PustakAI: Curriculum-Aligned and Interactive Textbooks Using Large Language Models

Shivam Sharma, Riya Naik, Tejas Gawas, Heramb Patil, and Kunal Korgaonkar

CSIS Department, BITS Pilani K K Birla Goa Campus, India
{kunalk}@goa.bits-pilani.ac.in

Abstract. Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding and generating human-like content. This has revolutionized various sectors such as healthcare, software development, and education. In education, LLMs offer potential for personalized and interactive learning experiences, especially in regions with limited teaching resources. However, adapting these models effectively to curriculum-specific content, such as the National Council of Educational Research and Training (NCERT) syllabus in India, presents unique challenges in terms of accuracy, alignment, and pedagogical relevance. In this paper, we present the framework "PustakAI"[1] for the design and evaluation of a novel question-answering dataset "NCERT-QA" aligned with the NCERT curriculum for English and Science subjects of grades 6 to 8. We classify the curated QA pairs as Factoid, Inferential, and Others (evaluative and reasoning). We evaluate the dataset with various prompting techniques, such as meta-prompt, few-shot, and CoT-style prompting, using diverse evaluation metrics to understand which approach aligns more efficiently with the structure and demands of the curriculum. Along with the usability of the dataset, we analyze the strengths and limitations of current open-source LLMs (Gemma3:1b, Llama3.2:3b, and Nemotron-mini:4b) and high-end LLMs (Llama-4-Scout-17B and Deepseek-r1-70B) as AI-based learning tools in formal education systems.

Keywords: Large Language Model · QA Systems · Educational AI.

1 Introduction

The idea of a machine that could answer questions, write essays, or translate natural language like humans seemed fiction. But with the advent of Large Language Models (LLMs), that vision has become reality. LLMs built on a vast amount of data learn using transformer architectures. As they evolved, LLMs have been used to assist in various tasks such as composing emails, writing code, scientific discovery, and tutoring. It is the last category that is of interest to us. Trained on a diverse range of subjects and answering styles, LLMs can assist students by answering questions, explaining complex concepts, and even offering feedback on their submissions. On the other hand, for educators, it can help create

[1] Pustak means 'book' in many Indian languages.

lesson plans, automate learning tasks such as creating balanced questionnaires, and also explain concepts with additional research. LLMs have the potential to benefit education by extending learning beyond standard teaching-learning and help bridge the educational gap. To facilitate this, increasing integration of language models into educational contexts has prompted a wave of research exploring their capabilities, limitations, and impact.

Researchers and developers are utilizing LLMs to create interactive learning platforms that can adapt to student needs. These models have been used to generate practice questions, summarize complex topics, provide coding help, and translate languages, thereby supporting diverse learners across disciplines. However, the challenge of efficiently training LLMs for educational purposes remains. This is largely due to the quality of training data and the inference methods employed. Recent progress in dataset development has focused on general educational content, but for practical use in institutional settings, the data must be tailored to specific curriculam. Therefore, it is equally important to improve LLMs for educational question answering by introducing more effective inference methods and creating high-quality, curriculum-aligned datasets. This will enable the fine-tuning of both LLMs and traditional language models, leading to more accurate and contextually relevant educational responses.

Our framework PustakAI aims to contribute as follows:

- Present the **NCERT-QA** dataset to implement a curriculum-aligned Q&A system. This study validates the NCERT-QA dataset as a foundation for building a curriculum-aligned Q&A system. We develop a QA dataset derived directly from curriculum content and conduct a comprehensive evaluation of its effectiveness through various inference prompting techniques.
- Demonstrate the unique challenge posed by our curriculum-specific dataset by baselining it against a general-domain benchmark such as SQuAD.
- Perform an in-depth analysis to identify the optimal models and prompting strategies for the NCERT-QA task, including a comparison between subjects
- Analyze the practical trade-offs between performance and efficiency to make a case for cost-effective deployment in real-world school settings.

The rest of the paper is organized as follows. Section 2 summarizes existing educational datasets, Sections 3 and 4 elaborate the dataset curation steps and dataset analysis, Section 5 describes the implementation of the pipeline to evaluate the dataset and inference strategies, and Section 6 summarizes the experimental evaluation and results. Finally, Section 7 summarizes the conclusions drawn.

2 Background and Related Work

The application of LLMs to question answering has evolved from general benchmarks like SQuAD [8] to high-stakes domains like education, where the risks of "hallucination" demand high factual accuracy and pedagogical alignment. This has driven research in two key areas: the creation of curriculum-aligned datasets

and the development of methodologies like advanced prompting to ensure model outputs are faithful to reliable sources.

Early educational datasets focused on broad reasoning, such as ARC for science [4] and FairytaleQA for narrative comprehension [16]. A move toward direct curriculum alignment was marked by datasets like RACE, sourced from student examinations [8]. This trend has intensified with resources tightly coupled to specific textbook content, such as CK12-QA for science [1] and PeerQA for scientific reviews [2]. Concurrently, the focus on evaluation has sharpened, with benchmarks like SyllabusQA introducing fact-checking metrics [6] and TruthfulQA testing models against common falsehoods [11]. Despite this progress, a significant gap remains for a large-scale dataset aligned with a major non-Western curriculum like India's NCERT, a gap our work aims to fill. A comparative analysis of these datasets is provided in (Table 1).

Aligning LLMs to be pedagogically sound is a key challenge, as general-purpose models are not inherently suited for the classroom. Frameworks like COGENT demonstrate how to generate grade-appropriate content by providing structured guidance on learning objectives and readability [12]. This has also prompted architectural debates, contrasting large unified models with more efficient Mixture-of-Experts (MoE) architectures tailored to specific curricula [13]. Our evaluation of a wide spectrum of models contributes directly to this investigation, providing empirical data on the performance-cost trade-offs for deploying practical AI tools in school systems.

In education, ensuring faithfulness (grounding answers in trusted sources) is non-negotiable. The standard approach combines Retrieval-Augmented Generation (RAG) for its architecture [9] with advanced prompting to control the model's reasoning process. While Chain-of-Thought (CoT) was an early breakthrough for eliciting reasoning [7], the efficacy of popular frameworks like ReAct has been challenged. A recent critical evaluation found that ReAct's performance gains stem from exemplar similarity rather than genuine reasoning, revealing a failure to generalize [3]. This critique highlights the need for robust alternatives like meta-prompting, which uses high-level, structural guidance to enforce a faithful reasoning process [17]. Our work directly investigates if this structural approach is more effective than the content-based guidance of few-shot or CoT prompts in a curriculum-aligned context.

3 Dataset Curation

Our NCERT-QA data curation process includes three major steps: collection of NCERT text documents, data extraction, and answer mapping. Each of the steps is detailed below and is visualized in Fig. 1(a). Our objective is to collect high-quality, authentic data. To achieve this, we gathered 35 documents (chapters) from the English curriculum and 48 documents from the Science curriculum for classes 6 to 8. Each document's textual content was meticulously extracted from PDFs, with all images, captions, and tables removed to ensure the purity of the text. Our overall refined corpus comprised a total of 83 documents.

Table 1: Comparative Analysis of Key Question-Answering Datasets.

Dataset	Focus	Target Age/Grade	Subjects	Lang.	Curriculum Alignment	Key Features / Relevance
SQuAD	Extractive QA	General Adult	General (Wikipedia)	English	None	Foundational extractive QA benchmark.
RACE	Reading Comprehension	Grades 7-12 (Ages 12-18)	English	English	High (Chinese Examinations)	Precedent for curriculum-aligned QA from exams.
ARC	Science Reasoning	Grades 3-9	Science	English	Loose (Grade-level science)	Benchmark for complex science reasoning.
SciQ	Science QA	General	Science (Physics, Chem, Bio)	English	Loose (General science topics)	Provides supporting evidence text with questions.
TruthfulQA	Factual Faithfulness	General Adult	General (38 categories)	English	N/A	Measures model's ability to avoid common falsehoods.
FairytaleQA	Narrative Comprehension	Grades K-8	Reading/ Stories	English	None	Expert-generated questions for younger students.
CK12-QA	Multimodal Textbook QA	Middle School	Science	English	High (CK-12 Textbooks)	Direct parallel for RAG on science textbooks.
SyllabusQA	Course Logistics QA	University	General (36 majors)	English	High (University Syllabi)	Introduces Fact-QA metric for factual accuracy.
PeerQA	Scientific Document QA	Graduate+	STEM/NLP	English	N/A (Scientific Papers)	Expert-generated questions from authentic sources.
NCERT-QA	**Curriculum-aligned QA**	**Grades 6-8**	**Science, English**	**English**	**High (Indian NCERT)**	**Addresses the gap for a major non-Western curriculum.**

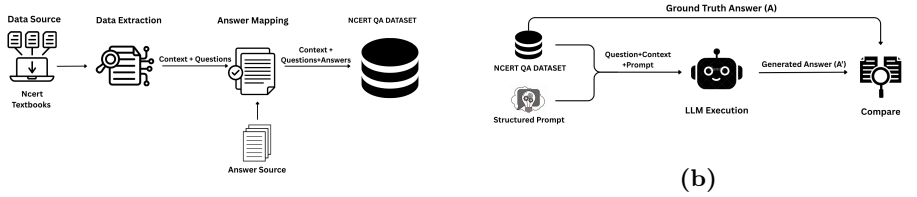

Fig. 1: (a) NCERT-QA dataset curation process. NCERT textbooks are parsed to extract chapters as context and the respective questions. Answers are retrieved from various authentic public online sources and aligned based on chapter and question indices. These answers are used as ground truth. The resulting QA dataset is structured as a collection of context-question-answer tuples. (b) LLM prompting and evaluation pipeline. LLM is presented with chapter and its corresponding question by employing a variety of prompting strategies. Model then generates a response to the question which is compared against ground truth using various evaluation matrices.

3.1 Data Extraction

In each chapter, we systematically extracted the chapter text as context and the questions provided in the exercises of the respective chapter. Specifically, a total of 451 questions were extracted for English documents, while 288 questions were obtained from the science documents, bringing the dataset to a total of 739 question-answer pairs. These extracted questions were subsequently categorized into three distinct types: Factoid, Inferential, and a third category denoted as Others. The **Factoid** category consists of questions whose answers can be directly extracted from the passage, identified as specific spans of text. The **Inferential** category includes questions that necessitate logical reasoning and inferential thinking to develop a comprehensive response based on the passage content. The **Other** category is distinct from the first two; it encompasses questions that are boolean in nature or require answers extracted from multiple paragraphs, among other traits. Examples for each category are illustrated in Table 2(a)

Table 2: (a) Examples of different question categories with corresponding answers; (b) Distribution of Question Types in the NCERT-QA Dataset.

Category	Question	Answer
Factoid	What did Patrick think his cat was playing with? What was it really?	Patrick thought his cat was playing with a doll, but it was actually a tiny man.
Inferential	Why did the little man grant Patrick a wish?	Because Patrick saved him from the cat and the elf wanted to return the favor.
Others	In what way did the shopkeeper make a fool of Rasheed?	The shopkeeper pretended Rasheed could win prizes, but tricked him with cheap goods.

Question Category	Count	Percentage
Factoid	405	55%
Inferential	258	35%
Other	76	10%
Total	**739**	**100%**

3.2 Answer Mapping

In the final phase of dataset curation, we execute the process of correlating answers to the questions extracted from the exercises. We utilize the solutions to each specified question sourced from authentic public online sources. The mapping of questions is achieved by utilizing the chapter index and question number as reference points. Subsequently, we manually assess the accuracy and appropriateness of answer mappings to ensure their correctness.

4 Exploratory Dataset Analysis

In this section, we conduct an in-depth analysis of the NCERT-QA dataset to better understand its defining characteristics and underlying structure. Our

primary focus lies in examining the diversity of the data and the formulation of question-answer pairs. This analysis is crucial for evaluating the dataset's suitability for building robust educational QA systems and for highlighting how its features can be used to test different model capabilities.

4.1 Data Diversity

The NCERT-QA dataset is intentionally diverse, covering two distinct subjects English and Science across three consecutive grade levels (6, 7, and 8). This subject diversity introduces a variety of text complexities and styles. The English texts are primarily narrative and literary, requiring comprehension of plot, character, and thematic elements. In contrast, the Science texts are descriptive and explanatory, demanding an understanding of concepts, processes, and factual information. This is evident from the Q&A length distribution shown in Fig 2. The data indicate that Science subject documents tend to feature more elaborate questions and answers compared to those in English. Most English questions fall within the 5–10 word range, while Science questions typically range from 10–20 words, reflecting their more conceptually driven nature. Similarly, the answer length distribution highlights the explanatory style of Science responses, with answer lengths commonly falling between 10-40 words. This dichotomy provides a comprehensive testbed for evaluating an LLM's adaptability to different linguistic domains and reasoning types within a single, coherent educational framework. Although the two subjects differ in the structure and flow of their texts, they also exhibit notable similarities. As shown in Fig 2, both English and Science texts frequently reference entities such as PERSON, CARDINAL, DATE, and ORG. This suggests that while English and Science differ in textual structure and cognitive demands, they converge in their reference to certain key real-world entities which may aid in the development of cross-domain entity recognition and information extraction capabilities in language models.

4.2 Question-Answer Formulation

The formulation of the question-answer pairs is designed to mirror real-world educational exercises and assess a range of cognitive skills. As introduced in Section 3, questions are categorized as Factoid, Inferential, and Other. A quantitative breakdown of the 739 questions in our dataset reveals the distribution shown in Table 2(b). The prevalence of Factoid questions (55%) ensures a solid baseline for testing a model's core reading comprehension and information retrieval abilities. A significant portion of Inferential questions (35%) is critical for evaluating deeper reasoning, requiring models to connect ideas and make logical deductions that are not explicitly stated in the text. The Other category (10%) includes more complex questions that often require synthesizing information from multiple paragraphs, providing a challenge for advanced reasoning. This balanced distribution ensures that our evaluation rigorously tests models on a spectrum of tasks, from simple lookup to complex synthesis, which is essential for a versatile educational assistant.

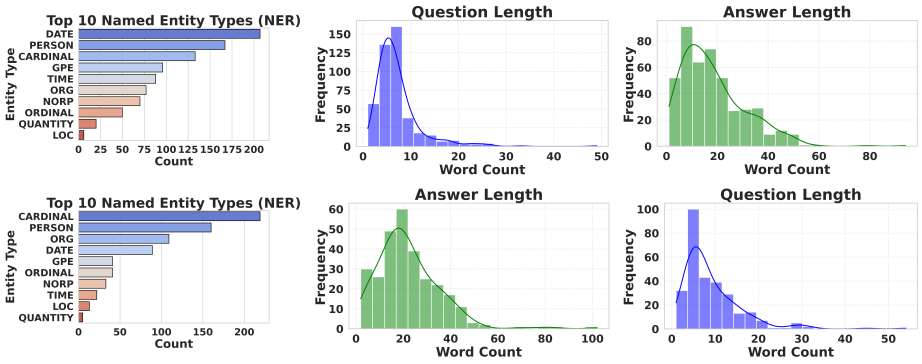

Fig. 2: Subject-wise analysis of the NCERT-QA dataset showing question&answer length distributions and top named entities, highlighting differences in complexity and focus between English and Science.

5 Methodology

In this study, we establish a pipeline founded on LLMs to exhibit the utilization of the NCERT-QA dataset. A comparative analysis is conducted on the performance of various LLMs, which include (gemma3:1b, llama3.2:3b, and nemotron-mini:4b) and high-end LLMs (Llama-4-Scout-17B-16E-Instruct and deepseek-r1-distill-llama-70b) with the objective of evaluating the enhancement in performance facilitated by this dataset, as shown in the Fig 1(b).

Each input to the LLM consists of a question paired with its corresponding context. We wrap this combination into a prompt to guide the model through a reasoning process. Prompt is stitched in a way that the model first determines the type of question—whether it is factoid, inferential, or others. Based on the identified question type, the model is then expected to follow appropriate reasoning steps using the given context to arrive at an accurate answer.

After the LLM generates an answer, we evaluate the answer against the ground truth answers provided in the NCERT-QA dataset. We employ a range of evaluation metrics to assess the quality of the responses. These metrics go beyond simple word overlap measures like ROUGE-L[10] to contextual relevance using BERTScore-f1[18].

We extend the comparison by evaluating the answer quality in terms of faithfulness and semantic relevance. We calculate semantic relevance as cosine similarity between sentence embeddings using SentenceTransformer models[14]:

$$\text{Faithfulness} = \frac{|tokens_{\text{answer}} \cap tokens_{\text{context}}|}{|tokens_{\text{answer}}|} \qquad \text{SemanticSim} = \frac{e_{\text{ref}} \cdot e_{\text{gen}}}{\|e_{\text{ref}}\| \, \|e_{\text{gen}}\|}$$

SemanticSim (SS) measures how well the generated answer aligns with the intended meaning of the reference answer, regardless of exact word matches. Unlike traditional lexical metrics such as ROUGE, semantic similarity evaluates

conceptual coherence between responses using cosine similarity between sentence embeddings.

Faithfulness (FF) measures the extent to which generated answers remain grounded in the provided source context, critical for educational applications where accuracy is paramount. This multilayered evaluation ensures a robust assessment of the LLM's ability to understand, reason, and respond accurately to NCERT-QA.

5.1 Prompt Strategies

We implement the framework detailed in 5 using diverse state-of-the-art prompting strategies [15, 5]. This implementation is designed to evaluate the underlying inference process utilizing our dataset.
The prompting strategies used are as follows:

1. Shot-based: We provide one example of each question category for the LLM to understand the answering pattern before asking the model to perform answer retrieval. This helps guide the model by showing the format and logic needed to come up with a response, without overloading it with lots of interleaved instructions. We extend this prompt to 3 and 5 shots by increasing the number of examples such that LLM can learn the diverse spectrum of each category.
2. CoT: In this strategy, we provide a method that encourages the model to reason step by step rather than jumping straight to the answer. We utilize this method to solve complex problems that require logical reasoning, intermediate steps, or multi-hop reasoning within the context.
3. Meta Prompt (MP): Our approach involves employing the large language model, Claude-Sonnet due to its structured meta framework, to construct a prompt or template that is used in our subsequent NCERT question and answer task. The goal of incorporating a meta_prompt is to evaluate whether Instructions by a meta-language model can enhance generating responses from the answering LLM, as opposed to relying solely on human-devised instructions.
4. Meta One-shot (MP-1S): In this work, we utilize a dual approach by integrating a meta prompt with one illustrative example of question types. To construct the prompt, we begin by drawing upon instructions sourced from a meta-level language model and subsequently embed an instance for each category of question within the prompt as examples.

6 Results and Analysis

Our experimental results are presented in three parts. We first establish the unique value of the NCERT-QA dataset, then analyze model and prompt performance on our task, and finally discuss the practical implications for deployment.
Part 1: The Unique Challenge of Curriculum-Aligned QA: To demonstrate that answering curriculum-specific questions is a distinct challenge that

general-purpose models cannot solve from pre-trained knowledge alone, we conducted a baseline experiment. We evaluated Llama4-Scout-17B on the well-known SQuAD dataset and on our NCERT-QA dataset without providing the textbook context. This setup forces the model to rely solely on its internal knowledge.

The results in Table 3 are unequivocal. The model performs exceptionally well on SQuAD, a general-knowledge benchmark and possibly used for LLM training, but its performance collapses on NCERT-QA questions when deprived of the textbook context. The F1 score plummets, indicating an inability to generate precise answers. This is because NCERT questions are deeply tied to the specific phrasing, narratives, and vocabulary of the curriculum, which is not adequately represented in the model's general training data. This experiment confirms our core hypothesis: a specialized dataset and prompting with the right context are not just helpful but essential for building a reliable educational assistant.

Table 3: Baseline performance of Llama-4-Scout-17B on SQuAD vs. NCERT-QA (no context), showing the performance drop and the need for a curriculum-specific, context-aware approach.

Metric	SQuAD	NCERT QA (Eng)	NCERT QA (Sci)	NCERT-QA (Overall)
F1	**0.894**	0.32	0.47	0.395
Semantic Sim.	**0.920**	0.78	0.88	0.830

We further assessed the models using each prompt type with and without incorporating contextual information on NCERT-QA. As shown in Tables 4a and 4b, there is a noticeable improvement in metric scores when context is included during retrieval. This highlights the critical role of external knowledge in enhancing model performance and underscores the significance of our proposed dataset.

Part 2: In-depth Analysis of Model and Prompt Performance: Having established the need for our approach of prompting with the right context, we now analyze the performance of various models and prompting strategies on the NCERT-QA dataset with the full context provided. As shown in the Table 4b, there is a clear correlation between model size and performance. The larger models, Llama4-Scout-17B and DeepSeek-70B, significantly outperform their smaller, open-source counterparts. Between the two models, Llama4-Scout-17B shows higher performance, and among the small open-sourced counterparts, Llama3.2-3B leads largely in F1 and faithfulness.

The choice of prompting strategy also had a profound impact. The meta oneshot prompt emerged as the most consistently effective strategy, delivering the best F1 scores for most models (See Table 4b) in both English and Science subsets. This suggests that a high-quality, machine-generated instruction combined with a single, clear example offers an optimal balance of guidance. Conversely, the Chain-of-Thought strategy yielded surprisingly poor results, partic-

ularly on the Faithfulness metric across models(e.g., 0.532 for Llama4-Scout; See Appendix A). This indicates that encouraging detailed step-by-step reasoning for this task caused the models to "hallucinate" details beyond the provided context. Comparing performance across subjects, models consistently scored slightly higher on the Science dataset than the English dataset. For instance, Llama4-Scout achieved a faithfulness score of 0.87 on Science, while its performance on English was closer to 0.85. This suggests that the factual, descriptive nature of the science texts is more amenable to prompting with context than the inferential and narrative complexities of the English literary texts.

Table 4: Comparison of Small (S) and Large (L) models: (a) Performance without contextual data; (b) Best model and prompt type on English and Science with context.

(a)

NCERT	Model	F1	B-F1	R-L	SS
English	Llama4 (L)	0.32	0.81	0.27	0.78
	DS (L)	0.13	0.78	0.07	0.77
	Gemma3 (S)	0.26	0.79	0.22	0.77
	Nemo (S)	0.24	0.79	0.19	0.77
	Llama3.2 (S)	0.18	0.76	0.14	0.70
Science	Llama4 (L)	0.47	0.81	0.41	0.88
	DS (L)	0.46	0.88	0.40	0.89
	Gemma3 (S)	0.27	0.81	0.23	0.79
	Nemo (S)	0.25	0.89	0.20	0.79
	Llama3.2 (S)	0.19	0.78	0.15	0.72

(b)

NCERT	Model	Prompt	F1	B-F1	R-L	SS	FF
English	Llama4 (L)	MP-1S	**0.46**	0.86	0.40	0.86	**0.85**
	DS (L)	MP	0.45	0.86	0.39	0.87	0.79
	Llama3.2 (S)	MP-1S	**0.40**	0.84	0.35	0.83	**0.81**
	Nemo (S)	MP	0.36	0.83	0.28	0.82	0.76
	Gemma3 (S)	MP-1S	0.35	0.83	0.31	0.81	0.80
Science	Llama4 (L)	MP-1S	**0.47**	0.88	0.41	0.88	**0.87**
	DS (L)	MP-1S	0.46	0.87	0.40	0.89	0.81
	Llama3.2 (S)	MP-1S	**0.41**	0.86	0.35	0.85	**0.83**
	Nemo (S)	MP-1S	0.37	0.85	0.29	0.84	0.78
	Gemma3 (S)	MP-1S	0.36	0.85	0.32	0.83	0.82

Part 3: Practical Implications for Real-World Deployment While larger models deliver higher performance, a practical educational tool must also be efficient and cost-effective. In this section, we analyze the trade-off between performance and inference speed to identify the most viable model for deployment in a school setting. Fig: 3 (a) and (b) compares our two top-performing models. While DeepSeek-70B holds a slight edge in semantic similarity, Llama4-Scout-17B achieves a better F1 score and is significantly more faithful, all while being over 6 times faster. An average inference time of 2 seconds is well within the acceptable range for a real-time interactive assistant, whereas a 13 second wait is likely too slow for an engaging student experience. This analysis demonstrates that Llama4-Scout-17B is not just a compromise but arguably the superior choice for this application. It delivers state-of-the-art results on the metrics that matter most (accuracy and faithfulness) with the efficiency required for practical, cost-effective deployment in schools, where computational resources may be limited. Additionally as can be seen from Fig. 3 (c), among the smaller open-source models, Gemma3-1B and Llama3.2-3B, exhibit lower latency, with Llama3.2-3B offering a more optimal trade-off between latency and overall performance.

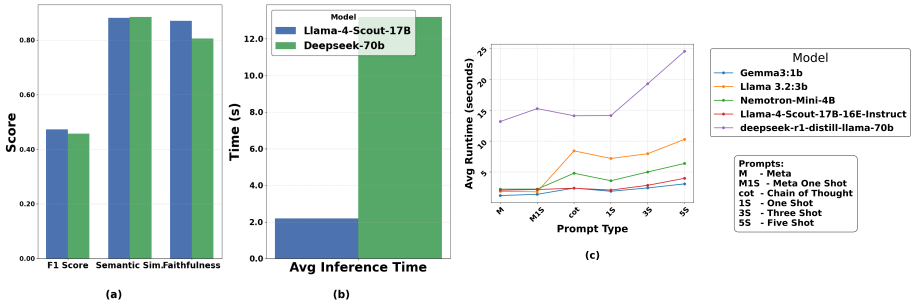

Fig. 3: Performance–efficiency trade-off: (a,b) Llama-4-Scout matches DeepSeek-70B performance at far lower inference time. (c) Average runtime per prompt type shows meta and meta-one-shot with reduced latency; Llama models and Gemma3-1B are fastest, DeepSeek-70B slowest.

7 Conclusion

In this study, we introduce the NCERT-QA dataset to bridge the gap between curriculum content and educational Q&A systems. Our categorization of question types highlights the diversity present in the dataset, ensuring a comprehensive representation of curriculum-based queries. This work marks an initial step toward expanding the dataset across additional grades and subjects. Through extensive evaluations across multiple models and prompting strategies, we emphasize the vital role of high-quality, curriculum-aligned data in enhancing the accuracy and relevance of responses. Our findings demonstrate that incorporating contextual knowledge significantly improves model performance, reinforcing the importance of structured retrieval and carefully curated datasets such as NCERT-QA. Our analysis of models of varying scales demonstrates the potential of smaller open-source models for practical deployment in resource-constrained environments. Our dataset and pipeline, PustakAI, establishes a foundation for building more robust and scalable educational AI systems that are closely aligned with academic curriculum. More detailed observations and exhaustive analysis will be provided on the ArXiv version of this paper.

A Appendix

A.1 Comparison of Model Performance Under Contextual and Non-Contextual Prompts

We visualize the comparative performance of each model across the employed prompting strategies. As illustrated in the graphs below, Llama-4-Scout combined with the Meta one-shot prompt consistently achieves superior performance across both the English and Science datasets. Furthermore, the inclusion of contextual information exerts a positive effect, yielding a notable performance improvement relative to the no-context condition.

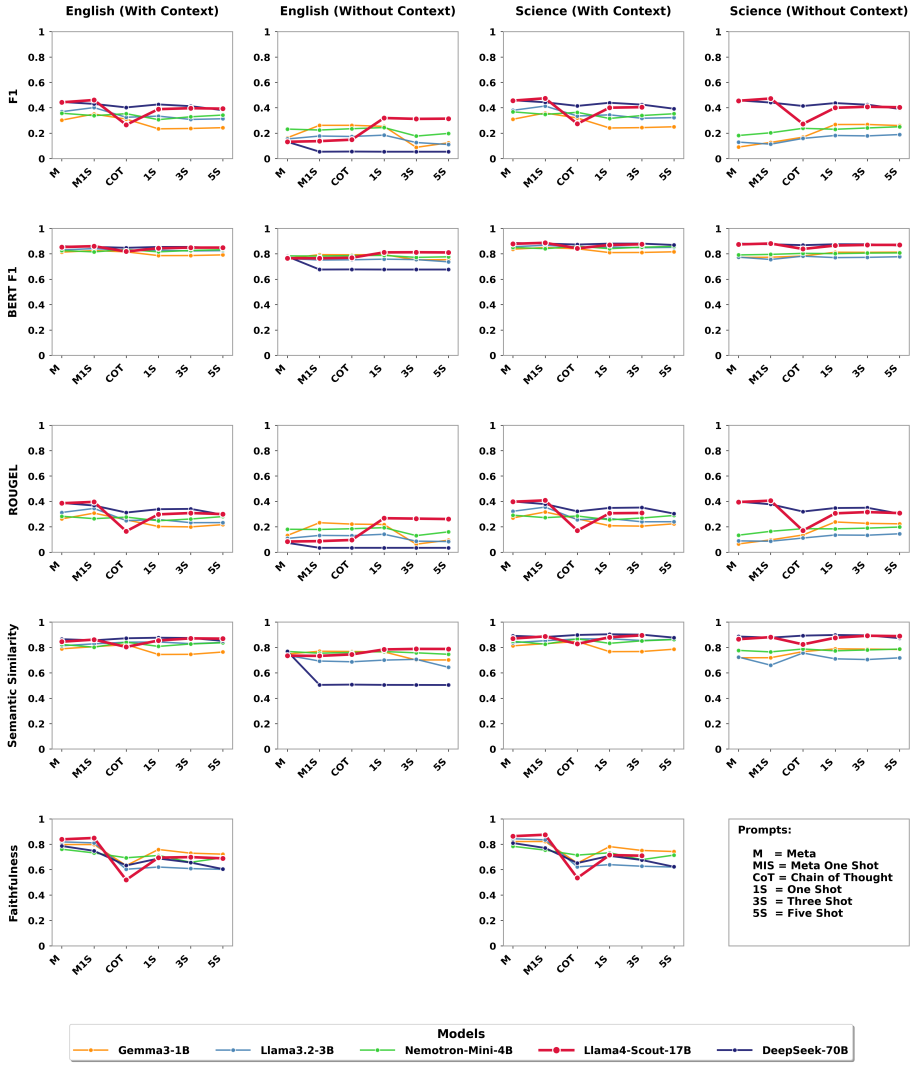

Fig. 4: Performance of models on contextual vs. non-contextual prompts. Faithfulness metric is absent for non-contextual prompts since faithfulness measures the match between the provided context and the LLM's generated answer, which cannot be computed without context.

References

1. Alawwad, H.A., Zafar, A., Jamal, A., Alhothali, A., Alharbi, A., Kawsar, F.: Evaluating multimodal large language models on educational textbook question answering. arXiv preprint arXiv:2506.21596 (2025)
2. Baumgärtner, T., Briscoe, T., Gurevych, I.: PeerQA: A scientific question answering dataset from peer reviews. In: Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2025)
3. Bhambri, S., Verma, M., Kambhampati, S.: Do think tags really help LLMs plan? a critical evaluation of ReAct-style prompting. Transactions on Machine Learning Research (2025)
4. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O.: Think you have solved question answering? try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457 (2018)
5. Dang, H., Mecke, L., Lehmann, F., Goller, S., Buschek, D.: How to prompt? opportunities and challenges of zero-and few-shot learning for human-ai interaction in creative applications of generative models. arXiv preprint arXiv:2209.01390 (2022)
6. Fernandez, N., Scarlatos, A., Lan, A.: SyllabusQA: A course logistics question answering dataset. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 10344–10369 (2024)
7. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: Advances in Neural Information Processing Systems. vol. 35, pp. 22199–22213 (2022)
8. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: RACE: Large-scale reading comprehension dataset from examinations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 785–794 (2017)
9. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Ott, M., Chen, D., Yih, W.t., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474 (2020)
10. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/
11. Lin, S., Hilton, J., Evans, O.: TruthfulQA: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958 (2021)
12. Liu, Z., Yin, S.X., Goh, D.H.L., Chen, N.F.: COGENT: A curriculum-oriented framework for generating grade-appropriate educational content. In: Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (2025)
13. Razafinirina, M.A., et al.: Pedagogical alignment of large language models (LLM) for personalized learning: A survey, trends and challenges. Journal of Intelligent Learning Systems and Applications **16**(04), 448 (2024)
14. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
15. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

16. Xu, Y., Yao, B., Wu, T., Zhang, Z., Yu, M., Ma, X., Wang, D., Hou, Y., Peng, N., Li, T.J.J., et al.: Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 835–853 (2022)
17. Zhang, C., et al.: Meta-prompting: A structure-oriented approach for large language models. arXiv preprint (2024)
18. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)