer. 2021. On the linguistic and pedagogical quality of automatic question generation via neural machine translation. In *EC-TEL 2021, Proceedings*, pages 289–294. Springer.

Tamil Nadu Textbook and Educational Services Corporation. State Council of Educational Research and Training. 2022. *Standard Ten, Social Science*. Directorate of School Education Tamil Nadu.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2022. Generative Language Models for Paragraph-Level Question Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, U.A.E. Association for Computational Linguistics.

Kristiyan Vachev, Momchil Hardalov, Georgi Karadzhov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2022. Leaf: Multiple-choice question generation. In *European Conference on Information Retrieval*, pages 321–328. Springer.

Chi Wang, Xueqing Liu, and Ahmed Hassan Awadallah. 2023. Cost-effective hyperparameter optimization for large language model generation inference. In *International Conference on Automated Machine Learning*, pages 21–1. PMLR.

Pei-Hsin Wang, Sheng-Iou Hsieh, Shih-Chieh Chang, Yu-Ting Chen, Jia-Yu Pan, Wei Wei, and Da-Chang Juan. 2020. Contextual temperature for language modeling. *arXiv preprint arXiv:2012.13575*.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*.

Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM TOIS*, 40(1):1–43.

# A Appendix

## A.1 Example prompt with a specific context

The example prompt for a specific context given to the LLMs to generate the questions is given below. All the instructions and other details remained the same for other prompts except for the context information.

*Please read through the following context and instructions to create high quality questions based on the context and as per the instructions.*

***Context**:*

*In 1857, British rule witnessed the biggest challenge to its existence. Initially, it began as a mutiny of Bengal presidency sepoys but later expanded to the other parts of India involving a large number of civilians, especially peasants. The events of 1857–58 are significant for the following reasons: 1. This was the first major revolt of armed forces accompanied by civilian rebellion. 2. The revolt witnessed unprecedented violence, perpetrated by both sides. 3. The revolt ended the role of the East India Company and the governance of the Indian subcontinent was taken over by the British Crown.*

*(a) Causes*

1. *Annexation Policy of British India*
   *In the 1840s and 1850s, more territories were annexed through two major policies: The Doctrine of Paramountcy. British claimed themselves as paramount, exercising supreme authority. New territories were annexed on the grounds that the native rulers were inept, and the Doctrine of Lapse. If a native ruler did not have male heir to the throne, the territory was to 'lapse' into British India upon the death of the ruler. Satara, Sambalpur, parts of the Punjab, Jhansi and Nagpur were annexed by the British through the Doctrine of Lapse.*
2. *Insensitivity to Indian Cultural Sentiments*
   *In 1806 the sepoys at Vellore mutinied against the new dress code, which prohibited Indians from wearing religious marks on their foreheads and having whiskers on their chin, while proposing to replace their turbans with a round hat. It was feared that the dress code was part of their effort to convert soldiers to Christianity. Similarly, in 1824, the sepoys at Barrackpur near Calcutta refused to go to Burma by sea, since crossing the sea meant the loss of their caste. The sepoys were also upset with discrimination in salary and promotion. Indian sepoys were paid much less than their European counterparts. They felt humiliated and racially abused by their seniors.*

*(b) The Revolt of 1857*

*The precursor to the revolt was the circulation of rumors about the cartridges of the new Enfield rifle. There was strong suspicion that the new cartridges had been greased with cow and pig fat. The cartridge had to be bitten off before loading (pork is forbidden to the Muslims and the cow is sacred to a large section of Hindus). On 29 March a sepoy named Mangal Pandey assaulted his European officer. His fellow soldiers refused to arrest him when ordered to do so. Mangal Pandey along with others were court-martialled and hanged. This only fuelled the anger and in the following days there were increasing incidents of disobedience. Burning and arson were reported from the army cantonments in Ambala, Lucknow, and Meerut.*

***Instructions:***

1. *Create a question for each cognitive level in Bloom's taxonomy: remember, understand, apply, analyze, evaluate, and create from the context.*
2. *Ensure the questions use B2 level words or below of the Common European Framework of Reference for the English Language.*
3. *Make sure the questions relate to the students in India.*
4. *Make sure to connect events within the context while creating questions.*

## A.2 Examples of LLM generated questions

Some questions generated by LLMs that are relevant, high-quality, and adhered to instructions are given along with Bloom's skill associated with the question.

- **Remember**: Name the three leaders referred to as Lal-Bal-Pal during the Swadeshi period.

- **Understand**: How did the Swadeshi movement help to promote Indian industries?
- **Apply**: How would you promote the concept of Swadeshi today, especially given the globalized world we live in?
- **Analyze**: How did the development of Swadeshi industries relate to the wider goals of the Swadeshi Movement?
- **Evaluate**: Considering the importance of self-sufficiency, do you think the boycott of foreign goods was an effective method in promoting Swadeshi industries? Provide reasons for your answer.
- **Create**: Compose a short speech or paragraph encouraging fellow students to support Swadeshi industries, drawing inspiration from the historical events mentioned.