EduAdapt: A Question Answer Benchmark Dataset for Evaluating Grade-Level Adaptability in LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) are transforming education by answering questions, explaining complex concepts, and generating content across a wide range of subjects. However, despite strong performance on academic benchmarks, they often fail to adapt responses to students' grade levels. This is a critical need in K–12 education, where age-appropriate vocabulary and explanation are essential for effective learning. Existing models frequently produce outputs that are too advanced or vague for younger learners, and there are no standardized benchmarks to evaluate their ability to adjust across cognitive and developmental stages. To address this gap, we introduce a benchmark of nearly 48k grade-labeled QA pairs across 9 science subjects, spanning Grades 1–12 and grouped into four grade levels. We evaluate a diverse set of open-source LLMs and find that while larger models generally perform better, they still struggle with generating suitable responses for early-grade students (Grades 1–5). Our work presents the first dataset and evaluation framework for assessing grade-level adaptability in LLMs, aiming to foster more developmentally aligned educational AI systems through better training and prompting strategies. EduAdapt's code and datasets are open-sourced and publicly available at [URLredacted].

1 Introduction

Recent research has shown that LLMs can perform at a student level on standardized tests across subjects like mathematics, physics, and computer science, often achieving high accuracy on both multiple-choice and open-ended questions (OpenAI et al., 2024). For example, studies demonstrate that tools like ChatGPT are capable of generating logically coherent responses that reflect a strong grasp of subject matter across a wide range of disciplines (Susnjak, 2022). While these abilities are impressive, they primarily benefit older students and professionals. As highlighted by (Rooein et al., 2023), LLMs often fail to adapt their explanations to suit different grade levels, providing answers that may be too complex for younger students or overly simplified for advanced learners. This lack of grade-specific adaptability is a consistent limitation among many state-of-the-art LLMs. Even when explicitly prompted, most models struggle to adjust their language, tone, and complexity to match the cognitive level of different age groups. This is particularly concerning, given the high level of digital engagement among children. According to UNICEF, one in three internet users globally is a child (Keeley and Little, 2017), and children aged 8–12 spend over five hours per day on screens on average (Rideout et al., 2022). This presents a major opportunity to enrich learning through AI,

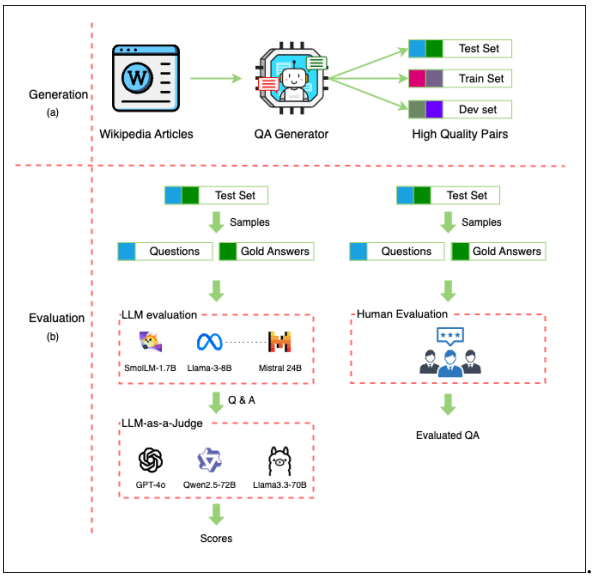

Figure 1: Overview of the full methodology pipeline. The process consists of two main stages: (a) **Generation**, where Wikipedia articles are used to generate high-quality QA pairs; and (b) **Evaluation**, involving human verification of dataset quality followed by models evaluation on the test set.

but also a risk if content is not age-appropriate or understandable. The key concerns include a lack of contextual relevance for younger users (Nayeem and Rafiei, 2024; Seo et al., 2024) and difficulties in maintaining the right level of lexical simplicity across grade levels (Valentini et al., 2023).

To address these challenges, researchers have begun developing specialized language models such as KidLM (Nayeem and Rafiei, 2024), an encoder-based model trained with child-appropriate data and objectives to improve the readability, safety, and developmental suitability of language representations for children. KidLM is trained on a curated dataset of child-friendly texts and employs innovative techniques like stratified masking to improve vocabulary relevance while minimizing the reinforcement of stereotypes. These domain-specific efforts highlight the need for LLMs that are not only accurate but also context-aware and adaptive to the diverse educational needs of younger audiences. This research tackles a key challenge: the inability of current LLMs to effectively adapt their responses to students at different grade levels.

To bridge this gap, we developed a high-quality benchmark dataset spanning Grades 1 through 12, comprising nearly 48k question–answer pairs across 9 educational subjects. Based on the K–12 framework, we reorganized the grades into four finer developmental levels, Grades 1–2, 3–5, 6–8, and 9–12, to better capture the cognitive and linguistic progression of learners. This stratification enables more precise modeling and evaluation of educational content. The dataset design is guided by the Next Generation Science Standards (NGSS) (NGSS Lead States, 2013), ensuring that questions align with appropriate cognitive skill levels, from basic recall to higher-order reasoning. We evaluated multiple LLMs of varying sizes and found that even the most advanced models struggled to adapt their outputs effectively across grade levels. To the best of our knowledge, this is the first benchmark specifically developed to evaluate grade-level adaptability of LLMs across nine subject areas for full K–12 educational system.

2 Methodology

The pipeline for this study consists of two main stages: the **Generation Process** and the **Evaluation Process**. The **Generation Process** begins with extracting clean, domain-specific text from Wikipedia articles, followed by the generation of question-answer (QA) pairs tailored to each educational level. To ensure quality, a self-reflection mechanism (Renze and Guven, 2024) is applied, enabling the model to evaluate and refine its outputs based on pedagogical criteria. The **Evaluation Process** involves two key steps. First, human reviewers assess the quality and grade-level appropriateness of a subset of the dataset to ensure educational validity. Second, the validated test split is used to evaluate various open-source LLMs on their ability to generate accurate, grade-aligned responses. This includes standard accuracy metrics for multiple-choice questions and an LLM-as-a-judge framework for open-ended QA evaluation (Zheng et al., 2023). An overview of the complete methodology is shown in Figure 1, summarizing the process from data collection to evaluation.

2.1 Stage 1: Generation Process

The first stage of our pipeline involves extracting and cleaning Wikipedia articles, which serve as input for the question-answer generation process, as shown in part (a) of Figure 1. Each article is processed by the QA Generator, as shown in Figure 2, to produce grade-appropriate educational QA pairs. This is implemented using the text-generation module of the Distilabel framework (Argilla.io, 2025), which supports iterative refinement through AI-generated feedback, enhancing both data quality and model behavior.

2.1.1 Content Collection from Wikipedia

We began by collecting source material from Wikipedia using Wikipedia dumps, focusing on articles related to key academic disciplines. Specifically, we targeted nine subject areas: (1) **Chemistry**, (2) **Computer Science**, (3) **Meteorology**, (4) **Ecology**, (5) **Geology**, (6) **Biology**, (7) **Physics**, (8) **Medicine**, (9) **Geography**. These fields were carefully selected to provide broad coverage across scientific and technical domains, ensuring the dataset is rich and versatile for different educational contexts. Once the relevant articles were collected, we applied a series of cleaning and preprocessing steps to prepare the data for downstream use. This process resulted in a clean, well-structured dataset of text segments, categorized by subject and ready for educational alignment task.

2.1.2 Question-Answer Generation

We leveraged Distilabel's `text-generation` module to generate QA pairs from our curated content,

employing Phi-4 (Abdin et al., 2024), a 14B parameter model developed by Microsoft. Phi-4 is specifically designed for educational and reasoning tasks, with training data curated to span a wide range of educational levels, from elementary to graduate. It demonstrates strong performance across benchmarks and notably outperforms its teacher model, GPT-4o, on several tasks despite its smaller size. Its alignment with educational content, high efficiency, and strong output quality under limited computational resources made Phi-4 an ideal choice for our QA generation pipeline.

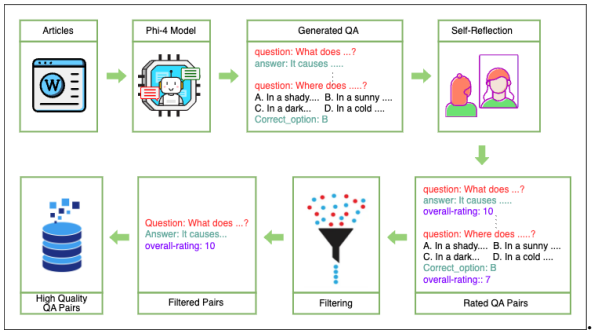

Figure 2: **QA Generator:** A pipeline that generates QA pairs, applies self-reflection for quality assessment, and filters results to create a high-quality dataset.

To ensure each QA pair aligned with the language and cognitive abilities of students, we grouped K–12 into four grade bands: Grades 1–2, 3–5, 6–8, and 9–12. For each group, we designed tailored prompts based on NGSS guidelines (NGSS Lead States, 2013), reflecting students' comprehension and reasoning skills. These prompts underwent multiple rounds of refinement to ensure age-appropriateness and effectiveness, and were then integrated into a structured QA generation pipeline. During QA generation using `phi-4`, we experimented with various model settings to optimize output quality. The best results were achieved with a temperature of 0.3 and a top_p value of 0.9. To accelerate the generation process, we hosted the model using vLLM (Kwon et al., 2023), which enabled faster and more efficient inference. The finalized grade-specific prompts are shown in Listings 1 through 4.This phase generated approximately **166k** QA pairs across all subjects and grade levels.

2.1.3 Self-Reflection for Quality Assessment

Following the generation of QA pairs across different grades and subjects, we applied a self-reflection mechanism to ensure high-quality data. This step used the Phi-4 model to evaluate its own generated pairs, aiming to retain only those that met rigorous pedagogical and linguistic standards based on NGSS guidelines. Prior work has shown that using the same model for both generation and evaluation can be highly effective (Renze and Guven, 2024). We implemented a customized UltraFeedback-style reflection pipeline (Cui et al., 2024) using the Distilabel framework. Building on the original UltraFeedback framework, which assesses responses across multiple dimensions, we adapted it for evaluating educational QA. We designed separate prompts for each grade group (Grades 1–2, 3–5, 6–8, and 9–12), ensuring consistency in evaluation criteria such as language appropriateness, grade alignment, relevance, clarity, and subject fit, each tailored to the developmental stage of the respective grade level. Each QA pair received a score from 1 to 10 on each criterion, and an average was computed as the overall rating, following the **overall-rating** scheme of UltraFeedback. This holistic scoring avoids bias toward any single criterion. Based on our analysis, we found that QA pairs with an average score of 8 or higher consistently demonstrated high quality, so only these were retained. We have used the same temperature and top_p as we used in QA generation. The full prompts and criteria are shown in Listings 5 to 8. Out of the initial **166k** QA pairs, only **48,123** were retained after an aggressive filtering process based on strict quality and grade-level appropriateness criteria. The final dataset was split into 60% for training, 20% for development, and 20% for testing. Detailed statistics are provided in Table 3.

2.2 Stage 2: Evaluation Process

The second stage of our pipeline evaluates the quality and effectiveness of the generated QA dataset using two complementary approaches: (1) Human evaluation of QA pairs (approximately 10% of the test set) to verify quality, appropriateness, and grade-level alignment; and (2) Model-based evaluation, where a diverse set of instruction-tuned LLMs are evaluated on the validated gold test set to assess their alignment with grade-specific requirements.

2.2.1 Human Evaluation

To ensure the quality of our generated dataset, we carried out a human evaluation on a randomly selected subset of **1,000** QA pairs. This sample covered all nine scientific subjects and represented roughly 10% of the test set, helping us verify both

the content's pedagogical soundness and grade-level suitability of the content. We hired three expert reviewers with backgrounds in educational content development via the Upwork platform to independently evaluate the QA pairs. Each annotator was compensated at a rate of $10 per 100 words for reviewing the complete set of 1k QA pairs. They followed a detailed evaluation criteria defined in Listings 5 to 8, rating each pair on a 1–10 scale across several criteria: language appropriateness, grade alignment, relevance, clarity, and subject-fit. For each QA pair, the overall score was calculated by averaging the ratings across all criteria. These overall scores were then used to evaluate quality of dataset at the grade level by aggregating them accordingly. Table 1 presents the average scores from each reviewer as well as the overall average for each grade group.

Grade Level	Human 1	Human 2	Human 3	Average
Grade 1 and 2	7.18	7.71	8.19	7.69
Grade 3 to 5	8.14	7.56	8.32	8.00
Grade 6 to 8	8.20	7.58	8.63	8.14
Grade 9 to 12	9.00	8.86	8.71	8.86

Table 1: Human evaluation scores across grade levels on sample of testset.

The human evaluation results provide a comprehensive view of the dataset's quality, based on expert, human-centered judgment. To further validate the reliability of these evaluations, we calculated inter-annotator agreement using Fleiss' Kappa (Fleiss, 1971) across all three reviewers. Specifically, we computed the Kappa scores per grade level by averaging agreement scores across all nine scientific fields. Table 2 presents the average Fleiss' Kappa scores for each grade level. The consistently high values indicate strong agreement among the reviewers, reinforcing the trustworthiness of the human ratings as a benchmark for assessing dataset quality. In summary, our detailed human evaluation and strong inter-rater agreement confirm the dataset's reliability and highlight key areas for future improvement.

2.2.2 Model Evaluation

We conducted a comprehensive evaluation of open-source instruction-tuned language models to understand how effectively they adapt to different grade levels. Using our curated test set of approximately **9,624** QA pairs (both mcq and open-ended questions), spanning a wide range of subjects and

Grade Level	Fleiss' Kappa
Grade 1 and 2	0.668
Grade 3 to 5	0.706
Grade 6 to 8	0.741
Grade 9 to 12	0.860

Table 2: Average Fleiss' Kappa scores across grade levels, over all scientific fields.

educational levels, we assessed model performance with a focus on grade-level appropriateness. A detailed breakdown of the test set distribution is shown in Table 8.

To capture the effects of model architecture and scale, we evaluated a diverse set of language models, including Qwen2.5 models at 1.5B, 3B, 7B, and 14B parameters (Team, 2024), SmolLM-1.7B (Allal et al., 2024), Gemma-2B-it (Team et al., 2024), LLaMA3.2-3B and LLaMA3-8B (AI, 2024; AI@Meta, 2024), and the larger Mistral-Small-24B (AI, 2025). This model lineup enabled us to analyze the impact of scaling on accuracy, linguistic suitability, and educational relevance across various grade levels. Each model was evaluated using the same set of questions, with prompts explicitly tailored to indicate the intended grade level of the target learners. The prompt used during evaluation is shown in Listing 10.

For open-ended questions, models received only the question text and were expected to generate an answer aligned with the intended grade level. For multiple-choice questions, the full question and answer options were provided, and models were required to select the correct choice. To evaluate outputs, we used two approaches. For MCQs, accuracy was computed by comparing the model's selected option with the correct answer. For open-ended questions, we adopted an LLM-as-a-judge framework using three independent judges, including both proprietary and open-source models: **Qwen2.5-72B** (Qwen et al., 2025), **LLaMA3.3-70B** (Grattafiori et al., 2024), and **GPT-4o** (OpenAI et al., 2024). Each judge received the question, reference answer, and model-generated response, and scored it across multiple qualitative criteria on a 1–10 scale. This setup is illustrated in part (b) of Figure 1, and the prompt used is shown in Listing 9.

Results of these evaluations are presented in Table 5. By evaluating a diverse set of models, we analyzed how differences in architecture and size

influence the ability to produce accurate, grade-aligned, and pedagogically sound responses. This also revealed trade-offs between model size, computational cost, and response quality in educational settings. Table 9 summarizes the Hugging Face identifiers and roles of each model used in our pipeline.

3 Dataset

This section presents the dataset we developed to evaluate LLM adaptability across grade levels. It consists of question-answer pairs spanning Grades 1 to 12 across 9 academic subjects. Each subject is organized into four grade bands following the K–12 system: Grades 1–2, 3–5, 6–8, and 9–12. This structure captures a broad range of student proficiency levels, supporting multi-level educational evaluation. The QA pairs were designed to be accurate and grade-appropriate. Listings 11 to 46 present sample testset examples for each subject across all grade levels.

3.1 Dataset Statistics

The final dataset comprises **48,123** high-quality question-answer (QA) pairs, structured to support the development and evaluation of LLMs in educational contexts. It is divided into three subsets: **28,875** for training, **9,624** for development, and **9,624** for testing, enabling both robust benchmarking and fine-tuning for grade-specific tasks. The dataset features a balanced mix of open-ended and multiple-choice questions (MCQs). Table 3 summarizes the overall distribution across subjects and grade levels, ensuring fair representation. Detailed splits by subject and grade for each subset are provided in Tables 6, 7, and 8. This structure facilitates rigorous experimentation for educational AI applications. We intend to publicly release our dataset, enabling the research community to build upon it and further evaluate grade-level adaptability in language models.

4 Experiments and Results

Building on the methodology outlined in Section 2, this section benchmarks various LLMs using our curated educational dataset. The primary objective is to evaluate how well current models adapt their responses to different grade levels, addressing student-specific comprehension and developmental needs. We tested models of varying sizes, from 1.5B to 24B parameters, on the test split. Each model was assessed on its ability to generate responses aligned with the cognitive and developmental stage of the target grade. These evaluations highlight both the capabilities and current limitations of LLMs in education, emphasizing the need for models that are not only factually accurate but also pedagogically aligned.

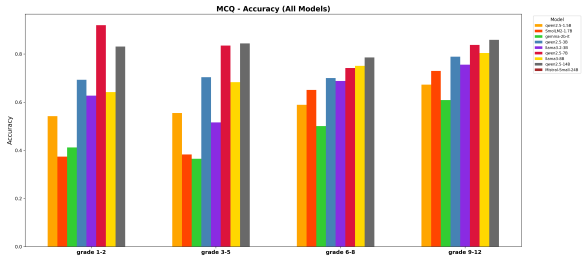

Figure 3: Accuracy on MCQs for the test split across grade levels

Results The model-wise average accuracy for multiple-choice questions (MCQs) across grade levels is reported in Table 4 and illustrated in Figure 3. Larger models such as Qwen2.5-14B and Mistral-Small-24B consistently achieve higher accuracy across all grades. Smaller models (1.5B–3B) perform poorly, particularly on lower grade levels (Grades 1–5), where their accuracy ranges between 50–60%. Their performance improves to 70–80% on higher grades, indicating difficulty in adapting to simpler, age-appropriate content. Mid-sized models like Qwen2.5-7B and LLaMA3-8B perform significantly better than smaller models and are often comparable to large models. Notably, the Qwen series consistently outperforms other models of the same size, with Qwen2.5-14B performing on par with the much larger Mistral-Small-24B. Although MCQs are relatively constrained and should be easier to answer, many models, especially smaller ones, still underperform. This highlights the diversity and challenge of our dataset, demonstrating gaps in current LLMs' ability to handle grade-specific educational content. To evaluate open-ended question answers, we employed three LLMs, Qwen2.5-72B, LLaMA3.3-70B, and GPT-4o, as independent judges. Each judge rated model outputs per grade level, and the average score per judge was computed. Table 5 presents these scores across grade levels, and the trends are visualized in Figure 4. The results show that all models, regardless of size, struggle to generate grade-appropriate responses for lower grades (Grades 1–5) compared

	Grade Levels	Biology	Physics	Chemistry	Computer Science	Ecology	Geography	Geology	Medicine	Metrology
QA Count	1 and 2	456	67	133	162	840	1250	256	146	92
	3 to 5	1004	100	263	438	1510	2086	236	379	124
	6 to 8	344	89	144	302	759	463	299	175	121
	9 to 12	1407	2475	2706	2263	1159	687	1248	1621	1660
MCQ Count	1 and 2	125	47	75	40	193	252	40	53	37
	3 to 5	393	43	82	111	736	913	50	85	100
	6 to 8	409	95	150	311	925	463	417	187	118
	9 to 12	1125	1972	2421	1983	1105	592	1323	1251	1851
Total		**5263**	**4888**	**5974**	**5410**	**7227**	**6706**	**3869**	**3897**	**4103**

Table 3: **Full Dataset**: Distribution of question-answer pairs across all subjects and grade levels in the full dataset.

Model	Grade 1-2	Grade 3-5	Grade 6-8	Grade 9-12
qwen2.5-1.5B	0.542	0.555	0.589	0.673
SmolLM-1.7B	0.374	0.383	0.651	0.730
gemma-2b-it	0.412	0.365	0.501	0.609
qwen2.5-3B	0.693	0.704	0.700	0.789
llama3.2-3B	0.627	0.516	0.688	0.756
qwen2.5-7B	**0.920**	0.835	0.742	0.838
llama3-8B	0.642	0.683	0.751	0.804
qwen2.5-14B	0.831	0.844	0.786	0.859
mistral24B	0.862	**0.858**	**0.805**	**0.863**

Table 4: Model-wise accuracy on MCQ questions across grade levels

to higher grades. Larger models like Qwen2.5-14B and Mistral-Small-24B consistently outperform others across all levels but still exhibit weaker performance on early-grade content. Mid-sized models, such as Qwen2.5-7B and LLaMA3-8B, perform slightly below the large models and follow a similar trend of reduced effectiveness in lower grades. Smaller models (1–3B) perform noticeably worse than both mid-sized and large models across all grade levels. However, their performance gradually improves as the grade level increases, indicating better alignment with higher, grade content despite overall lower effectiveness.

Analysis These findings confirm a critical gap: current LLMs are better aligned with content for older students and less effective at adapting to early-grade needs. Our dataset is the first benchmark to comprehensively cover the full K–12 range, enabling systematic evaluation of grade-level adaptability.

In addition to human and LLM-based evaluations, we assessed the quality of open-ended question-answer (QA) pairs using standard automated metrics, including BLEU (Papineni et al., 2002), ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) (Lin, 2004), and BERTScore (Zhang et al., 2020). These metrics are commonly used in natural language generation tasks to evaluate surface-level similarity between generated and reference texts. Our analysis revealed that even smaller models performed reasonably well on metrics such as ROUGE and BLEU, which focus on n-gram and lexical overlap. However, these models still exhibited lower semantic accuracy and weaker alignment with the intended educational goals, particularly for younger grade levels.

Interestingly, while BERTScore often reported high similarity values, frequently exceeding 90%, manual inspection showed that it was not a reliable indicator of answer quality in educational contexts. The metric tended to assign high scores to answers that were semantically incorrect, incomplete, or misaligned with the cognitive needs of the target grade level. This disconnect between surface-level similarity and pedagogical validity calls into question the applicability of such metrics for educational QA evaluation. Similarly, ROUGE and BLEU despite of their popularity, showed limitations when applied to our grade-specific dataset. These metrics prioritize lexical matching and n-gram overlap, which do not adequately capture the depth, correctness, or developmental appropriateness required for high-quality educational responses. In summary, our findings highlight the inadequacy of standard automated metrics for evaluating educational QA, especially across varying grade levels where cognitive and linguistic expectations differ significantly. Metrics like Accuracy (for multiple-choice questions) and LLM-as-a-Judge scoring (for open-ended responses) provide more pedagogically meaningful assessments and are better suited for evaluating the quality and appropriateness of educational content.

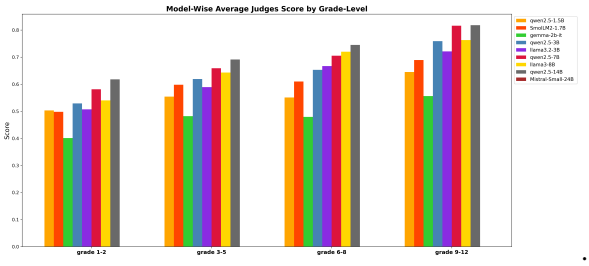

Figure 4: Average model scores across grade levels for open-ended QA tasks, evaluated independently by GPT-4o, Qwen2.5-72B, and LLaMA3.3-70B

Implementation Details All generation and evaluation experiments were conducted using NVIDIA RTX A6000 GPUs (48GB). Large models like Mistral-Small-24B, Qwen2.5-72B, and LLaMA3.3-70B were run on 2 GPUs, while the remaining models used a single GPU. Throughout the model evaluations and the LLM-as-a-judge setup, we used a temperature of 0.3, as it consistently produced more stable and reasonable outputs compared to other settings

5 Literature Review

This section reviews benchmark datasets developed to evaluate LLMs on educational tasks, focusing on math-centric, interdisciplinary, and multilingual evaluations. These benchmarks are crucial for assessing model performance and guiding the development of AI systems that support diverse learning needs.

In mathematics, several high-quality datasets have been introduced to evaluate LLM reasoning across grade levels. GSM8K (Cobbe et al., 2021) offers 8.5K human-written grade school problems (up to grade 8), emphasizing verification over scaling. For high school, the MATH dataset (Hendrycks et al., 2021b) provides 12.5K competition-style problems in algebra, geometry, and number theory, with step-by-step solutions. To broaden coverage, Dolphin18K (Huang et al., 2016) compiles real-world math questions from community Q&A forums using automated equation extraction and annotations. DRAW-1K (Upadhyay and Chang, 2017) emphasizes evaluating derivations, noting that correct answers can stem from flawed logic. Math23K (Wang et al., 2017), with over 23K elementary word problems, showed that deep learning outperforms statistical methods in this domain. MathQA (Amini et al., 2019) introduces multi-choice math word problems across disciplines, with interpretable programs to guide reasoning.

Beyond mathematics, several benchmarks target specialized and interdisciplinary domains. MedMCQA (Pal et al., 2022) includes 194K multiple-

Models	Gpt-4o	Qwen2.5-72B	Llama3.3-70B	Average
Grade 1–2				
qwen2.5-1.5B	0.436	0.542	0.532	0.503
SmolLM-1.7B	0.425	0.536	0.533	0.498
gemma-2b-it	0.335	0.434	0.435	0.401
qwen2.5-3B	0.440	0.596	0.552	0.529
llama3.2-3B	0.436	0.537	0.548	0.507
qwen2.5-7B	0.499	0.634	0.609	0.580
llama3-8B	0.477	0.575	0.569	0.540
qwen2.5-14B	0.541	0.657	0.655	0.617
mistral24B	0.553	0.655	0.659	0.622
Grade 3-5				
qwen2.5-1.5B	0.484	0.595	0.584	0.554
SmolLM-1.7B	0.527	0.632	0.634	0.597
gemma-2b-it	0.418	0.514	0.513	0.481
qwen2.5-3B	0.530	0.679	0.649	0.619
llama3.2-3B	0.506	0.622	0.639	0.589
qwen2.5-7B	0.584	0.701	0.693	0.659
llama3-8B	0.570	0.677	0.683	0.643
qwen2.5-14B	0.631	0.722	0.720	0.691
mistral24B	0.646	0.744	0.737	0.709
Grade 6-8				
qwen2.5-1.5B	0.489	0.589	0.574	0.550
SmolLM-1.7B	0.547	0.651	0.632	0.61
gemma-2b-it	0.396	0.501	0.541	0.479
qwen2.5-3B	0.567	0.700	0.692	0.653
llama3.2-3B	0.596	0.688	0.717	0.667
qwen2.5-7B	0.614	0.742	0.759	0.705
llama3-8B	0.660	0.751	0.749	0.720
qwen2.5-14B	0.651	0.786	0.798	0.745
mistral24B	0.696	0.805	0.809	0.770
Grade 9-12				
qwen2.5-1.5B	0.584	0.673	0.678	0.645
SmolLM-1.7B	0.596	0.730	0.742	0.689
gemma-2b-it	0.488	0.609	0.572	0.556
qwen2.5-3B	0.691	0.789	0.797	0.759
llama3.2-3B	0.676	0.756	0.732	0.721
qwen2.5-7B	0.754	0.838	0.856	0.816
llama3-8B	0.674	0.804	0.812	0.763
qwen2.5-14B	0.768	0.859	0.827	0.818
mistral24B	0.794	0.863	0.858	0.838

Table 5: Model-wise average scores for open-ended QA tasks across grade levels, as evaluated by three LLM judges: **GPT-4o**, **Qwen2.5-72B**, and **Llama3.3-70B**. The scores reflect the alignment of each model's responses with grade-specific expectations.

choice questions from Indian medical exams, emphasizing domain-specific reasoning. TheoremQA (Chen et al., 2023) assesses application of 354 scientific theorems across physics, electrical engineering, and finance. MathSum (Yuan et al., 2020) focuses on summarizing math questions from Stack Exchange, while TABMWP (Lu et al., 2023) features 38K grade-level problems requiring table-based reasoning, a known challenge for LLMs. In science education, ARC (Clark et al., 2018) provides 7,787 multiple-choice questions for Grades 3–9, distinguishing between simple retrieval and complex reasoning. In programming education, Defects4J (Just et al., 2014) catalogs 357 real-world Java bugs, while ManyBugs and IntroClass (Le Goues et al., 2015) target C language and student-written code errors, supporting research in program repair and automated feedback. More recent efforts like CodeReviewer (Li et al., 2022) and follow-up work by (Guo et al., 2023) evaluate LLMs on code review and refinement tasks. In science QA, SciQ (Welbl et al., 2017) provides 13.7K textbook-based multiple-choice questions, while FairytaleQA (Xu et al., 2022) offers 10K QA pairs from classic children's stories, supporting comprehension assessment for kindergarten through Grade 8.

Multilingual and global benchmarks are also gaining importance. C-EVAL (Huang et al., 2023) features 13,948 Chinese questions across 52 disciplines and difficulty levels, while GAOKAO-Bench (Zhang et al., 2024) assesses LLMs using China's high-stakes college entrance exam, exposing consistent underperformance in STEM fields. Similarly, AGIEval (Zhong et al., 2023) compiles real-world exam questions from the SAT, LSAT, and civil service tests to evaluate cognitive and domain-specific reasoning. MMLU (Hendrycks et al., 2021a) and CMMLU (Li et al., 2024) further broaden this scope, offering diverse academic challenges across dozens of subjects in English and Chinese, respectively, and revealing LLM shortcomings in areas like negation and multi-step logic.

Evaluating LLM performance across educational levels is vital for building systems that support diverse learners. Multi-level benchmarks assess a model's adaptability across subjects and grade ranges. MATH-Vision (Wang et al., 2024) tests mathematical reasoning using both text and visual inputs across varying complexities. C-EVAL (Huang et al., 2023) includes 13,948 multiple-choice questions in Chinese, spanning 52 subjects and four difficulty levels. Despite progress, even advanced models like GPT-4 show limited accuracy, especially in STEM subjects requiring deeper reasoning. The AI2 Reasoning Challenge (ARC) (Clark et al., 2018) features grade-school science questions aimed at testing beyond surface-level retrieval, yet most models struggle to outperform simple heuristics. These results highlight a persistent gap in LLMs' ability to generalize across academic domains and educational levels. Realizing the full educational potential of LLMs requires models capable of both subject-specific reasoning and grade-level adaptability.

6 Conclusion and Future Work

This work introduces the first comprehensive benchmark for evaluating educational QA across all K–12 grade levels. High-quality QA pairs were validated through a combination of LLM-based and expert human review and used to assess a range of language models. Using accuracy for MCQs and LLM-as-a-Judge scoring for open-ended responses, we evaluated how well models align with the linguistic and cognitive needs of students at different stages. Results show a clear performance gap: models struggle significantly with lower-grade content (Grades 1–5), achieving only 60–70% on open-ended questions, compared to up to 85% for higher grades. Smaller models, in particular, showed poor performance across both MCQs and descriptive answers. These findings underscore the need for grade-aware training, prompting, and fine-tuning strategies tailored to younger learners.

Looking ahead, several directions can further improve educational language models. Expanding subject coverage will enable broader curriculum alignment, while incorporating multimodal QA (e.g., image or diagram-based) will better reflect real-world assessments. Supporting multilingual QA will increase accessibility for non-English-speaking students. Finally, addressing lower-grade performance through data augmentation, curriculum-aligned pretraining, and targeted fine-tuning is critical. Together, these efforts aim to build more reliable pedagogically grounded educational AI systems.

7 Limitations

While this study focused on building and evaluating a grade-specific benchmark dataset across multiple language models, several limitations should

be noted to contextualize the findings and guide future improvements. First, the dataset shows an imbalance in grade-level distribution, with fewer question-answer pairs for lower grades (Grades 1–5) compared to upper grades (Grades 6–12). This skew may affect the reliability of model evaluations for early-grade content and contribute to poorer performance in those categories. Future work should aim to create more balanced datasets to enable fairer and more comprehensive assessments across all grade levels. Second, the dataset is based on a K–12 curriculum framework, which may limit its generalizability to other educational systems. As curricular standards and cognitive expectations vary globally, adapting and extending the dataset for international contexts is essential for broader applicability. These limitations point to key areas for refinement, including more balanced data generation, enhanced cross-curricular alignment, and deeper integration of human judgment to support more accurate and inclusive educational evaluations.

8 Ethical Statement

Ethical responsibility was a core principle throughout this study. From data generation using Wikipedia articles to evaluating educational QA pairs, all stages were designed to ensure transparency, fairness, and minimal bias. Using publicly available sources like Wikipedia promoted reproducibility and avoided risks associated with private or sensitive data, aligning with the broader goals of openness and accountability in educational AI research.

Recognizing potential biases in LLMs due to pretraining data, we employed a diverse set of models varying in size and architecture to reduce reliance on any single system. For evaluation, we used an LLM-as-a-Judge framework with three independent models, supplemented by manual review to ensure reliability and consistency. No personal or identifiable student data was used. All generated content and evaluations were conducted solely for academic research. This study aims to contribute responsibly to the development of educational AI systems, emphasizing fairness, transparency, and trust.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open models. llama3.2.

Mistral AI. 2025. Mistral small 3: A latency-optimized 24b-parameter model. `https://mistral.ai/news/mistral-small-3`.

AI@Meta. 2024. Llama 3 model card.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024. Smollm - blazingly fast and remarkably powerful.

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards Interpretable math word problem solving with operation-based formalisms. *Preprint*, arXiv:1905.13319.

Argilla.io. 2025. Distilabel: An open-source framework for controllable labeling and data annotation. Accessed: March 10, 2025.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, Singapore. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. *Preprint*, arXiv:2310.01377.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh

Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Qi Guo, Junming Cao, Xiaofei Xie, Shangqing Liu, Xiaohong Li, Bihuan Chen, and Xin Peng. 2023. Exploring the potential of chatgpt in automated code refinement: An empirical study. *Preprint*, arXiv:2309.08221.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Danqing Huang, Shuming Shi, Chin-Yew Lin, Jian Yin, and Wei-Ying Ma. 2016. How well do computers solve math word problems? large-scale dataset construction and evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 887–896, Berlin, Germany. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Preprint*, arXiv:2305.08322.

René Just, Darioush Jalali, and Michael D. Ernst. 2014. Defects4j: a database of existing faults to enable controlled testing studies for java programs.

Brian Keeley and Céline Little, editors. 2017. *The State of the World's Children 2017: Children in a Digital World*. United Nations Children's Fund (UNICEF), New York, NY. ERIC Number: ED590013.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Claire Le Goues, Neal Holtschulte, Edward K. Smith, Yuriy Brun, Premkumar Devanbu, Stephanie Forrest, and Westley Weimer. 2015. The manybugs and introclass benchmarks for automated repair of c programs. *IEEE Transactions on Software Engineering*, 41(12):1236–1256.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

Zhiyu Li, Shuai Lu, Daya Guo, Nan Duan, Shailesh Jannu, Grant Jenks, Deep Majumder, Jared Green, Alexey Svyatkovskiy, Shengyu Fu, and Neel Sundaresan. 2022. Automating code review activities by large-scale pre-training. *Preprint*, arXiv:2203.09095.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *Preprint*, arXiv:2209.14610.

Mir Tafseer Nayeem and Davood Rafiei. 2024. Kidlm: Advancing language models for children–early insights and future directions. *arXiv preprint arXiv:2410.03884*.

NGSS Lead States. 2013. Next generation science standards: For states, by states. Accessed: 2025-05-16.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Matthew Renze and Erhan Guven. 2024. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*.

Victoria Rideout, Alanna Peebles, Supreet Mann, and Michael B. Robb. 2022. Common sense census: Media use by tweens and teens.

Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. 2023. Know your audience: Do llms adapt to different age and education levels? *Preprint*, arXiv:2312.02065.

Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. Chacha: Leveraging large language models to prompt children to share their emotions about personal events. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *Preprint*, arXiv:2212.09292.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Shyam Upadhyay and Ming-Wei Chang. 2017. Annotating derivations: A new evaluation strategy and dataset for algebra word problems. *Preprint*, arXiv:1609.07197.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina Kann. 2023. On the automatic generation and simplification of children's stories. *arXiv preprint arXiv:2310.18502*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Preprint*, arXiv:2402.14804.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 845–854, Copenhagen, Denmark. Association for Computational Linguistics.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *Preprint*, arXiv:1707.06209.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: Fairytaleqa – an authentic dataset for narrative comprehension. *Preprint*, arXiv:2203.13947.

Kun Yuan, Di He, Zihang Jiang, Liang Gao, Zhaowei Tang, and C. Lee Giles. 2020. Automatic generation of headlines for online math questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9490–9497.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2024. Evaluating the performance of large language models on gaokao benchmark. *Preprint*, arXiv:2305.12474.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

A Prompts

A.1 Grade-Level QA Generation Prompts

```
"You are an AI assistant specializing in creating educational content for young learners. Your task is
↪  to generate two simple, question-answer (QA) pairs (one mcq type and one qa type) based on the
↪  given text, suitable for Grade 1 and 2 students.

Instructions:
    - Use simple, short sentences with easy vocabulary appropriate for 6-8-year-old children.
    - Ask about observable things (what something looks like, where it lives, etc.).
    - Avoid reasoning or multi-step thinking
    - Keep the tone friendly, fun, and age-appropriate."
```

Listing 1: Prompt for grade 1 and 2 question-answer generation

```
"You are an AI assistant specializing in creating educational content for students in Grades 3 to 5.
↪  Your task is to generate two question-answer (QA) pairs (one mcq type and one qa type) based on
↪  the given text.
Instructions:
    - Use clear language suitable for ages 8-11
    - Use clear and concise language. Avoid overly complex words, but encourage age-appropriate
    ↪  critical thinking and explanation.
    - Focus on helping students understand important facts and cause-and-effect relationships.
    - Encourage observational or factual reasoning, not abstract modeling.
    - Keep the tone engaging, educational, and appropriate for upper elementary school learners."
```

Listing 2: Prompt for grade 3 to 5 question-answer generation

```
"You are a AI assistant specializing in creating a Question-Answer (QA) pair for middle school
↪  students (Grades 6-8) based on the provided text. Your task is to generate a {qa_or_mcq} based on
↪  the given text.
Instructions:
    - Use vocabulary and complexity suitable for students aged 12-14.
    - Ask questions that require students to interpret information, reason through cause-and-effect,
    ↪  apply models, or predict outcomes.
    - Focus on scientific relationships, system interactions, and basic modeling of processes or
    ↪  phenomena.
    - Simplify complex or abstract ideas into familiar contexts that students can reason about.
    - Maintain an educational tone that encourages scientific thinking and exploration."
```

Listing 3: Prompt for grade 6 to 8 question-answer generation

```
"You are a AI assistant specializing in creating a Question-Answer (QA) pair for high school students
↪  (Grades 9-12) based on the provided text. Your task is to generate a {qa_or_mcq} based on the
↪  given text.
Instructions:
    - Use academically precise language appropriate for students aged 14-18 preparing for advanced or
    ↪  college-level studies.
    - Focus on modeling, applying laws, analyzing systems, and using quantitative or qualitative
    ↪  relationships.
    - Ask questions that require students to analyze, model, predict, calculate, or critically
    ↪  evaluate scenarios.
    - Ensure the question and answer are fully self-contained and understandable without needing to
    ↪  reference the original text.
    - Maintain an academic, analytical tone suited for high school science learners."
```

Listing 4: Prompt for grade 9 to 12 question-answer generation

A.2 Self-Reflection and Human Evaluation Prompts


```
"Your role is to evaluate each question-answer pair for Grade {grade_level} students in the subject of
↪ {subject}, focusing on the following criteria:

Evaluation Criteria:

    1. language-appropriateness: Is the language simple, short, and easy for 6-8-year-old children to
    ↪  understand?
    2. grade-alignment: Does the question reflect what students at this age typically observe or
    ↪  experience.
    3. relevance: Is the question-answer pair based on observable actions or phenomena, and
    ↪  understandable on its own without needing to refer back to any source?
    4. clarity: Is the question phrased clearly, with an unambiguous answer?
    5. subject-fit ({subject}): Does the question relate to age-appropriate scientific concepts in
    ↪  this subject, without factual inaccuracies or misconceptions?

    Rate each criteria on 1 to 10.
    "
```

Listing 5: Prompt for evaluating the quality of grade 1-2 QA pairs through self-reflection

```
"Your role is to evaluate each question-answer pair for Grade {grade_level} students in the subject of
↪ {subject}, focusing on the following criteria:

    Evaluation Criteria:

        1. language-appropriateness: Is the language clear, age-appropriate (for 8-11-year-old
        ↪  students), avoiding overly complex vocabulary but encouraging basic reasoning?
        2. grade-alignment: Does the question match the cognitive and curriculum expectations for
        ↪  Grades 3-5, focusing on understanding facts, cause-and-effect, or simple scientific
        ↪  reasoning?
        3. relevance: Is the QA pair directly related to observable phenomena, simple explanations, or
        ↪  important scientific facts appropriate to the grade level?
        4. clarity: Is the question clearly phrased, guiding students to provide or recognize a
        ↪  straightforward explanation or prediction?
        5. subject-fit ({subject}): Does the content accurately reflect important concepts from the
        ↪  subject suitable for upper elementary learners?

    Please rate each criteria on 1 to 10 scale.
    "
```

Listing 6: Prompt for evaluating the quality of grade 3-5 QA pairs through self-reflection

```
"Your role is to evaluate each question-answer pair for Grade {grade_level} students in the subject of
↪ {subject}, focusing on the following criteria:

    Evaluation Criteria:

        1. language-appropriateness: Is the language clear, precise, and appropriate for
        ↪  12-14-year-old students, supporting intermediate scientific reasoning?
        2. grade-alignment: Does the question match the cognitive expectations for middle school
        ↪  learners, involving interpretation, cause-and-effect analysis, simple system modeling, or
        ↪  predictions?
        3. relevance: Is the QA pair rooted in scientific phenomena, relationships, or system-level
        ↪  interactions appropriate to the grade level?
        4. clarity: Is the question phrased clearly, guiding students to reason, analyze, or predict
        ↪  in a focused and understandable way?
        5. subject-fit ({subject}): Does the content reflect accurate and important scientific
        ↪  concepts appropriate for middle school science in this subject?

        Please rate each criteria on 1 to 10 scale.
    "
```

Listing 7: Prompt for evaluating the quality of grade 6-8 QA pairs through self-reflection

```
"Your role is to evaluate each question-answer pair for Grade {grade_level} students in the subject of
↪ {subject}, focusing on the following criteria:

    Evaluation Criteria:

        1. language-appropriateness: Is the language academically precise and appropriate for students
        ↪  aged 14-18 preparing for advanced science studies?
        2. grade-alignment: Does the question meet the cognitive expectations for high school learners,
        ↪  requiring multi-step reasoning, quantitative analysis, modeling, or critical evaluation?
        3. relevance: Is the QA pair grounded in substantial scientific concepts, systems modeling, or
        ↪  data-driven explanations appropriate for high school science?
        4. clarity: Is the question phrased clearly and at a cognitive depth suitable for high school
        ↪  students?
        5. subject-fit ({subject}): Does the content align with advanced high school curriculum topics
        ↪  within the subject, and maintain scientific accuracy?

        Please rate each criteria on 1 to 10 scale."
```

Listing 8: Prompt for evaluating the quality of grade 9-12 QA pairs through self-reflection

A.3 Prompt for LLM-as-a-Judge Evaluation

```
"""Your role is to evaluate the model's response for a student of Grade {{ grade_level }} by comparing
↪  it to the gold answer.

Evaluation Criteria:
  Evaluate the model's response in relation to the gold answer, based on the following criteria:

  - Vocabulary Alignment: Does the model use vocabulary that closely matches the complexity,
  ↪  accessibility, and tone of the gold answer, assuming the gold answer is already
  ↪  grade-appropriate?
  - Conceptual Alignment: Does the model's response reflect a similar level of cognitive and
  ↪  conceptual depth as the gold answer?
  - Scientific Language Alignment: Does the model use scientific or technical terms in a way that
  ↪  aligns with the gold answer in terms of complexity and usage?
  - Correctness: Is the model's answer factually accurate and consistent with the gold answer?
  - Clarity: Is the model's response as clear, coherent, and well-structured as the gold answer?
  - Completeness: Does the model's answer cover the same key ideas, explanations, or observations as
  ↪  the gold answer?

  Assign a rating from 1 to 10 to each criteria based on how well the model's answer aligns with the
  ↪  gold answer across each criterion:

Question:
{{ question_text }}

Gold Answer:
{{ gold_answer }}

Model Answer:
{{ model_answer }}
"""
```

Listing 9: Prompt for LLM-as-a-Judge evaluation

A.4 Prompt for Answer Generation from LLMs

```
"You are an experienced educator answering questions for students in {grade_level}. Please give a
↪  clear and developmentally appropriate answer to the question below."
```

Listing 10: Prompt for evaluating different LLMs on testset

B Grade-Level QA Pairs

Below sections contain example qa pairs from our dataset for each field and grade level.

B.1 Biology

```
"grade_level: grade 1 and 2
question: What kind of animals live in the Pigsties?
answer: Pigs live in the Pigsties."
```

Listing 11: Grade 1 and 2 QA pair for biology

```
"grade_level: grade 3 to 5
question: What does it mean if a species is omnivorous?
answer: If a species is omnivorous, it means it eats both plants and animals. This allows the species
↪  to have a varied diet and adapt to different food sources available in its habitat."
```

Listing 12: Grade 3 to 5 QA pair for biology

```
"grade_level: grade 6 to 8
question: How do vampire bats locate blood vessels in their prey, and why is this adaptation important
↪  for their feeding habits?
answer: Vampire bats use heat sensors in their noses to detect blood vessels near the surface of the
↪  skin. This adaptation is important because it allows them to accurately find and target areas rich
↪  in blood, making their feeding process more efficient."
```

Listing 13: Grade 6 to 8 QA pair for biology

```
"grade_level: grade 9 to 12
question: Explain why trypan blue is used as a vital stain in biosciences and how it helps
↪  differentiate between live and dead cells.
answer: Trypan blue is used as a vital stain in biosciences because it selectively colors dead tissues
↪  or cells blue, while live cells with intact cell membranes remain unstained. This is due to the
↪  selective permeability of cell membranes, which allows trypan blue to pass through and stain dead
↪  cells, but not live cells. This property makes it a useful tool for distinguishing between live and
↪  dead cells under a microscope, as dead cells appear blue while live cells do not take up the dye."
```

Listing 14: Grade 9 to 12 QA pair for biology

B.2 Physics

```
"grade_level: grade 1 and 2
question: What is a whirlpool?
answer: A whirlpool is a swirling movement of water."
```

Listing 15: Grade 1 and 2 QA pair for physics

```
"grade_level: grade 3 to 5
question: How are sound waves in water detected by a receiver like the human ear or a hydrophone?
answer: Sound waves in water are detected by a receiver as changes in pressure. The receiver senses
↪  the alternating compressions and rarefactions of the water, which are changes in how tightly the
↪  water molecules are packed together."
```

Listing 16: Grade 3 to 5 QA pair for physics

```
"grade_level: grade 6 to 8
question: How does the pressure exerted by a glacier affect the melting point of ice at its base, and
↪  what is the result of this process?
answer: The pressure exerted by a glacier on its lower surface lowers the melting point of the ice,
↪  causing it to melt. This melting allows the glacier to move from a higher elevation to a lower
↪  elevation, and at lower elevations, the liquid water may flow from the base of the glacier when
↪  the air temperature is above the freezing point of water."
```

Listing 17: Grade 6 to 8 QA pair for physics

```
"grade_level: grade 9 to 12
question: Explain how pressure piling can lead to a deflagration to detonation transition in
↪  connected vessels, and discuss the measures taken to prevent this in electrical equipment.
answer: Pressure piling occurs when a flame front propagates along a tube, compressing and heating the
↪  unburned gases ahead of it. This compression can significantly increase the pressure, ranging
↪  from twice to eight times the initial pressure. In systems where multiple vessels are connected by
↪  piping, this can lead to a deflagration to detonation transition, resulting in a very large
↪  explosion pressure. In electrical equipment in hazardous areas, this risk is mitigated by
↪  avoiding the use of conduits to connect classified equipment and by using barrier glands on cables
↪  entering enclosures. These measures ensure that compartments remain separate, preventing the
↪  transmission of explosions from one compartment to another."
```

Listing 18: Grade 9 to 12 QA pair for physics

B.3 Chemistry

```
"grade_level: grade 1 and 2
question: What does moisture mean?
answer: Moisture means the presence of water, often in small amounts."
```

Listing 19: Grade 1 and 2 QA pair for chemistry

```
"grade_level: grade 3 to 5
question: Why is it important to know if a substance is soluble in water?
answer: Knowing if a substance is soluble in water helps us understand how it can be used or handled.
↪  For example, if a substance dissolves in water, it can be mixed into drinks or used in cooking. It
↪  also helps scientists and engineers in creating solutions for cleaning, medicine, and other
↪  applications."
```

Listing 20: Grade 3 to 5 QA pair for chemistry

```
"grade_level: grade 6 to 8
question: What is the difference between an accepted value and an experimental value in chemistry?
answer: An accepted value is a value of a substance's properties that is agreed upon by almost all
↪  scientists, while an experimental value is the value of a substance's properties that is
↪  determined in a specific laboratory setting."
```

Listing 21: Grade 6 to 8 QA pair for chemistry

```
"grade_level: grade 9 to 12
question: Explain how acidosis affects the pH level of blood or body fluids, and why this change
↪  occurs.
answer: Acidosis increases the concentration of hydrogen ions in blood or body fluids. Since pH is the
↪  negative logarithm of hydrogen ion concentration, an increase in hydrogen ions results in a
↪  decrease in pH. This occurs because the pH scale is inversely related to hydrogen ion
↪  concentration; more hydrogen ions mean a lower pH, indicating increased acidity."
```

Listing 22: Grade 9 to 12 QA pair for chemistry

B.4 Computer Science

```
"grade_level: grade 1 and 2
question: What can you do with the Game Boy?
answer: "You can play games on the Game Boy."
```

Listing 23: Grade 1 and 2 QA pair for computer science

```
"grade_level: grade 3 to 5
question: What is the main purpose of the Little Professor calculator?
options: [To provide the answer to a mathematical expression, To generate unsolved expressions and
↪  prompt the user for the answer, To teach children how to use a regular calculator, To help
↪  children memorize multiplication tables.]
correct_answer: To generate unsolved expressions and prompt the user for the answer"
```

Listing 24: Grade 3 to 5 QA pair for computer science

```
"grade_level: grade 6 to 8
question: Why did Apple decide to change the case design of the first production Apple II computers,
↪  and what solution did they implement?
answer: Apple decided to change the case design because the initial design had no vent openings, which
↪  caused high heat buildup from the PCB. This heat caused the plastic to soften and sag. To solve
↪  this problem, Apple added vent holes to the case within three months of production. Customers with
↪  the original case could have them replaced at no charge."
```

Listing 25: Grade 6 to 8 QA pair for computer science

```
"grade_level: grade 9 to 12
question: How did the development of mod_qos evolve to enhance the security of Apache servers over
↪  time?
answer: The development of mod_qos evolved to enhance the security of Apache servers by initially
↪  limiting the number of concurrent HTTP requests for specified resources. Over time, it introduced
↪  features such as connection level controls, request/response throttling, per client (IP) control
↪  mechanisms, and anomaly detection using client characteristic measurement. Additionally, it added
↪  geolocation features and improved response throttling, along with supporting differentiated
↪  services by controlling the DSCP IP field, all of which contributed to protecting servers against
↪  DoS attacks."
```

Listing 26: Grade 9 to 12 QA pair for computer science

B.5 Medicine

```
"grade_level: grade 1 and 2
question: What should you avoid to stay healthy? answer: Avoid touching people who are sick."
```

Listing 27: Grade 1 and 2 QA pair for medicine

```
"grade_level: grade 3 to 5
question: Why is it important for people to get vaccinated against COVID-19?
answer: Getting vaccinated helps protect people from getting sick with COVID-19. It also helps stop
↪  the virus from spreading to others, which can keep communities safer and help everyone return to
↪  normal activities more quickly."
```

Listing 28: Grade 3 to 5 QA pair for medicine

```
"grade_level: grade 6 to 8
question: Why might people with severe COVID-19 need respiratory support, and what are some forms of
↪  this support?
answer: People with severe COVID-19 might need respiratory support because the virus can severely
↪  affect their lungs, making it difficult for them to breathe. Some forms of respiratory support
↪  include oxygen therapy, mechanical ventilation, and intravenous fluids."
```

Listing 29: Grade 6 to 8 QA pair for medicine

```
"grade_level: grade 9 to 12
question: Analyze the impact of Dame Kate Isabel Campbell's discovery on the treatment of premature
↪  babies worldwide. How did her findings change medical practices?
answer: Dame Kate Isabel Campbell's discovery that blindness in premature babies was caused by high
↪  concentrations of oxygen led to a significant change in medical practices worldwide. Her findings
↪  prompted a reevaluation of the treatment protocols for premature babies, specifically regarding
↪  the administration of oxygen. As a result, medical professionals adjusted the oxygen levels used
↪  in neonatal care to prevent blindness, thereby improving the health outcomes for premature
↪  infants globally."
```

Listing 30: Grade 9 to 12 QA pair for medicine

B.6 Metrology

```
"grade_level: grade 1 and 2
question: Can you name something that might be measured using a unit of volume?
answer: Water, rice, sugar, grain, or flour."
```

Listing 31: Grade 1 and 2 QA pair for metrology

```
"grade_level: grade 3 to 5
question: Why do graduated cylinders have marked lines? answer: Graduated cylinders have marked lines
↪  to show the amount of liquid that has been measured. These lines help people accurately measure
↪  the volume of liquids in the cylinder."
```

Listing 32: Grade 3 to 5 QA pair for metrology

```
"grade_level: grade 6 to 8
question: Explain how the volume of a cubic inch is related to a US gallon and why this might be
↪  useful in understanding volume conversions.
answer: A cubic inch is 1/231 of a US gallon. This relationship is useful for understanding volume
↪  conversions because it provides a way to translate between smaller units of volume (like cubic
↪  inches) and larger, more commonly used units (like gallons), which can be helpful in various
↪  practical applications such as cooking, fuel measurements, and fluid storage."
```

Listing 33: Grade 6 to 8 QA pair for metrology

```
"grade_level: grade 9 to 12
question: How do enhanced geothermal systems (EGS) differ from traditional oil and gas fracking
↪  techniques in terms of environmental impact, and what measures are taken to minimize potential
↪  damage?
answer: Enhanced geothermal systems (EGS) differ from traditional oil and gas fracking techniques
↪  primarily in their environmental impact. While both techniques involve injecting fluids under
↪  high pressure to expand rock fissures, EGS does not use toxic chemicals, reducing the possibility
↪  of environmental damage. Instead, EGS uses proppants like sand or ceramic particles to keep the
↪  cracks open and ensure optimal flow rates. Additionally, the geologic formations targeted by EGS
↪  are deeper, which further minimizes the risk of environmental harm."
```

Listing 34: Grade 9 to 12 QA pair for metrology

B.7 Ecology

```
"grade_level: grade 1 and 2
question: What does the gecko mostly eat?,
answer: The gecko mostly eats insects."
```

Listing 35: Grade 1 and 2 QA pair for ecology

```
"grade_level: grade 3 to 5
question: Why do you think the white-winged swallow builds its nest a few meters above water?,
answer: The white-winged swallow likely builds its nest a few meters above water to protect its eggs
↪  and young from predators and to ensure easy access to food, as swallows often feed on insects
↪  found near water."
```

Listing 36: Grade 3 to 5 QA pair for ecology

```
"grade_level: grade 6 to 8
question: What are the components included within the boundaries of the MPA, and why might it be
↪  important to protect all of these components?
answer: The MPA includes the water column, the seabed, and the subsoil. Protecting all these
↪  components is important because they are interconnected ecosystems that support marine life,
↪  maintain biodiversity, and ensure the health of the marine environment. The water column provides
↪  habitat and resources for marine organisms, the seabed is home to various species and supports
↪  ecological processes, and the subsoil contains nutrients and minerals crucial for the overall
↪  ecosystem."
```

Listing 37: Grade 6 to 8 QA pair for ecology

```
"grade_level: grade 9 to 12
question: How does the long-eared myotis adapt its feeding strategy when hunting beetles, and why is
↪  this adaptation beneficial?
answer: The long-eared myotis adapts its feeding strategy by using its robust molars and highly placed
↪  articular process to exert more crushing force, allowing it to penetrate the hard carapace of
↪  beetles. This adaptation is beneficial because it enables the bat to effectively consume beetles,
↪  which have a protective shell that would otherwise be difficult to penetrate."
```

Listing 38: Grade 9 to 12 QA pair for ecology

B.8 Geology

```
"grade_level: grade 1 and 2
question: What does the Ely Limestone preserve?
answer: It preserves fossils"
```

Listing 39: Grade 1 and 2 QA pair for geology

```
"grade_level: grade 3 to 5
question: Why do aftershocks occur after a big earthquake?
answer: Aftershocks occur because the ground is still adjusting after the main earthquake. The big
↪  earthquake changes the stress in the Earth's crust, and the aftershocks are the Earth's way of
↪  settling back into a new balance."
```

Listing 40: Grade 3 to 5 QA pair for geology

```
"grade_level: grade 6 to 8
question: What processes contribute to the formation of floodplains, and how do they impact the
↪  landscape? answer: Floodplains are formed by the deposition of suspended load from overbank flow,
↪  bedload deposition from lateral river migration, and landscape processes such as landslides.
↪  These processes contribute to the buildup of land adjacent to river channels and shape the
↪  landscape by adding new layers of soil and altering the river's path."
```

Listing 41: Grade 6 to 8 QA pair for geology

```
"grade_level: grade 9 to 12
question: Explain how erosional sheltering contributes to the formation of a crag and tail structure
↪  in rocks. answer: Erosional sheltering occurs when rocks contain particles that are harder than
↪  the surrounding material. As the rock is worn down, these harder particles resist erosion more
↪  effectively than the softer rock. This resistance protects the rock on the lee side of the hard
↪  particle from further wear. Over time, this process results in the formation of a crag, where the
↪  hard particle was located, and a tail that extends parallel to the direction of movement down-slip
↪  from the particle. The crag and tail structure is thus a direct result of the differential erosion
↪  rates between the hard particles and the surrounding softer rock."
```

Listing 42: Grade 9 to 12 QA pair for geology

B.9 Geography

```
"grade_level: grade 1 and 2
question: Is France in Europe?
answer: Yes, France is in Europe."
```

Listing 43: Grade 1 and 2 QA pair for geography

```
"grade_level: grade 3 to 5
question: Why might knowing the altitude and area of a municipality be important?
answer: Knowing the altitude and area of a municipality can help us understand its climate, the types
↪  of plants and animals that live there, and how people might use the land. For example, higher
↪  altitudes might have cooler temperatures, and larger areas might have more space for homes, parks,
↪  or farms."
```

Listing 44: Grade 3 to 5 QA pair for geography

```
"grade_level: grade 6 to 8
question: Why is Île-de-France no longer considered an official wine region, and what recent
↪  developments suggest a revival in its viticulture?
answer: Île-de-France is no longer considered an official wine region due to changes in wine region
↪  classifications. However, its viticulture is experiencing a revival, as evidenced by the
↪  establishment of more than 200 small recreational vineyards in recent decades, covering about 12
↪  hectares, and the involvement of 5 villages in the Champagne area."
```

Listing 45: Grade 6 to 8 QA pair for geography

```
"grade_level: grade 9 to 12
question: Analyze how the historical significance of the Payap area has influenced its current role in
↪ Bangkok. Consider the changes in its name, infrastructure, and community over time.
answer: The historical significance of the Payap area has greatly influenced its current role in
↪ Bangkok. Originally known as 'Payap,' meaning 'northwestern' or 'northern,' it was a significant
↪ location for the Chet Ton Dynasty, particularly as the residence of Prince Dilok Nopparat, son of
↪ King Chulalongkorn. This royal connection established Payap as a notable area in Bangkok. After
↪ the 1932 revolution, the area was renamed 'Si Yan,' meaning 'auspicious quarter,' reflecting a
↪ desire for good fortune and a new identity. Over time, the infrastructure evolved with the
↪ establishment of the Payap pier, which became a vital transport link via the Chao Phraya Express
↪ Boat. The community also transformed, becoming a diverse neighborhood with a mix of Thai and
↪ Chinese descent, reflecting its long history and resilience, especially during World War II when
↪ it housed a bomb shelter. Today, the area continues to serve as a functional and historical part
↪ of Bangkok, with the Ministry of Interior's Operation Centre for Displaced Persons and the
↪ Metropolitan Electricity Authority Hospital Samsen, indicating its ongoing importance in the
↪ city's social and administrative landscape."
```

Listing 46: Grade 9 to 12 QA pair for geography

C Tables

	Grade Levels	Biology	Physics	Chemistry	Computer Science	Ecology	Geography	Geology	Medicine	Metrology
QA Count	1 and 2	269	41	81	99	510	746	159	85	55
	3 to 5	601	59	158	263	906	1252	483	236	127
	6 to 8	212	55	90	185	440	272	180	106	72
	9 to 12	845	1485	1624	1238	696	412	749	973	996
MCQ Count	1 and 2	80	29	45	23	111	156	19	36	24
	3 to 5	238	28	49	68	419	543	145	44	36
	6 to 8	241	57	88	184	572	284	250	112	73
	9 to 12	675	1183	1453	1190	648	350	794	751	1111
Total		**3161**	**2937**	**3594**	**3250**	**4302**	**4015**	**2779**	**2343**	**2494**

Table 6: **Training Set:** Subject-wise and grade-level distribution of question-answer pairs in the training split.

	Grade Levels	Biology	Physics	Chemistry	Computer Science	Ecology	Geography	Geology	Medicine	Metrology
QA Count	1 and 2	95	13	21	34	169	253	44	31	15
	3 to 5	199	20	51	86	302	417	158	71	45
	6 to 8	71	20	24	64	172	109	61	31	24
	9 to 12	282	495	541	413	225	136	250	324	332
MCQ Count	1 and 2	21	9	20	6	37	47	15	8	10
	3 to 5	80	8	18	23	152	187	51	21	8
	6 to 8	79	16	34	58	164	76	82	41	23
	9 to 12	225	395	484	397	227	119	265	250	370
Total		**1052**	**976**	**1193**	**1081**	**1448**	**1344**	**926**	**777**	**827**

Table 7: **Development Set:** Subject-wise and grade-level distribution of question-answer pairs in the development split.

	Grade Levels	Biology	Physics	Chemistry	Computer Science	Ecology	Geography	Geology	Medicine	Metrology
QA Count	1 and 2	92	13	31	29	161	251	53	30	22
	3 to 5	204	21	54	89	302	417	169	72	37
	6 to 8	61	14	30	53	147	82	58	38	25
	9 to 12	282	495	541	413	222	132	250	324	332
MCQ Count	1 and 2	24	9	10	11	45	49	6	9	3
	3 to 5	75	7	15	20	165	183	40	20	16
	6 to 8	89	22	28	69	189	103	85	34	22
	9 to 12	225	395	484	397	230	123	265	250	370
Total		**1052**	**976**	**1193**	**1081**	**1448**	**1344**	**926**	**777**	**827**

Table 8: **Test Set:** Subject-wise and grade-level distribution of question-answer pairs in the test split.

Models	Huggingface Identifiers	Usage
phi-4	microsoft/phi-4	QA Generation/Self-Reflection
Qwen2.5-1.5B	Qwen/Qwen2.5-1.5B-Instruct	Evaluation
SmolLM-1.7B	HuggingFaceTB/SmolLM-1.7B-Instruct	Evaluation
Gemma-2b-it	google/gemma-2b-it	Evaluation
Qwen2.5-3B	Qwen/Qwen2.5-3B-Instruct	Evaluation
Llama3.2-3B	meta-llama/Llama-3.2-3B-Instruct	Evaluation
Qwen2.5-7B	Qwen/Qwen2.5-7B-Instruct	Evaluation
Llama3-8B	meta-llama/Meta-Llama-3-8B-Instruct	Evaluation
Qwen2.5-14B	Qwen/Qwen2.5-14B-Instruct	Evaluation
Mistral24B	mistralai/Mistral-Small-24B-Instruct-2501	Evaluation
GPT4o	gpt-4o-2024-08-06	LLM-as-a-Judge
Qwen2.5-72B	Qwen/Qwen2.5-72B-Instruct-GPTQ-Int8	LLM-as-a-Judge
Llama-3.3-70B	meta-llama/Llama-3.3-70B-Instruct	LLM-as-a-Judge

Table 9: Huggingface identifiers of our models and their usage point across the pipeline