

Classroom AI: Large Language Models as Grade-Specific Teachers

Jio Oh^{124*}, Steven Euijong Whang¹, James Evans³, Jindong Wang^{4†}

¹KAIST, ²Microsoft Research Asia, ³University of Chicago, ⁴William & Mary

Abstract

Large Language Models (LLMs) offer a promising solution to complement traditional teaching and address global teacher shortages that affect hundreds of millions of children, but they fail to provide grade-appropriate responses for students at different educational levels. We introduce a framework for finetuning LLMs to generate age-appropriate educational content across six grade levels, from lower elementary to adult education. Our framework successfully adapts explanations to match students’ comprehension capacities without sacrificing factual correctness. This approach integrates seven established readability metrics through a clustering method and builds a comprehensive dataset for grade-specific content generation. Evaluations across multiple datasets with 208 human participants demonstrate substantial improvements in grade-level alignment, achieving a 35.64 percentage point increase compared to prompt-based methods while maintaining response accuracy. AI-assisted learning tailored to different grade levels has the potential to advance educational engagement and equity.

Introduction

Large Language Models (LLMs) have expanded beyond traditional natural language processing into diverse domains ranging from finance [1, 2] and healthcare [3, 4] to education [5, 6, 7]. With models like GPT [8], Gemini [9], and LLaMA [10] now accessible worldwide, LLMs transform education by assisting teachers and students with question-solving, confusion clarification, material creation, and content personalization [6, 7].

Teacher shortages present a large and growing global challenge. UNESCO estimates 44 million additional teachers are needed to achieve universal primary and secondary education by 2030 [11], while 35% of U.S. public schools report at least one teaching vacancy [12]. Worldwide, 244 million children lack access to school [11], with shortages most acute in rural and high-poverty areas [13]. In Pakistan, 44% of students drop out between ages 5-16 [14, 15], with 18.6% leaving before completing primary education. Limited access to qualified teachers exacerbates educational disparities [13, 15], with over 70% of teachers in sub-Saharan Africa classified as inadequately trained [16].

With 67.9% of the global population connected to the internet, LLM-based educational tools could benefit over 100 million children currently without school access and provide enhanced assistance to countless others. LLMs can deliver consistent explanations and personalized assistance regardless of geographic location, potentially increasing learning engagement and reducing educational inequity worldwide.

Effective teachers require both subject knowledge and pedagogical skills tailored to different grade levels. Despite their capabilities, LLMs struggle to provide grade-appropriate answers [17]. Even with explicit prompts like “Answer for 3rd graders”, LLMs generate responses that systematically exceed the target grade’s comprehension level. Existing works mostly focus on prompt-based evaluations that fail to achieve a satisfactory level due to the lack of comprehensive evaluation criteria and appropriately curated training corpora [18, 19, 20, 21]. To serve as effective educational tools, LLMs must produce content students can understand and directly engage with.

Previous work on finetuning LLMs for specific reading levels focused primarily on summarization or paraphrasing [22, 23, 24]. However, real classroom settings involve open-ended questions without source

*Work done during internship in Microsoft Research Asia and William & Mary <harryoh99@kaist.ac.kr>

†Corresponding author: Jindong Wang <jdw@wm.edu>

This paper was published in npj Artificial Intelligence (2026).

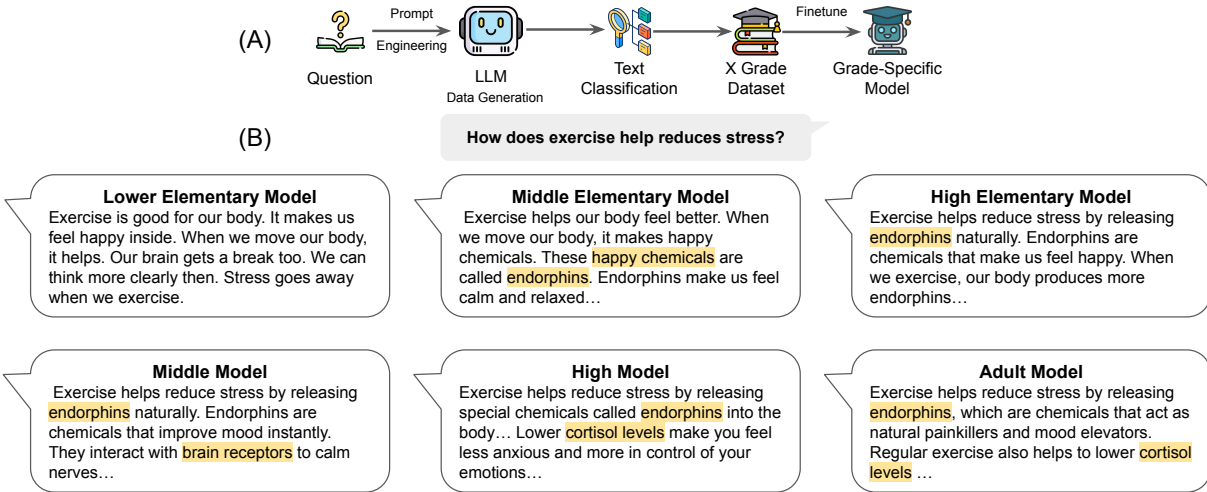


Figure 1: (A) Overall framework of our approach. (B) Exemplar responses across different models.

texts to summarize or paraphrase. For example, when a student asks “What is gravity?” and a teacher needs LLM assistance, summarization approaches fail. This becomes critical in AI tutoring scenarios where students interact directly with LLMs. Even high-quality LLM responses provide no benefit if students cannot comprehend them.

We propose a framework for grade-level targeted finetuning of LLMs that handles open-ended educational queries across various subjects. Our approach enables grade-appropriate content generation for natural questions across six educational levels: lower elementary (grades 1-2), middle elementary (grades 3-4), upper elementary (grades 5-6), middle school (grades 7-9), high school (grades 10-12), and college/adult (grade 13+). This granular classification reflects research in educational psychology showing that reading and comprehension skills develop rapidly during early education [25, 26]. This classification can provide a more accurate measure of grade-level suitability, enabling our evaluation framework to detect subtle yet significant shifts in linguistic complexity that might be overlooked. Moreover, our approach advocates for models that behave appropriately given teaching context [27].

To assess text complexity, we integrate seven established readability metrics: Flesch Reading Ease [28], Flesch-Kincaid Grade Level [29], the Coleman-Liau Index [30], Linsear Write [31], the Gunning Fog Index [32], Dale-Chall [33], and the Spache Readability Formula [34]. Each metric captures different aspects of readability, and we group them based on their underlying characteristics to create a more reliable integrated measure (see [Readability Metrics Integration](#) for further details).

To address the challenge of limited training data for grade-specific content, we generate data using LLMs, a technique increasingly used for data collection and generation [35, 36]. LLMs are known to generate high-quality text data that aligns with user instructions, thereby improving model performance when finetuned on such data [37, 38, 39]. Notably, Orca-math [40] presents a math dataset with GPT by guiding the model to adopt a teacher–student paradigm, which highlights the potential of instruction-aligned synthetic data to enhance educational resources. We categorize 54 subjects across 8 fields to create diverse questions, then use state-of-the-art LLMs (GPT [8], Gemini [9], LLaMA [10], and Mixtral [41]) to generate questions answerable across all grade levels (see Supplementary Figure 1, for details on subjects). We craft tailored prompts for LLaMA3.1:70B to produce outputs for each grade level, then classify generated text using our integrated metrics algorithm (see [Answer Generation](#) for prompt details).

Our experiments demonstrate that this approach significantly improves grade-level alignment compared to prompt-based methods while maintaining response accuracy. Human studies with 208 participants confirm that our framework aligns with human perceptions of grade-appropriate content.

Our contributions include:

- A framework for developing grade-specific LLMs to enhance educational equity and deliver social benefits globally.

- Extensive evaluation with 208 participants validating our framework’s alignment with human perceptions of difficulty, showing that finetuned models can explain complex concepts at targeted grade levels.
- A model-agnostic dataset for finetuning LLMs in educational contexts, integrating multiple educational metrics with grade-appropriate responses for open-ended questions.

Results

We evaluate our method on two main dimensions: **compatibility** and **accuracy**. **Compatibility** measures whether finetuned models’ outputs align with target grade students’ comprehension capability. We assess compatibility using: (1) an integrated measure using seven readability metrics (see Supplementary Section B, for definitions of the metrics), (2) individual evaluations for each metric, and an (3) Automated Readability Index (ARI) [42] as a held-out metric to test generalizability. **Accuracy** measures whether the model maintains its ability to generate correct and relevant responses.

We also measure perplexity and diversity gain [43], which reflect linguistic variety relative to the training corpora and the base model. Finally, we conduct surveys with 208 human participants and GPT4o to validate our framework’s alignment with human perceptions.

Compatibility

We test the finetuned models’ compatibility on all four datasets. For \mathcal{D}_{GPT} , \mathcal{D}_{ELI} , and \mathcal{D}_{NQ} , we use all sampled questions. For \mathcal{D}_{SQ2} , we split questions based on their designated grade levels and analyze results accordingly. For example, when targeting lower elementary level, we focus exclusively on grades 1-2 questions.

As shown in Figure 2(A), our approach significantly increases target success rates for each grade level compared to prompt-based approaches, with an average improvement of 35.64 percentage points over the prompt-based baseline. The blue bars represent the success rate of each model in producing outputs at the intended grade level (higher is better), while red dots show corresponding ARI values discretized into six difficulty levels. Dots closer to the red dotted line indicate stronger alignment with ARI. Similar improvements appear for the held-out ARI metric, with our approach best aligning with intended grade levels. Detailed output grade-level distribution is shown in Supplementary Section C.3 and C.6.

Figure 2(B) shows that our approach successfully shifts all seven educational metrics toward their optimal values (1 for lower elementary through 6 for adult) compared to the base model or prompt-based approaches. (Note that DC, LW, FKGL, Fog, Sp, CLI, and FRES are measures for Dale-Chall, Linsear Write, Flesch-Kincaid Grade Level, Gunning Fox Index, Spache Readability Formula, Coleman-Liau Index, and Flesch Reading Ease respectively.) This improvement stands out for elementary school grade levels, which previous research identified as most challenging to target [17].

Accuracy

We test the finetuned models’ accuracy on \mathcal{D}_{SQ} , a multiple-choice dataset aligned with educational contexts. Figure 2(C) shows that finetuned models achieve performance comparable to the base model. While finetuning typically causes some accuracy reduction [44, 45, 46], our results show minimal performance degradation.

Perplexity and Diversity

We measure output unexpectedness through perplexity and diversity gain (see Supplementary Section B.2 and B.3, for formulas). Figure 2(D) shows that lower-grade models exhibit higher values for both metrics, suggesting that text comprehensible for lower grades appears less frequently in training corpora, explaining why traditional approaches struggle with these levels. Lower-grade models convey difficult concepts using simpler language, resulting in higher diversity compared to the more direct language of existing models.

Survey

We conduct two surveys using different datasets. Survey 1 uses questions from \mathcal{D}_{NQ} with 108 participants, while Survey 2 uses questions from \mathcal{D}_{SQ2} with 120 participants. All participants are English-speaking and

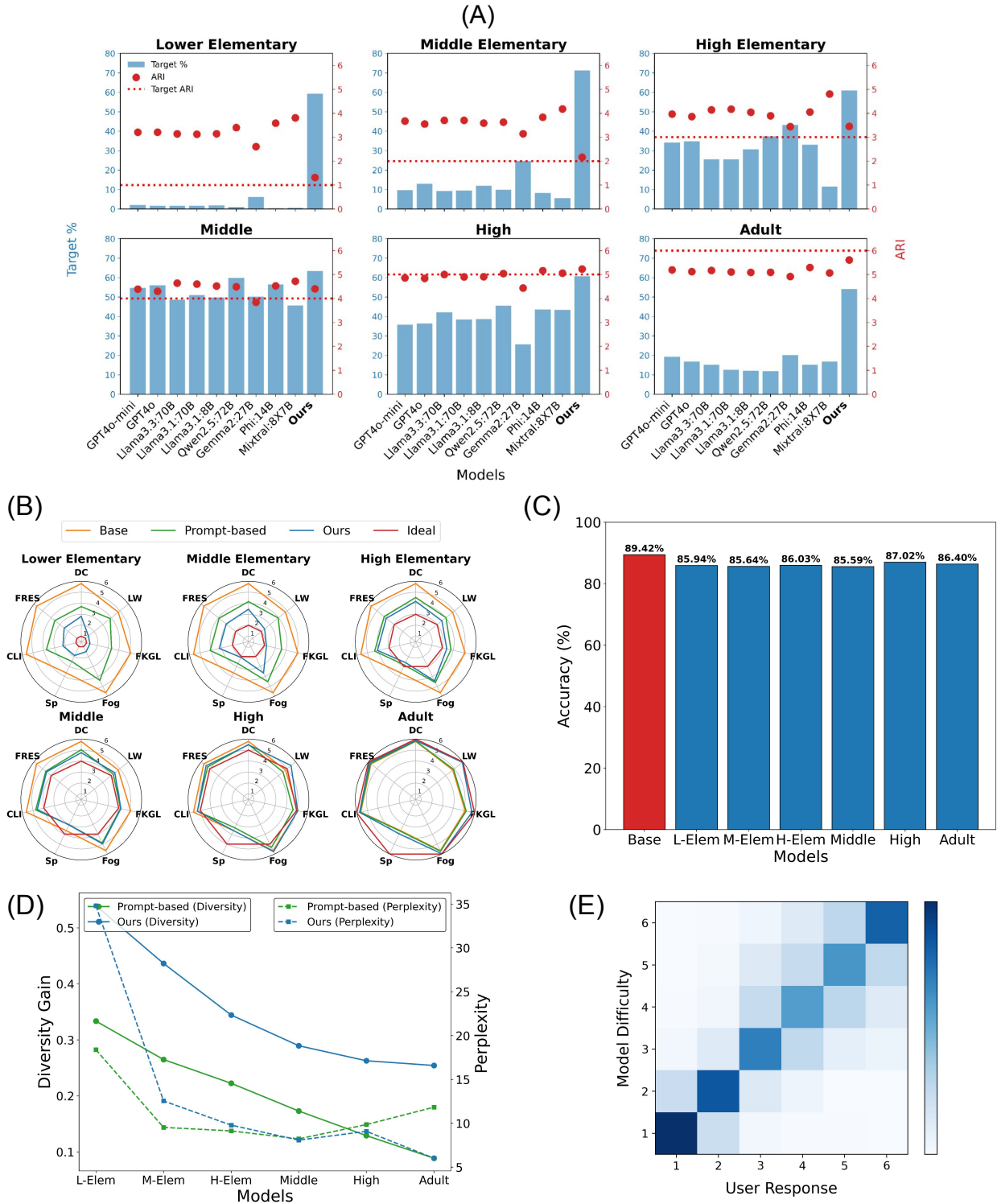


Figure 2: Results across evaluation criteria: (A) compatibility through integrated measure and ARI; (B) compatibility for each metric; (C) accuracy; (D) diversity gain and perplexity; and (E) survey results for type 1 questions.

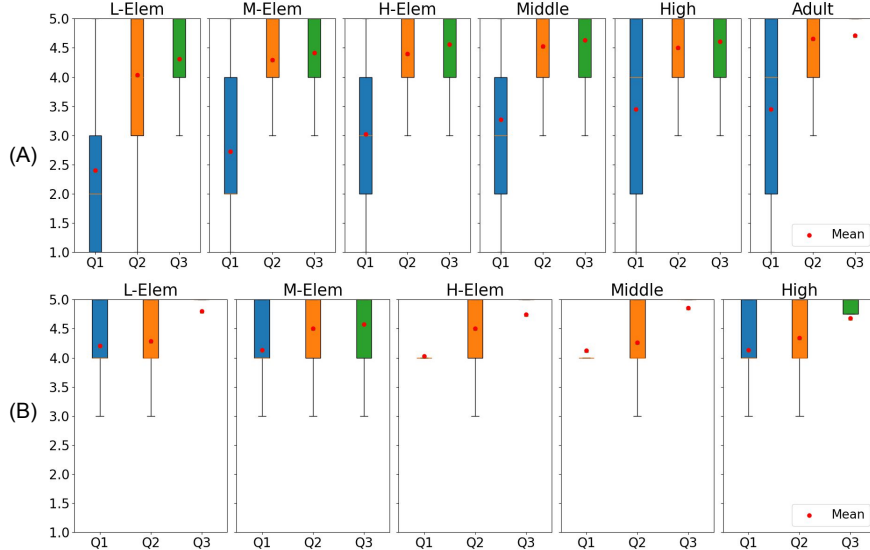


Figure 3: Survey results on Type 2 questions: (A) \mathcal{D}_{NQ} and (B) \mathcal{D}_{SQ2} . Box plots show Q1 (question difficulty), Q2 (answer comprehensibility), and Q3 (model accuracy) across grade levels, with mean values as red dots on a five-point scale. Higher Q1 and Q2 results indicate lower question and answer difficulty; higher Q3 results indicate stronger accuracy. While \mathcal{D}_{NQ} questions appear difficult for lower grades (low Q1 in A), answers remain comprehensible (high Q2 in A). Answer comprehensibility increases (higher Q2 in B vs. A) for grade-specific questions in \mathcal{D}_{SQ2} .

have completed high school, with most being undergraduate or graduate students who understand the relative difficulty across grade levels. We use two question types:

- Type 1: Each question includes six answers from different finetuned models ranging from lower elementary to adult. Participants assign each answer to a unique grade level, effectively ranking them.
- Type 2: Each question includes one answer from a finetuned model. Participants answer three five-point scale questions about question difficulty (Q1), answer comprehensibility (Q2), and answer accuracy (Q3).

For Type 1 questions, we measure the association between human-perceived difficulty rankings and model outputs using Kendall’s τ coefficient. The high coefficient of 0.76 across 108 participants demonstrates strong agreement between intended and perceived difficulty levels. We also compute L1 distances between participant rankings and ground truth ordering. For example, given six outputs sorted by ascending grade levels, the ground-truth ranking would be [1,2,3,4,5,6]. If a participant ranks them as [6,5,4,3,2,1], the L1 distances would be [5,3,1,1,3,5]. The L1 distances between rankings, [0.293, 0.398, 0.659, 0.693, 0.676, 0.578], all fall below 1. Combined with the high Kendall’s τ coefficient of 0.76, these results confirm our approach successfully generates grade-appropriate text aligned with human perception. Figure 2(E) shows dark cells along the diagonal, indicating strong alignment between model outputs and human perceptions.

For Type 2 questions, Figure 3 shows user scores on a five-point scale for Q1, Q2, and Q3 across two surveys. For \mathcal{D}_{NQ} , despite relatively difficult questions (low Q1 scores), models generate outputs comprehensible for each grade level. In post-survey feedback, participants noted that lower-grade models effectively explain concepts beyond their grade level using shorter, simpler sentences.

Regarding relatively lower Q2 scores (answer comprehensibility) for lower grades, participants attributed this to topic complexity rather than explanation difficulty. One noted, “Despite how clearly the model explains the concept of LLC sublayer in an operating system, lower elementary students will struggle with the concept itself, hence I give a 1 for Q2.” This observation is supported by higher Q2 scores in the second survey (\mathcal{D}_{SQ2}), where questions were designed for specific grade levels. High Q3 scores across all grade levels confirm our approach maintains factual correctness while adapting explanations.

We also evaluate surveys using GPT4o, following recent trends in using LLMs for human value alignment [47,

48]. Results from GPT4o align with human evaluations, further validating our approach (see Supplementary Section C.7, for more details).

Discussion

As students’ perspectives evolve with age, we investigate whether finetuned models develop distinct worldviews by examining model layers and outputs. We finetune LLaMA3.1:8B and compare results between lower elementary and adult models. Using logit lens [49] to analyze internal model layers, we observe that lower-grade models formulate ideas more directly and succinctly, while higher-grade models favor in-depth explanations as shown in Figure 4. This pattern mirrors human cognitive development, emphasizing clarity for younger students while preserving detail for advanced audiences. Plus, for lower-grade models, certain complex words (*e.g.*, “atmosphere”) are replaced by simpler synonyms (*e.g.*, “air”) and higher grade models internally show relatively complicated words such as “wavelength” or “dispersed”.

Moreover, analysis of vocabulary and sentence structure reveals that lower-grade models use simpler words and shorter sentences, while higher-grade models employ specialized terminology with longer explanations. These findings indicate our finetuning approach influences not only readability, but also the way models think and communicate, aligning each model with its target audience’s comprehension needs. Detailed visualizations appear in Supplementary Section D.

Our research addresses the critical global teacher shortage that impacts millions of children, aiming to improve educational equity. Our approach introduces a novel framework for training grade-specific LLMs to deliver age-appropriate educational content. These tools can supplement traditional teaching, providing personalized support to students of different grades and potentially increasing learning engagement worldwide. We believe that our work can contribute to a future where LLM-assisted learning can help mitigate educational disparities and create broader social benefits. The authors are responsible for all analyses and the final manuscript content.

While our approach successfully tailors textual complexity, it does not fully address conceptual difficulty. For example, even when written at a lower reading level, concepts like organizational culture may remain difficult for young students to comprehend based on their limited experience with the organizational world. Future work could incorporate domain-specific knowledge graphs or concept taxonomies to provide step-by-step explanations of challenging ideas. Combining readability metrics with conceptual difficulty frameworks would create truly adaptive LLMs that match both linguistic capacity and conceptual background.

Methods

We construct a grade-aligned question–answering dataset by generating diverse questions across eight educational fields ([Question Generation](#)) and prompting LLMs to produce answers with different readability levels ([Answer Generation](#)). We then assign grade-levels to the generated data using a novel integrated metric based on seven established readability formulas ([Readability Metrics Integration](#)) and finally fine-tune six grade-specific models to produce grade-appropriate responses ([Model Training](#)) and test the performance of the models across different datasets ([Datasets](#)). An overview of the full framework is illustrated in Figure 1.

Question Generation

We define eight educational fields based on K-12 curriculum frameworks: art, artificial intelligence, health education, literature, music, physical education, science, and social science. Each field contains five to eight subjects (see Supplementary Figure 1). To create a comprehensive question set, we prompt ChatGPT to generate sample questions answerable across all grades for each subject. Using these generated questions as few-shot demonstrations [8], we employ LLMs including Gemini [9], GPT [50], and LLaMA [10] to generate approximately 550 questions per subject.

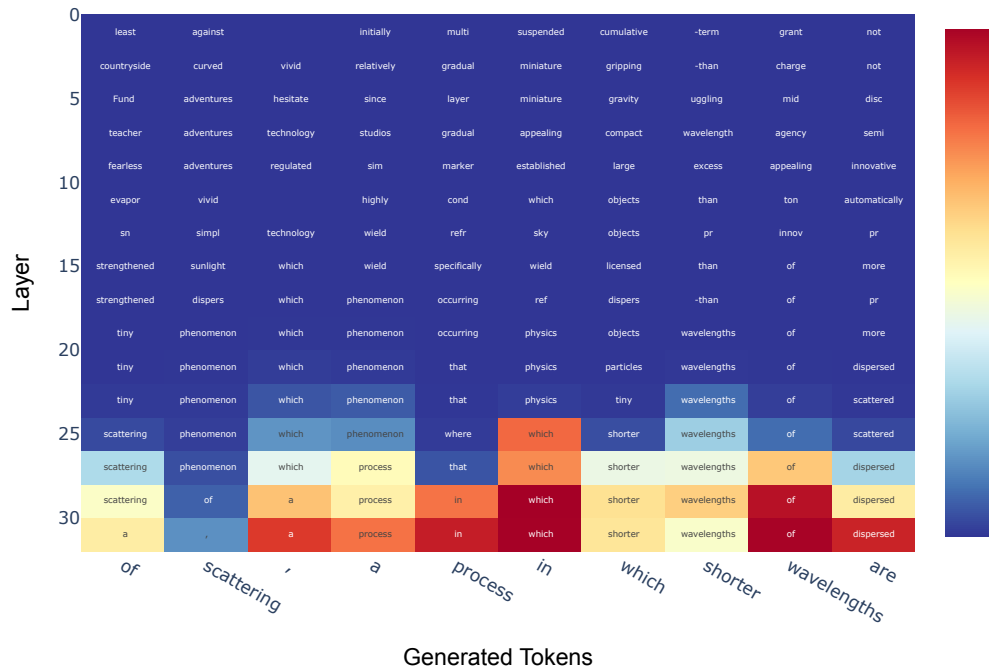
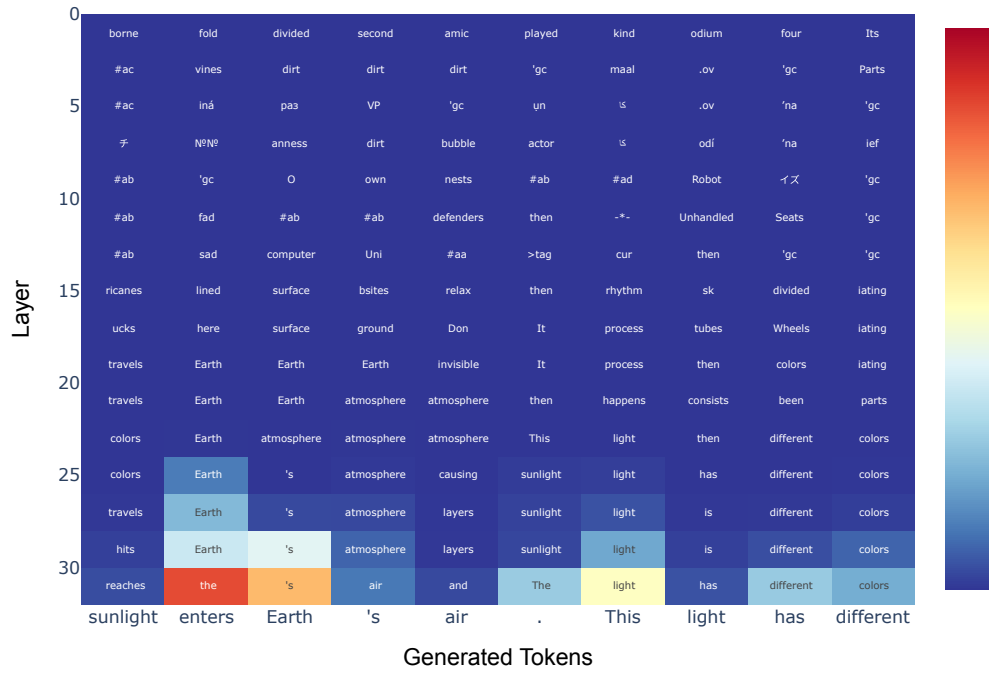


Figure 4: Logit-lens visualization for the lower-elementary (top) and adult (bottom) models of LLaMA3.1:8B on the prompt, “Why is the sky blue? The sky is blue because”. The bottom row for each figure shows the final output tokens, and each row above represents the top prediction at each transformer layer. Warmer colors (*e.g.*, red) indicate higher confidence.

Answer Generation

We design prompts to align with readability metrics and prior works [17, 20] by varying word difficulty, sentence length, and target audience. The prompt template follows: “Please provide the explanation in plain

text with no bullet points using <very easy, fairly easy, fairly difficult> words that <elementary school 1st grade, elementary school 3rd grade, elementary school 5th grade, middle school 7th grade, high school 10th grade, or college> students will know. Answer in detail with at a maximum of <4, 5, 6, 7, 8, 10, 12, 15, or 20> words per sentence.”. The output distribution for the corresponding prompts is shown in Supplementary Figure 3.

We vary grade level and maximum sentence length in the input prompts, creating 54 (6×9) distinct question-answer pairs for each question. Word difficulty matches grade level based on the Flesch Reading Ease Metric [28]. Using our integrated metric, we classify Q&A pairs into six grade levels. The distribution appears in Supplementary Table 1.

Readability Metrics Integration

We integrate seven readability metrics: Flesch Reading Ease (FRES) [28], Flesch-Kincaid Grade Level(FKGL) [29], the Coleman-Liau Index(CLI) [30], Linsear Write(LW) [31], the Gunning Fog Index(Fog) [32], Dale-Chall(DC) [33], and the Spache Readability Formula(Sp) [34].

Each metric captures distinct aspects of linguistic complexity through word difficulty, sentence length, and syntactic structure. Because their calculations differ substantially, we categorize them into three groups based on shared characteristics to create an integrated evaluation process:

- G_1 = Metrics utilizing predefined easy word lists (e.g., DC, Sp),
- G_2 = Metrics focusing on average sentence and word length (e.g., FRES, FKGL, CLI),
- G_3 = Metrics accounting for syllables per word (e.g., LW, Fog).

Each formula $f \in G_i$ provides a grade-level L_f . The final grade-level calculation follows Algorithm 1, which computes group-wise votes and determines the final grade through majority agreement or median value.

Algorithm 1: Grade-level Integration

Input: Grade-level estimates from G_1, G_2, G_3 , where each formula f produces a grade-level L_f .

Output: A final grade-level L_{final} .

```

foreach  $G_i$  in  $\{G_1, G_2, G_3\}$  do // Compute group-wise grade-level vote
  Compute the group’s grade-level vote  $L_{G_i}$  as:
      
$$L_{G_i} = \bigcap_{f \in G_i} L_f.$$

  if  $L_{G_i} \neq \emptyset$  then
    | Vote( $G_i$ )  $\leftarrow L_{G_i}$ ;
  else
    | Vote( $G_i$ )  $\leftarrow \min\{L_f \mid f \in G_i\}$ ;

// Determine final grade-level
Let  $\{L_{G_1}, L_{G_2}, L_{G_3}\}$  be the three computed votes;
if Two or more groups agree on the same grade-level then
  |  $L_{\text{final}} = \text{mode}\{L_{G_1}, L_{G_2}, L_{G_3}\}.$ 
else
  |  $L_{\text{final}} = \text{median}(\{L_{G_1}, L_{G_2}, L_{G_3}\}).$ 
return  $L_{\text{final}}$ .

```

The final result L_{final} belongs to the predefined set of grade levels:

$$\{[1, 2], [3, 4], [5, 6], [7, 8, 9], [10, 11, 12], [13+]\}.$$

Model Training

We perform supervised finetuning on GPT4o-mini via the OpenAI API to train six grade-specific models, spanning from lower elementary to adult (college+). We train each model on the corresponding subset of our grade-labeled question and answer corpus, classified based on our integrated readability metrics.

Datasets

We test our approach on real and synthetic datasets representing diverse grade levels. We use four datasets: ScienceQA (\mathcal{D}_{SQ}) [51], ELI5_Category (\mathcal{D}_{ELI}) [52], Natural Questions (\mathcal{D}_{NQ}) [53], and synthetic questions generated by GPT4o (\mathcal{D}_{GPT}).

- \mathcal{D}_{SQ} : ScienceQA comprises multiple-choice questions across 21 educational domains for grades 1-12. We use this dataset to measure accuracy and convert the questions to an open-ended format using GPT4o for compatibility testing (\mathcal{D}_{SQ2}). We sample 10,876 and 10,427 questions for \mathcal{D}_{SQ} and \mathcal{D}_{SQ2} respectively.
- \mathcal{D}_{ELI} : ELI5_Category contains questions from Reddit requiring explanatory multi-sentence answers. We sample 12,000 questions to evaluate compatibility for open-ended questions requiring detailed explanations.
- \mathcal{D}_{NQ} : Natural Questions contains real user questions submitted to Google search. We sample 24,000 questions to evaluate compatibility for naturally occurring questions.
- \mathcal{D}_{GPT} : We prompt GPT4o to create 740 questions across 54 subjects that require explanations and can be answered across all grade levels.

Acknowledgment

This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program)(RS-2024-00436680) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This project is supported by Microsoft Research Asia. Jindong Wang was partially supported by The Commonwealth Cyber Initiative (CCI) program (H-2Q25-020), William & Mary Faculty Research Award, and Modal Academic Compute Award. The authors acknowledge William & Mary Research Computing for providing computational resources and/or technical support that have contributed to the results reported within this paper. The authors thank Xing Xie and Lexin Zhou for fruitful discussions.

Author Contributions

J.O. and J.W. designed the main framework. J.O. proceeded with the experiments. All authors reviewed and wrote the main manuscript text.

Competing Interests

The authors declare no competing interests.

References

- [1] Li, Y., Wang, S., Ding, H. & Chen, H. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance, ICAIF '23*, 374–382 (Association for Computing Machinery, New York, NY, USA, 2023). URL <https://doi.org/10.1145/3604237.3626869>.
- [2] Lee, J., Stevens, N., Han, S. C. & Song, M. A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315* (2024).
- [3] Nazi, Z. A. & Peng, W. Large language models in healthcare and medical domain: A review (2024). URL <https://arxiv.org/abs/2401.06775>. 2401.06775.
- [4] Tam, T. Y. C. *et al.* A framework for human evaluation of large language models in healthcare derived from literature review (2024). URL <https://arxiv.org/abs/2405.02559>. 2405.02559.
- [5] Huang, Y. *et al.* Social science meets llms: How reliable are large language models in social simulations? (2024). URL <https://arxiv.org/abs/2410.23426>. 2410.23426.
- [6] Wang, S. *et al.* Large language models for education: A survey and outlook (2024). URL <https://arxiv.org/abs/2403.18105>. 2403.18105.
- [7] Gan, W., Qi, Z., Wu, J. & Lin, J. C.-W. Large language models in education: Vision and opportunities (2023). URL <https://arxiv.org/abs/2311.13160>. 2311.13160.
- [8] OpenAI. Gpt-4 technical report (2024). URL <https://arxiv.org/abs/2303.08774>. 2303.08774.
- [9] Team, G. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024). URL <https://arxiv.org/abs/2403.05530>. 2403.05530.
- [10] team, L. The llama 3 herd of models (2024). URL <https://arxiv.org/abs/2407.21783>. 2407.21783.
- [11] UNESCO. 244m children won't start the new school year (2022). URL <https://www.unesco.org/gem-report/en/articles/244m-children-wont-start-new-school-year>. Accessed: 2025-02-09.
- [12] National Center for Education Statistics. Public school leaders reports suggest no overall change in the number of teaching positions since last school year (2024). URL https://nces.ed.gov/whatsnew/press_releases/12_12_2024.asp. Accessed: 2025-02-08.
- [13] García, E. & Weiss, E. The teacher shortage is real, large and growing, and worse than we thought. the first report in " the perfect storm in the teacher labor market" series. *Economic policy institute* (2019).
- [14] Shah, D., Amin, N., Kakli, M. B., Piracha, Z. F. & Zia, M. A. Pakistan education statistics 2016-17. *National Education Management Information System (NEMIS)*. *Academy of Educational Planning and Management (AEPAM)* (2018).
- [15] Khan, M. J. & Ahmed, J. Child education in the time of pandemic: Learning loss and dropout. *Children and Youth Services Review* **127**, 106065 (2021).
- [16] SAGE Publishing. World teachers' day: The teachers we need for the education we want: The global imperative to reverse the teacher shortage (2024). URL <https://www.sagepub.com/explore-our-content/blogs/posts/asia-pacific-insights/2024/11/25/world-teachers-day-the-teachers-we-need-for-the-education-we-want-the-global-imperative-to-reverse-the-teacher-shortage>. Accessed: 2025-02-19.
- [17] Rooein, D., Curry, A. C. & Hovy, D. Know your audience: Do llms adapt to different age and education levels? (2023). URL <https://arxiv.org/abs/2312.02065>. 2312.02065.
- [18] Imperial, J. M. & Tayyar Madabushi, H. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, 205–223 (Association for Computational Linguistics, Singapore, 2023). URL <https://aclanthology.org/2023.gem-1.18/>.

- [19] Gobara, S., Kamigaito, H. & Watanabe, T. Do llms implicitly determine the suitable text difficulty for users? (2024). URL <https://arxiv.org/abs/2402.14453>. 2402.14453.
- [20] Hsu, Y.-S., Feldhus, N. & Hakimov, S. Free-text rationale generation under readability level control. *ArXiv* **abs/2407.01384** (2024). URL <https://api.semanticscholar.org/CorpusID:270870139>.
- [21] Roegiest, A. & Pinkosova, Z. Generative information systems are great if you can read. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval, CHIIR '24*, 165–177 (Association for Computing Machinery, New York, NY, USA, 2024). URL <https://doi.org/10.1145/3627508.3638345>.
- [22] Malik, A., Mayhew, S., Piech, C. & Bicknell, K. From tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, 15670–15693 (Association for Computational Linguistics, Bangkok, Thailand, 2024). URL <https://aclanthology.org/2024.findings-acl.926/>.
- [23] Tran, H., Yao, Z., Li, L. & Yu, H. Readctrl: Personalizing text generation with readability-controlled instruction learning (2024). URL <https://arxiv.org/abs/2406.09205>. 2406.09205.
- [24] Ribeiro, L. F. R., Bansal, M. & Dreyer, M. Generating summaries with controllable readability levels. In *The 2023 Conference on Empirical Methods in Natural Language Processing* (2023). URL <https://openreview.net/forum?id=IFNbElsnCi>.
- [25] Chall, J. S. Stages of reading development (1983).
- [26] (US), N. R. P., of Child Health, N. I. & (US), H. D. *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (National Institute of Child Health and Human Development, National ... , 2000).
- [27] Leibo, J. Z. *et al.* A theory of appropriateness with applications to generative artificial intelligence. *arXiv preprint arXiv:2412.19010* (2024).
- [28] Flesch, R. A new readability yardstick. *Journal of applied psychology* **32**, 221 (1948).
- [29] Kincaid, J. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Chief of Naval Technical Training* (1975).
- [30] Coleman, M. & Liau, T. L. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* **60**, 283 (1975).
- [31] O’hayre, J. *Gobbledygook has gotta go* (US Department of the Interior, Bureau of Land Management, 1966).
- [32] Gunning, R. *The technique of clear writing* (1952).
- [33] Chall, J. S. & Dale, E. Readability revisited: The new dale-chall readability formula. (*No Title*) (1995).
- [34] Spache, G. A new readability formula for primary-grade reading materials. *The Elementary School Journal* **53**, 410–413 (1953).
- [35] Li, Z., Zhu, H., Lu, Z. & Yin, M. Synthetic data generation with large language models for text classification: Potential and limitations (2023). URL <https://arxiv.org/abs/2310.07849>. 2310.07849.
- [36] Long, L. *et al.* On llms-driven synthetic data generation, curation, and evaluation: A survey (2024). URL <https://arxiv.org/abs/2406.15126>. 2406.15126.
- [37] Peng, B., Li, C., He, P., Galley, M. & Gao, J. Instruction tuning with gpt-4 (2023). URL <https://arxiv.org/abs/2304.03277>. 2304.03277.

- [38] Chen, Z., Deng, Y., Yuan, H., Ji, K. & Gu, Q. Self-play fine-tuning converts weak language models to strong language models (2024). URL <https://arxiv.org/abs/2401.01335>. 2401.01335.
- [39] Taori, R. *et al.* Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023).
- [40] Mitra, A., Khanpour, H., Rosset, C. & Awadallah, A. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830* (2024).
- [41] Jiang, A. Q. *et al.* Mixtral of experts (2024). URL <https://arxiv.org/abs/2401.04088>. 2401.04088.
- [42] Smith, E. A. & Senter, R. *Automated readability index*, vol. 66 (Aerospace Medical Research Laboratories, Aerospace Medical Division, Air . . . , 1967).
- [43] Bilmes, J. Submodularity in machine learning and artificial intelligence (2022). URL <https://arxiv.org/abs/2202.00132>. 2202.00132.
- [44] Wang, Y. *et al.* Two-stage llm fine-tuning with less specialization and more generalization (2024). URL <https://arxiv.org/abs/2211.00635>. 2211.00635.
- [45] Dodge, J. *et al.* Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305* (2020).
- [46] Luo, Y. *et al.* An empirical study of catastrophic forgetting in large language models during continual fine-tuning (2025). URL <https://arxiv.org/abs/2308.08747>. 2308.08747.
- [47] Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena (2023). URL <https://arxiv.org/abs/2306.05685>. 2306.05685.
- [48] Li, Z. *et al.* Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 16028–16045 (Association for Computational Linguistics, Miami, Florida, USA, 2024). URL <https://aclanthology.org/2024.emnlp-main.896/>.
- [49] Belrose, N. *et al.* Eliciting latent predictions from transformers with the tuned lens (2023). 2303.08112.
- [50] Brown, T. *et al.* Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, vol. 33, 1877–1901 (Curran Associates, Inc., 2020). URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- [51] Lu, P. *et al.* Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)* (2022).
- [52] Gao, J., Zhou, Q. & Qiu, R. ELI5-Category: a categorized open-domain qa dataset (2021).
- [53] Kwiatkowski, T. *et al.* Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics* (2019).
- [54] Balthazar, C. H. & Scott, C. M. Sentences are key: Helping school-age children and adolescents build sentence skills needed for real language. *American Journal of Speech-Language Pathology* **33**, 564–579 (2024).
- [55] Zipf, G. K. *The psycho-biology of language: An introduction to dynamic philology* (Routledge, 2013).
- [56] Robertson, S. E. *et al.* Okapi at trec-3. *Nist Special Publication Sp* **109**, 109 (1995).

Supplementary Information

A Related Work

LLMs are getting huge interest as a tool for applications on various educational scenarios [6]. However, recent studies show that current LLMs fail to output grade-appropriate answers [19] regardless of how informative the prompt is [18]. Current models consistently default to a high school [20] or college+ level [17], even when explicitly prompted to target a lower grade level. The root of this challenge lies in the training data itself, as foundational datasets such as Alpaca and Dolly exhibit a strong bias towards college+ level [21].

Beyond prompts, prior research has explored fine-tuning to align with different grade-levels. However, these efforts are confined to text simplification tasks that require a source text (e.g., summarization and paraphrasing) [22, 24] and have shown poor performance for lower grades [23]. Furthermore, the works have been limited to incorporating only a few educational metrics, given the difficulty of optimizing for multiple criteria simultaneously. In contrast, our work targets open-ended question answering, simulating a more realistic classroom setting of direct student-LLM interaction. By incorporating seven educational metrics, we achieve strong performance across all grade-levels, proving effective even for the lower-elementary level.

B Metrics

B.1 Educational Metrics

We compute text readability using multiple readability metrics. As each metric has different methods for mapping scores to grade levels, we standardize the results by classifying them into six predefined grade levels: 1 (lower elementary), 2 (middle elementary), 3 (high elementary), 4 (middle), 5 (high), and 6 (adult).

B.1.1 Flesch Reading Ease

The Flesch reading-ease score (FRES) [28], used by the U.S. Department of Defense, provides a numerical measure of text readability:

$$\text{FRES} = 206.835 - 1.015 \times \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \times \left(\frac{\text{total syllables}}{\text{total words}} \right).$$

Higher FRES values correspond to text that is easier to read.

B.1.2 Flesch–Kincaid grade level

The Flesch–Kincaid Grade Level (FKGL) [29] formula is widely employed in education to estimate the U.S. grade level at which a text can be understood. It is calculated as:

$$\text{FKGL} = 0.39 \times \left(\frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \times \left(\frac{\text{total syllables}}{\text{total words}} \right) - 15.59.$$

Higher FKGL scores correspond to texts requiring more advanced reading skills.

B.1.3 Coleman–Liau Index

The Coleman–Liau Index (CLI) [30] quantifies text readability by analyzing the average number of letters and sentences per 100 words. It is given by:

$$\text{CLI} = 0.0588 \times L - 0.296 \times S - 15.8,$$

where L is the average number of letters per 100 words and S is the average number of sentences per 100 words. Higher CLI scores correspond to texts requiring more advanced reading skills.

B.1.4 Linsear Write

Linsear Write (LW) [31] is a readability test developed for the U.S. Air Force to evaluate how easily technical manuals can be read. It is computed over a 100-word passage as follows:

1. Let E be the number of “easy” words (words with at most two syllables), and H be the number of “hard” words (three or more syllables).
2. Let S be the number of sentences in the 100-word sample.
3. Compute the intermediate score i :

$$i = \frac{E + 3H}{S}.$$

4. Determine the final Linsear Write score (LW):

$$\text{LW} = \begin{cases} \frac{i}{2}, & \text{if } i > 20, \\ \frac{i-2}{2}, & \text{otherwise.} \end{cases}$$

Higher LW scores correspond to texts requiring more advanced reading skills.

B.1.5 Gunning–Fog Index

The Gunning–Fog Index (Fog) [32] estimates the years of formal education required to comprehend a passage on the first reading. This index is defined as:

$$\text{Fog} = 0.4 \left(\frac{W}{S} + 100 \frac{C}{W} \right).$$

where W is the total number of words, S is the total number of sentences, and C is the number of “complex” words with three or more syllables. Higher Fog values indicate text requiring more advanced reading skills.

B.1.6 Dale–Chall Readability Formula

The Dale–Chall readability formula (DC) [33] assesses text complexity based on the proportion of unfamiliar words and sentence length, using a predefined list of 3000 words familiar to 80% of fifth-grade students. It is defined as:

$$\text{DC}_{\text{raw}} = 0.1579p + 0.0496 \frac{W}{S}.$$

The final (adjusted) score is:

$$\text{DC} = \begin{cases} \text{DC}_{\text{raw}} + 3.6365, & \text{if } p > 5, \\ \text{DC}_{\text{raw}}, & \text{otherwise.} \end{cases}$$

where D is the number of difficult words (words not in the predefined list), W is the total number of words, S is the total number of sentences, and p is the percentage of difficult words ($\frac{D}{W} \times 100$). Higher DC values indicate text requiring more advanced reading skills.

B.1.7 Spache Readability Formula

The Spache readability formula (Sp) [34] tests text readability for younger children. It uses a predefined list of approximately 1000 words and is defined as:

$$\text{Sp} = 0.141 S + 0.086 U + 0.839.$$

where S is the average sentence length (words per sentence) and U is the percentage of unique unfamiliar words (words not in the predefined list). Higher Sp values indicate text requiring more advanced reading skills.

B.1.8 Automated Readability Index

The Automated Readability Index (ARI) [42] estimates the U.S. grade level required to comprehend a text. It is defined as:

$$\text{ARI} = 4.71 \frac{C}{W} + 0.5 \frac{W}{S} - 21.43.$$

where C is the total number of characters (letters and digits), W is the total number of words, and S is the total number of sentences. Higher ARI scores indicate text requiring more advanced reading skills.

B.2 Perplexity

Perplexity (ppl) predicts the probability of a token w_i given its preceding context w_1, \dots, w_{i-1} . The perplexity is defined as:

$$\text{ppl} = \exp\left(-\frac{1}{N} \sum_{i=1}^N \log p(w_i | w_1, \dots, w_{i-1})\right),$$

where N is the number of tokens in the text dataset with \log as the natural logarithm. Higher perplexity implies a broader probability spread across potential outcomes, indicating more diverse outputs. We use GPT2-XL for computing the probability.

B.3 Diversity Gain

Diversity gain [43] quantifies the extent to which a generated dataset brings diversity to the original (base) dataset. The base dataset is defined as

$$D_{\text{base}} = \{x_i = (q_i, r_i, a_i)\}_{i=1}^N$$

with N samples. The new generated dataset is defined as

$$D_{\text{new}} = \{x_i = (q_i, r_i, a_i)\}_{i=1}^M$$

with M samples. The diversity gain of D_{new} relative to D_{base} is

$$d_{\text{gain}} = \frac{1}{M} \sum_{x_i \in D_{\text{new}}} \min_{x_j \in D_{\text{base}}} \|f(x_i) - f(x_j)\|,$$

where f is the feature extractor. We use D_{base} as the collection of text outputs from the base model (GPT4o-mini) before finetuning and D_{new} as the collection of text outputs from the grade-level finetuned models for all four datasets. We use `text-embedding-3-large` from OpenAI to extract the features.

C Extended Results

C.1 Fields and Subjects for Training Text Data

We present the fields and subjects of questions used for finetuning based on the K-12 curriculum.

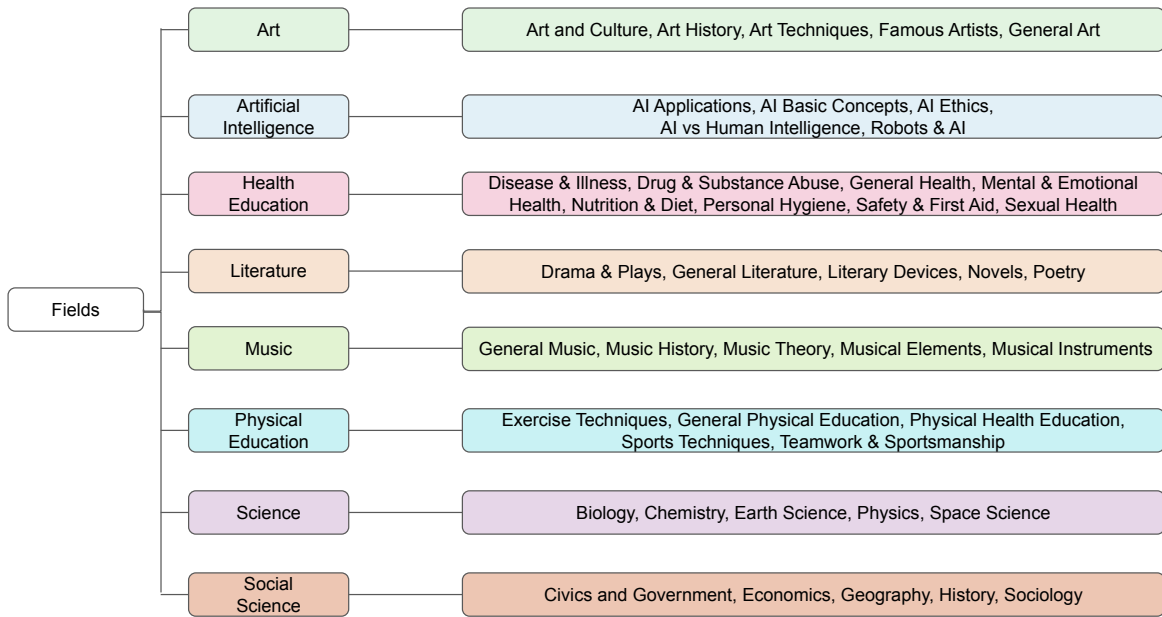


Figure 5: Fields and subjects of questions used for finetuning.

C.2 Number of Q&A for each Grade-level

Table 1 provides a detailed distribution of question & answer pairs across the six grade-levels. We collect a total of 1,287,149 pairs. The amount of data for lower elementary is comparatively smaller than those for other grad-levels, due to the relatively strict thresholds set by the educational metrics. However, the absolute number of data remains substantial enough to facilitate effective finetuning.

L-Elem	M-Elem	H-Elem	Middle	High	Adult	Total
71,524	324,910	302,485	275,131	130,779	182,590	1,287,149

Table 1: Number of question & answer pairs used for finetuning different models.

C.3 Extended Output Distribution

We compare the output grade-level distributions from the prompt-based approach and ours in Figure 6. Overall, our approach shows a diagonal pattern, indicating stronger alignment with the intended grade levels than the prompt-based approach.

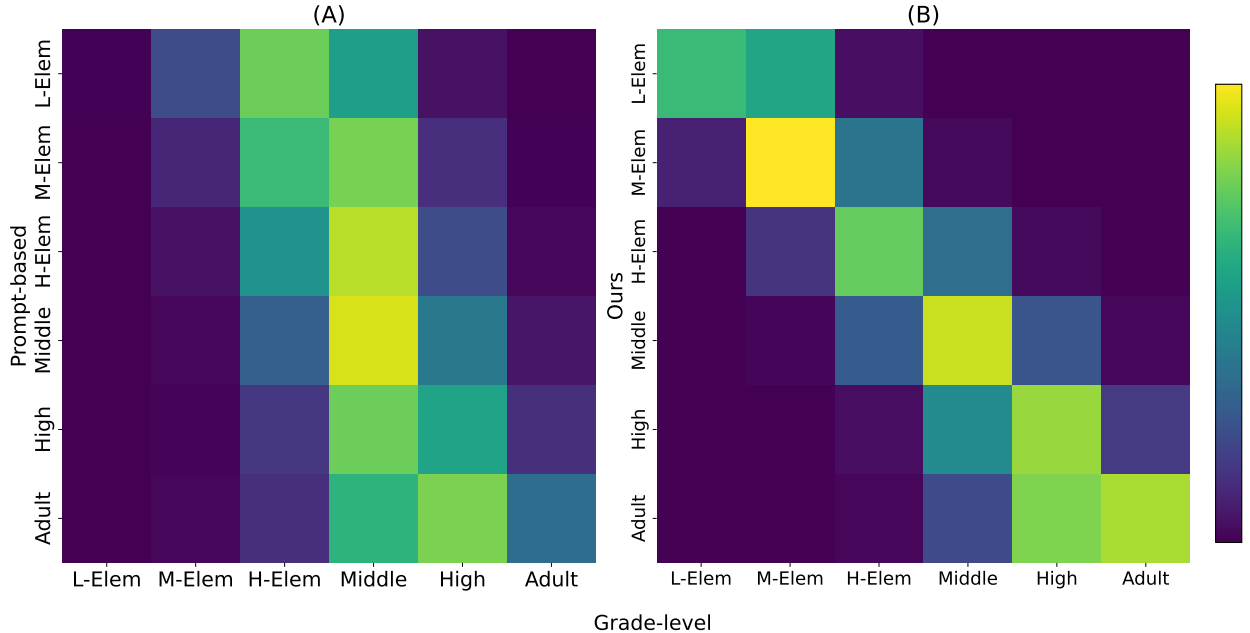


Figure 6: Comparison of output grade-level distributions from the prompt-based approach (A) and our method (B). The rows represent the targeted grade-level (prompt or model), and the columns show the classified grade-levels. Warmer colors (e.g., yellow) represent higher frequencies. For instance, the top row of (B) reveals the output classification results of the lower elementary model’s responses.

C.4 Text Distribution in Data Generation for Different Prompts

We show the output grade-level distribution across different prompts, which are used for data generation. No single prompt consistently aligns with the intended grade level, highlighting the need for our proposed finetuning method.

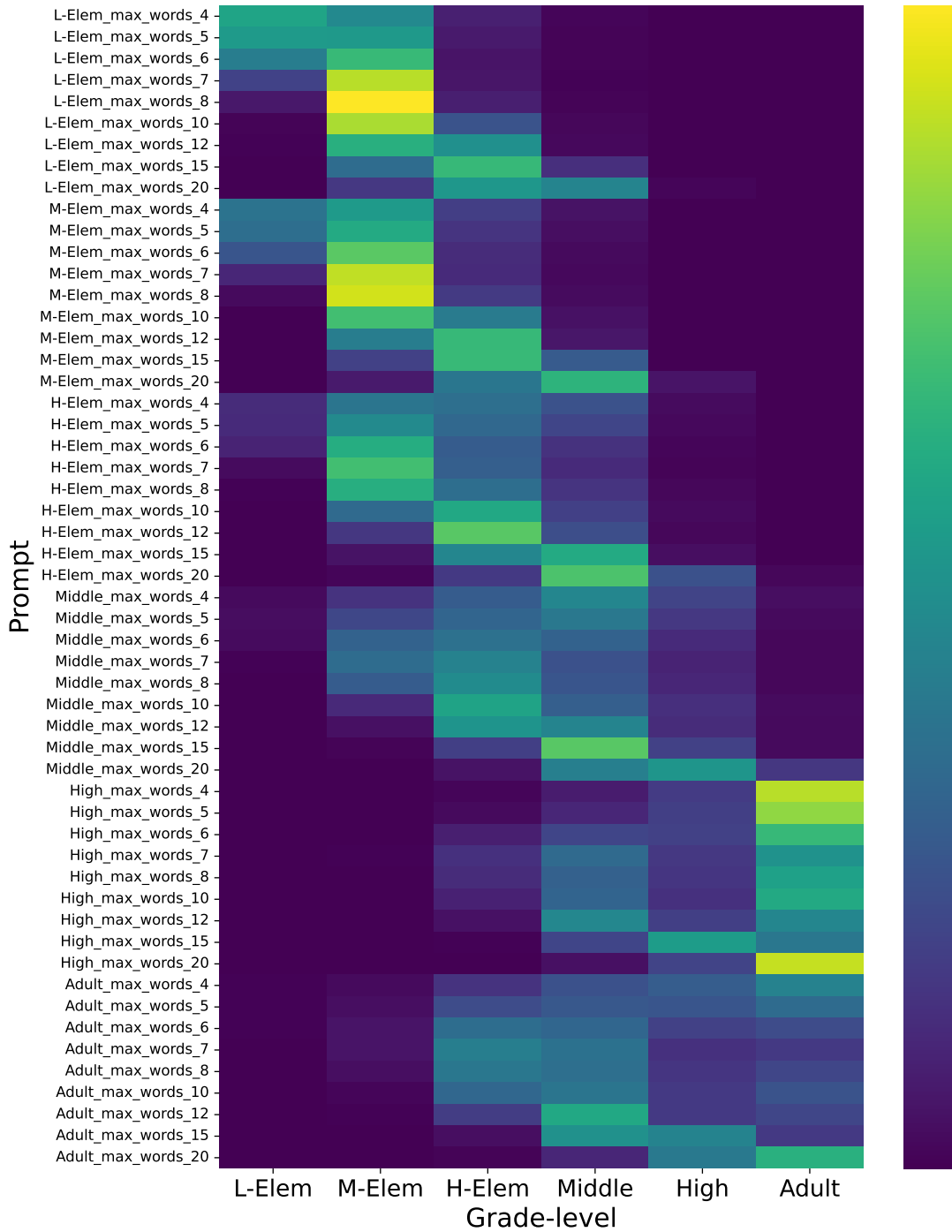


Figure 7: Grade distribution for different prompts in data generation. Each row represents a specific prompt, while the columns indicate the classified grade levels.

C.5 Compatibility Results for Base LLMs

In Table 2, we compare the grade-level distribution of outputs among different base LLMs and their ARI scores. The results show that the current state-of-the-art LLMs tend to output with higher-grade readability, highlighting the challenge of generating content suitable for lower grade students.

Model	Dataset	L-Elem (1)	M-Elem (2)	H-Elem (3)	Middle (4)	High (5)	Adult (6)	ARI
GPT4o-mini	\mathcal{D}_{GPT}	0.00	0.00	7.57	44.59	34.73	13.11	5.25
	\mathcal{D}_{SQ2}	0.67	3.99	8.02	31.53	43.81	14.74	4.21
	\mathcal{D}_{ELI}	0.00	0.08	1.59	37.93	49.26	11.14	5.24
	\mathcal{D}_{NQ}	0.03	1.14	11.00	35.69	36.78	15.37	4.82
GPT4o	\mathcal{D}_{GPT}	0.00	0.27	7.16	41.35	35.27	15.95	X
	\mathcal{D}_{SQ2}	1.15	3.09	10.96	32.91	45.51	13.63	X
	\mathcal{D}_{ELI}	0.00	0.08	1.67	38.09	47.49	12.67	X
	\mathcal{D}_{NQ}	0.07	1.33	11.55	36.86	35.84	14.34	X
LLaMA 3.3:70B	\mathcal{D}_{GPT}	0.00	0.68	11.89	46.35	32.16	8.92	5.06
	\mathcal{D}_{SQ2}	0.54	3.10	12.58	27.58	47.29	13.26	4.87
	\mathcal{D}_{ELI}	0.00	0.06	1.99	30.21	50.26	17.47	5.33
	\mathcal{D}_{NQ}	0.12	0.96	8.35	35.37	40.83	14.38	4.98
LLaMA 3.1:70B	\mathcal{D}_{GPT}	0.00	0.14	13.38	51.35	28.51	6.62	5.04
	\mathcal{D}_{SQ2}	1.16	2.72	10.16	26.73	44.47	16.41	4.92
	\mathcal{D}_{ELI}	0.00	0.08	2.44	33.64	48.51	15.32	5.35
	\mathcal{D}_{NQ}	0.15	0.86	7.95	34.60	41.20	15.24	4.98
LLaMA 3.1:8B	\mathcal{D}_{GPT}	0.00	0.41	10.95	48.92	32.16	7.57	5.09
	\mathcal{D}_{SQ2}	0.92	2.65	11.45	31.67	44.76	12.82	4.80
	\mathcal{D}_{ELI}	0.00	0.06	2.27	33.11	48.89	15.68	5.33
	\mathcal{D}_{NQ}	0.12	0.65	7.47	35.38	42.05	14.34	4.98
Qwen 2.5:72B	\mathcal{D}_{GPT}	0.00	0.00	6.22	39.32	42.03	12.43	5.24
	\mathcal{D}_{SQ2}	0.25	3.30	10.44	31.98	49.78	9.60	4.85
	\mathcal{D}_{ELI}	0.00	0.03	1.30	33.83	53.59	11.25	5.28
	\mathcal{D}_{NQ}	0.01	0.27	5.76	33.67	44.16	16.13	5.07
Gemma 2:27B	\mathcal{D}_{GPT}	0.00	0.41	8.24	29.73	37.43	24.19	5.29
	\mathcal{D}_{SQ2}	3.73	8.61	22.40	35.43	34.88	9.53	4.42
	\mathcal{D}_{ELI}	0.01	0.19	2.67	28.87	45.03	23.23	5.37
	\mathcal{D}_{NQ}	0.36	9.07	21.81	31.78	24.43	12.56	4.25
Phi 4:14B	\mathcal{D}_{GPT}	0.00	0.00	4.87	36.89	42.16	13.24	5.44
	\mathcal{D}_{SQ2}	0.00	1.63	7.12	26.65	40.76	16.66	5.16
	\mathcal{D}_{ELI}	0.00	0.00	0.83	23.39	58.65	17.13	5.58
	\mathcal{D}_{NQ}	0.03	0.70	8.56	30.44	39.13	21.15	5.04
Mixtral 8x7B	\mathcal{D}_{GPT}	0.00	1.36	15.20	37.04	29.44	16.96	5.07
	\mathcal{D}_{SQ2}	2.33	2.95	14.40	36.30	44.15	9.79	4.72
	\mathcal{D}_{ELI}	0.01	0.16	4.60	31.31	44.97	18.95	5.24
	\mathcal{D}_{NQ}	0.14	0.93	10.39	36.81	39.22	12.52	4.88

Table 2: Target % and ARI values for various base LLMs across four datasets. Each column (lower elementary to adult) shows the proportion of outputs classified at that grade level, while the final column presents the ARI value. Bolded entries indicate the highest proportion for each model and dataset pair.

C.6 Compatibility Results for Prior Approaches and Ours

We provide a comprehensive breakdown of grade-level output distribution for several models across four datasets, comparing prompt-based approach and ours. Table 3 includes both the percentage of outputs correctly aligned with each target grade level (Target %) and the ARI values. Notably, our finetuned models often achieve stronger alignment at the targeted grade-levels.

		L-Elem (1)		M-Elem (2)		H-Elem (3)		Middle (4)		High (5)		Adult (6)		
	Model	Dataset	Target %	ARI	Target %	ARI	Target %	ARI	Target %	ARI	Target %	ARI	Target %	ARI
Prompt Based	GPT4o mini (Baseline)	\mathcal{D}_{GPT}	0.13	3.35	11.22	3.77	36.22	3.99	50.68	4.36	27.30	4.87	22.16	5.40
		\mathcal{D}_{SQ2}	8.04	2.67	16.80	3.40	37.28	3.84	54.30	4.48	47.32	5.10	21.95	5.13
		\mathcal{D}_{ELI}	0.05	3.51	5.36	3.77	37.99	3.96	64.18	4.32	32.58	4.82	14.96	5.35
		\mathcal{D}_{NQ}	0.25	3.29	5.38	3.75	25.42	4.07	49.94	4.40	35.55	4.66	18.00	4.90
	GPT4o	\mathcal{D}_{GPT}	0.27	3.39	14.46	3.61	31.62	3.89	52.70	4.26	35.14	4.96	21.62	5.31
		\mathcal{D}_{SQ2}	4.70	2.98	14.17	3.60	32.33	3.95	52.39	4.51	48.74	5.14	16.95	5.04
		\mathcal{D}_{ELI}	0.13	3.49	8.71	3.63	41.29	3.83	62.38	4.21	26.47	4.60	13.06	5.22
		\mathcal{D}_{NQ}	1.20	2.97	14.31	3.36	34.09	3.78	56.67	4.22	35.34	4.67	15.53	4.92
	LLaMA 3.3:70B	\mathcal{D}_{GPT}	0.14	3.22	10.81	3.65	31.89	4.04	48.51	4.44	29.19	4.81	12.57	5.19
		\mathcal{D}_{SQ2}	6.10	2.48	17.69	3.38	34.50	3.89	47.81	4.68	49.86	5.16	14.66	5.06
		\mathcal{D}_{ELI}	0.02	3.54	2.56	3.92	18.30	4.29	53.89	4.73	47.74	5.11	18.87	5.35
		\mathcal{D}_{NQ}	0.59	3.31	6.04	3.86	17.73	4.34	44.25	4.72	41.80	4.93	14.53	5.08
	LLaMA 3.1:70B	\mathcal{D}_{GPT}	0.27	3.20	10.68	3.69	34.60	4.03	53.51	4.39	24.73	4.75	9.19	5.10
		\mathcal{D}_{SQ2}	5.37	2.49	18.98	3.35	31.09	3.96	49.43	4.65	46.57	4.96	11.74	4.97
		\mathcal{D}_{ELI}	0.03	3.52	2.76	3.87	19.89	4.29	56.10	4.70	43.50	5.06	13.99	5.30
		\mathcal{D}_{NQ}	0.66	3.28	5.40	3.90	16.32	4.40	45.08	4.67	39.04	4.84	15.45	5.07
	LLaMA 3.1:8B	\mathcal{D}_{GPT}	0.14	3.26	13.67	3.58	40.96	3.92	46.22	4.35	27.70	4.83	10.95	5.18
		\mathcal{D}_{SQ2}	6.80	2.50	23.76	3.17	38.69	3.78	49.52	4.48	44.19	4.91	10.21	4.85
		\mathcal{D}_{ELI}	0.02	3.49	4.10	3.77	24.85	4.16	57.44	4.64	43.17	5.04	12.85	5.27
		\mathcal{D}_{NQ}	0.49	3.32	6.10	3.82	17.98	4.31	45.94	4.61	39.77	4.86	14.12	5.06
	Qwen 2.5:72B	\mathcal{D}_{GPT}	0.00	3.59	9.87	3.70	42.97	3.80	63.38	4.43	37.97	5.07	13.24	5.27
		\mathcal{D}_{SQ2}	4.16	2.92	19.10	3.33	45.65	3.62	59.90	4.42	51.51	5.03	9.81	4.86
		\mathcal{D}_{ELI}	0.03	3.57	5.19	3.68	40.64	3.87	68.9	4.44	49.29	5.10	8.88	5.22
		\mathcal{D}_{NQ}	0.10	3.52	5.07	3.80	20.16	4.28	47.16	4.67	43.34	4.97	15.57	5.03
Gemma 2:27B	\mathcal{D}_{GPT}	1.14	2.97	17.16	3.36	41.76	3.64	54.73	3.94	25.14	4.55	29.46	5.38	
	\mathcal{D}_{SQ2}	15.53	1.92	37.91	2.91	47.49	3.28	46.32	3.74	28.12	4.47	12.65	4.65	
	\mathcal{D}_{ELI}	0.13	3.19	11.99	3.53	43.18	3.76	59.49	4.11	27.61	4.71	25.50	5.42	
	\mathcal{D}_{NQ}	7.58	2.34	31.97	2.77	40.49	3.08	40.26	3.58	21.74	4.02	12.70	4.22	
Phi 4:14B	\mathcal{D}_{GPT}	0.00	3.72	9.87	3.86	39.46	4.02	55.27	4.35	31.49	5.07	12.57	5.43	
	\mathcal{D}_{SQ2}	1.71	3.05	16.12	3.49	39.95	3.81	58.25	4.52	48.29	5.20	12.65	5.07	
	\mathcal{D}_{ELI}	0.00	3.84	3.48	3.93	35.27	4.05	67.16	4.47	46.45	5.18	12.56	5.51	
	\mathcal{D}_{NQ}	0.10	3.72	3.47	4.05	17.51	4.32	45.15	4.77	47.91	5.20	23.06	5.17	
Mixtral :8x7B	\mathcal{D}_{GPT}	0.68	3.83	7.57	3.97	18.02	4.62	50.81	4.54	34.39	4.92	21.94	5.22	
	\mathcal{D}_{SQ2}	1.96	3.16	10.60	3.87	15.96	4.55	45.54	4.61	45.73	5.03	11.17	4.75	
	\mathcal{D}_{ELI}	0.02	4.28	2.00	4.41	4.81	5.10	44.56	4.93	48.85	5.25	20.65	5.35	
	\mathcal{D}_{NQ}	0.09	3.96	2.16	4.46	7.01	4.94	41.49	4.81	44.66	5.04	13.32	4.95	
Ours	Finetuned GPT4o mini	\mathcal{D}_{GPT}	67.30	1.32	82.97	2.14	73.11	3.42	70.54	4.56	63.24	5.33	73.78	5.40
		\mathcal{D}_{SQ2}	65.99	1.21	67.08	2.20	52.74	3.56	58.83	4.45	65.40	5.26	43.16	5.67
		\mathcal{D}_{ELI}	57.41	1.36	72.82	2.24	60.00	3.55	65.18	4.38	64.68	5.34	66.16	5.88
		\mathcal{D}_{NQ}	46.65	1.37	61.90	2.07	58.21	3.27	58.71	4.23	49.07	4.99	33.54	5.49

Table 3: Extended performance comparison across multiple models, grade levels, and datasets. Each cell displays the Target% and the corresponding ARI value. Bold values indicate the highest performance for a particular dataset-grade combination.

C.7 Survey Results with GPT

We present the results on the survey results from GPT4o. We feed GPT4o with the same set of survey questions posed to human participants. For type 1 questions, we observe Kendall’s τ coefficient as 0.72 and an L1 distance of [0.419, 0.585, 0.635, 0.831, 0.724, 0.601], along with a strong alignment between model outputs and GPT4o’s perceptions (see Figure 8). For type 2 questions, answer comprehensibility and accuracy consistently remains high regardless of the question difficulty (see Figure 9). These high correlation between GPT4o’s output ratings and the ratings provided by human participants robustly validates our framework’s effectiveness.

C.7.1 Type 1 Question

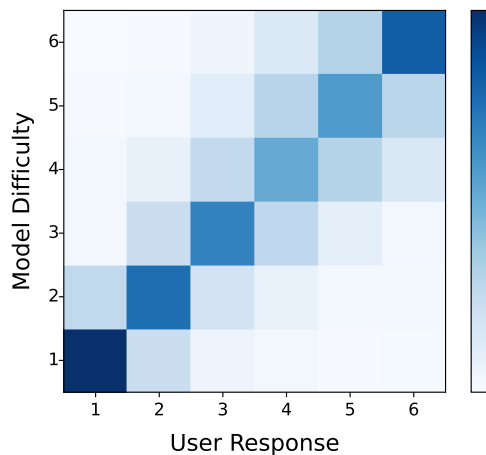


Figure 8: Survey results for type 1 questions by GPT4o.

C.7.2 Type 2 Question

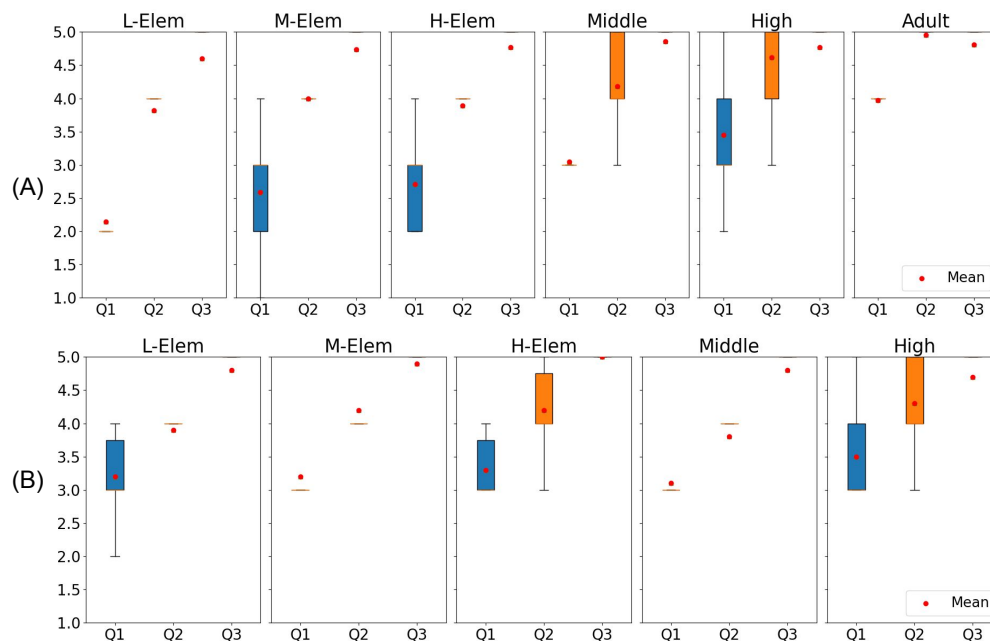


Figure 9: Survey results for type 2 questions by GPT4o.

D Exploring Worldview of Models–Output Distribution

We analyze the different worldviews of models through output analysis across the experimental datasets. Since educational metrics primarily capture sentence and word-level complexity, we present three analyses of these textual aspects.

D.1 Sentence Length Distribution

As shown in Figure 10, sentence length increases for higher-grade models. These models produce more detailed explanations that require advanced reading skills. Lower grade models generate shorter statements that match younger learners’ language needs [54]. Each model’s worldview shifts with its target grade level, from concise simplicity to detailed explanation.

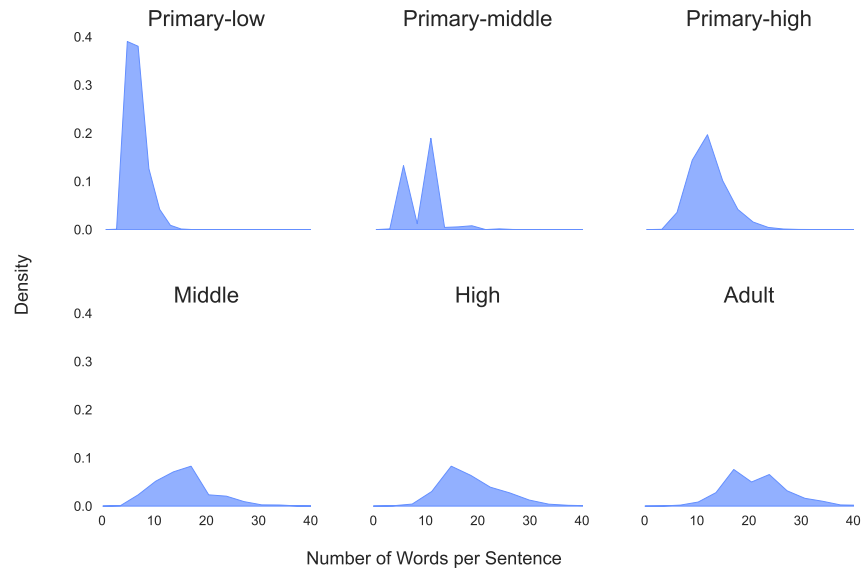


Figure 10: Sentence length distribution across different models.

D.2 Word Difficulty

A key component of model worldview analysis involves examining output vocabulary. Humans use different words as they develop, even when describing identical situations. We visualize outputs through word clouds and Zipfian distributions [55], applying BM25 [56] to highlight distinctive terms while minimizing common words like “people” and “like.” We exclude words appearing fewer than 30 times to eliminate outliers. Figure 11 reveals vocabulary shifts across different models, with higher-grade models using more complex terms that reflect advanced linguistic development.

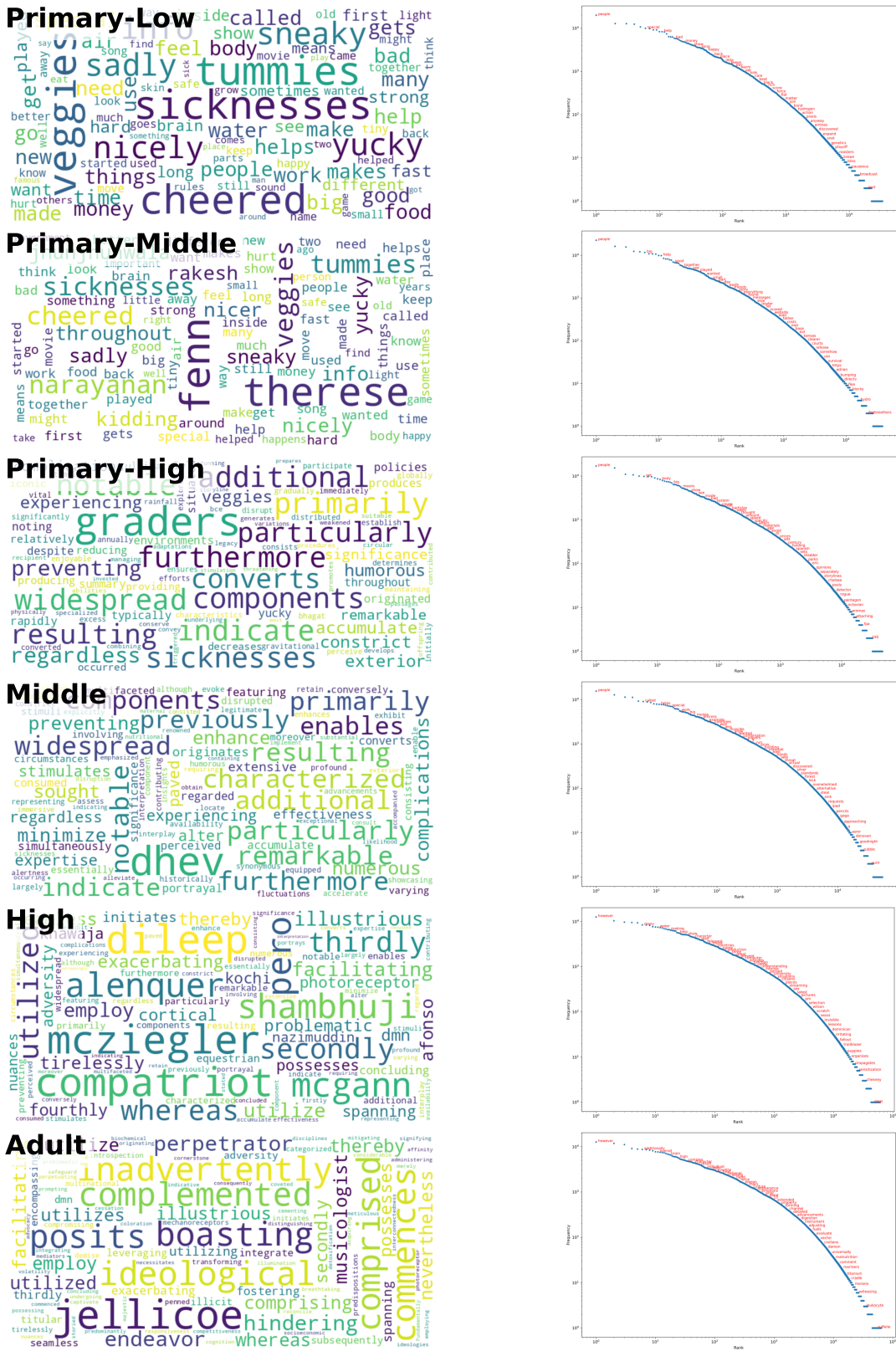


Figure 11: Word Cloud and Zipfian Distribution

E Survey Details

We provide further details on how we distribute and design the surveys.

E.1 Survey Distribution Details

For Survey 1, we sampled questions from \mathcal{D}_{NQ} , consisting of real user questions submitted to Google Search. We assessed whether finetuned model answers to natural questions were both comprehensible for target grades and factually accurate. We provided 5 Type 1 questions and 18 Type 2 questions to each participant with compensation of 7000 Korean Won (approximately 5 USD) for an estimated 30 minutes of work. We recruited 108 English-speaking participants from the Republic of Korea, the USA, and England. To ensure question diversity, we created 18 distinct forms, each completed by six participants.

For Survey 2, we sampled questions from \mathcal{D}_{SQ2} to measure the comprehensibility and accuracy of finetuned models for grade-appropriate questions. \mathcal{D}_{SQ2} contains questions classified for grades 1-12 based on difficulty. We provided 10 Type 2 questions per grade level to 100 participants from Amazon Mechanical Turk. Participants received 1.5 USD for an estimated 8 minutes of work.

E.2 Type 1 Question

- Q: Assign each answer to a unique grade level based on how comprehensible/understandable it would be to students of that grade.

E.3 Type 2 Questions

- Q1: Is the question answerable by X-grade students (regardless of the given answer)?
 - 5: Any X-grade student can answer this question.
 - 4: Most X-grade students can answer this question.
 - 3: It is likely that X-grade students can answer this question.
 - 2: Most X-grade students cannot answer this question.
 - 1: It is impossible for X-grade students to answer this question.
- Q2: Is the answer comprehensible by X-grade students (regardless of the given question)?
 - 5: Any X-grade student can comprehend this answer.
 - 4: Most X-grade students can comprehend this answer.
 - 3: It is likely that X-grade students can comprehend this answer.
 - 2: Most X-grade students cannot comprehend this answer.
 - 1: It is impossible for X-grade students to comprehend this answer.
- Q3: Is the answer consistently answering the question?
 - 5: The answer answers the corresponding question.
 - 4: The answer mostly follows the corresponding question.
 - 3: The answer roughly follows the corresponding question.
 - 2: The answer hardly follows the corresponding question.
 - 1: The answer is completely irrelevant to the corresponding question.