

<https://doi.org/10.1038/s44387-026-00081-7>

Classroom AI: large language models as grade-specific teachers

Check for updates

Jio Oh^{1,2,3}, Steven Euijong Whang¹, James Evans⁴ & Jindong Wang³✉

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.

Large Language Models (LLMs) have expanded beyond traditional natural language processing into diverse domains ranging from finance^{1,2} and healthcare^{3,4} to education⁵⁻⁷. With models like GPT⁸, Gemini⁹, and LLaMA¹⁰ 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 2030¹¹, while 35% of U.S. public schools report at least one teaching vacancy¹². Worldwide, 244 million children lack access to school¹¹, with shortages most acute in rural and high-poverty areas¹³. 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 trained¹⁶.

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 answers¹⁷. 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 corpora¹⁸⁻²¹. 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 paraphrasing²²⁻²⁴. However, real classroom settings involve open-ended questions without source 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.

¹Korea Advanced Institute of Science & Technology (KAIST), Daejeon, South Korea. ²Microsoft Research Asia, Beijing, PR China. ³William & Mary, Williamsburg, VA, USA. ⁴University of Chicago, Chicago, IL, USA. ✉e-mail: jdjw@wm.edu

Moreover, our approach advocates for models that behave appropriately given teaching context²⁷.

To assess text complexity, we integrate seven established readability metrics: Flesch Reading Ease²⁸, Flesch-Kincaid Grade Level²⁹, the Coleman-Liau Index³⁰, Linsear Write³¹, the Gunning Fog Index³², Dale-Chall³³, and the Spache Readability Formula³⁴. 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^{37–39}. Notably, Orca-math⁴⁰ 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 (GPT⁸, Gemini⁹, LLaMA¹⁰, and Mixtral⁴¹) to generate questions answerable across all grade levels (see Supplementary Fig. 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 gain⁴³, 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 Fig. 1A, 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 Sections C.3, C.6.

Figure 1B 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 target⁴⁷.

Accuracy

We test the finetuned models' accuracy on \mathcal{D}_{SQ} , a multiple-choice dataset aligned with educational contexts. Figure 1C shows that finetuned models achieve performance comparable to the base model. While finetuning typically causes some accuracy reduction^{44–46}, our results show minimal performance degradation.

Perplexity and diversity

We measure output unexpectedness through perplexity and diversity gain (see Supplementary Sections B.2, B.3, for formulas). Figure 1D 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 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 1E shows dark cells along the diagonal, indicating strong alignment between model outputs and human perceptions.

For Type 2 questions, Fig. 2 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

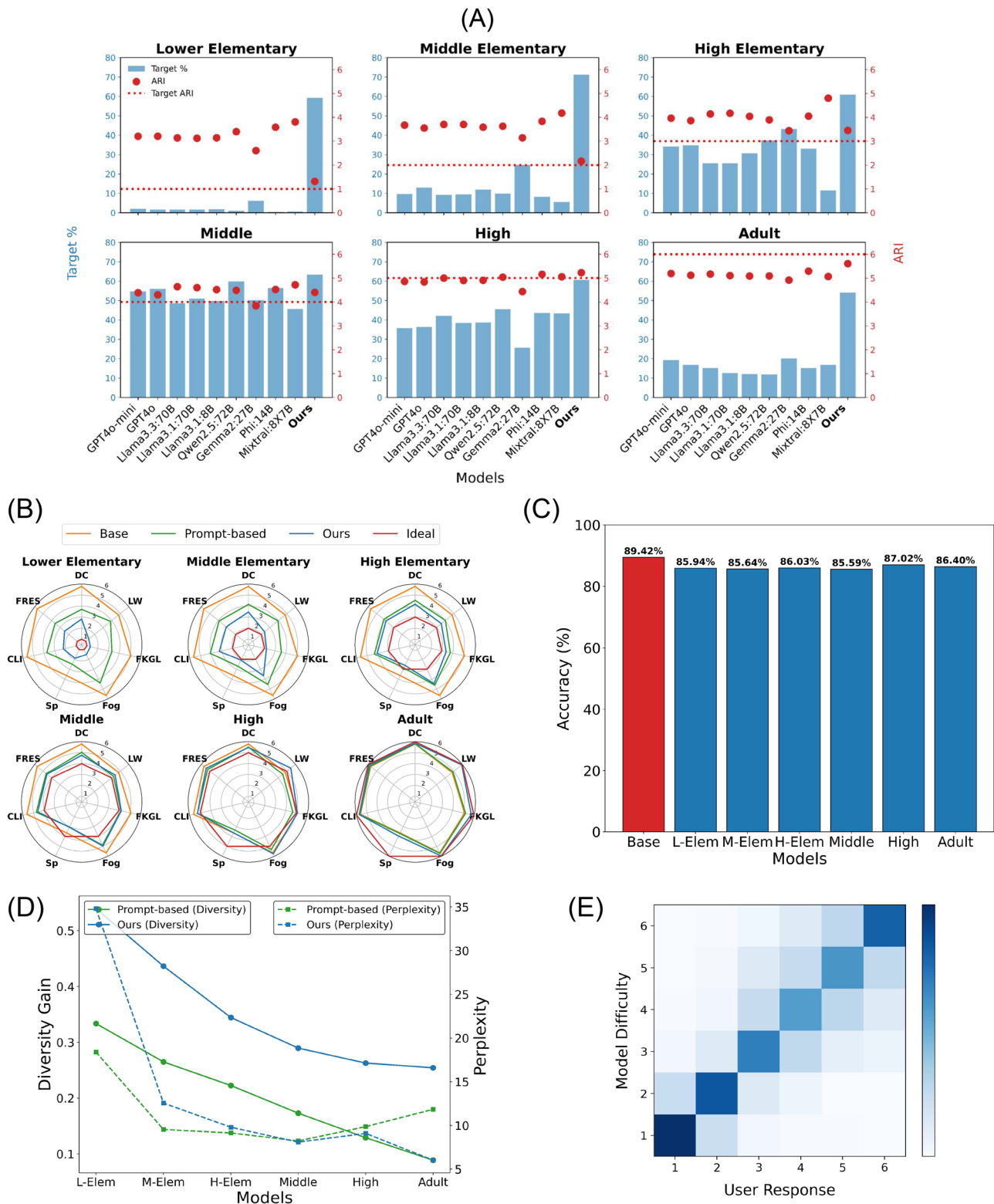


Fig. 1 | Results across evaluation criteria. **A** Compatibility through integrated measure and ARI; **B** compatibility for each metric; **C** accuracy; **D** diversity gain and perplexity; and **E** survey results for type 1 questions.

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 (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).

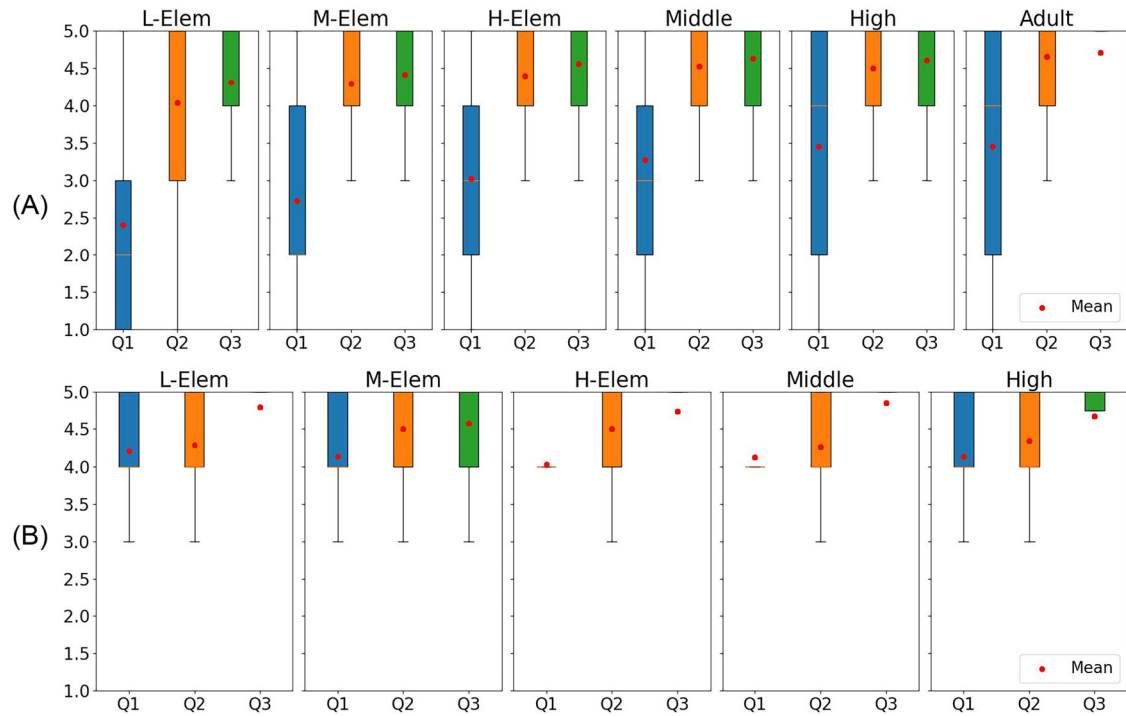


Fig. 2 | Survey results on Type 2 questions. A \mathcal{D}_{NQ} and B \mathcal{D}_{SQ2} . Box plots show Q1 (question difficulty), Q2 (answer comprehensibility), and Q3 (model accuracy) across grade levels, with mean values as red dots on a five-point scale. Higher Q1 and Q2 results indicate lower question 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} .

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 lens⁴⁹ 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 Fig. 3. 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 Fig. 4.

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 Fig. 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 demonstrations⁸, we employ LLMs including Gemini⁹, GPT⁵⁰, and LLaMA¹⁰ to generate ~550 questions per subject.

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

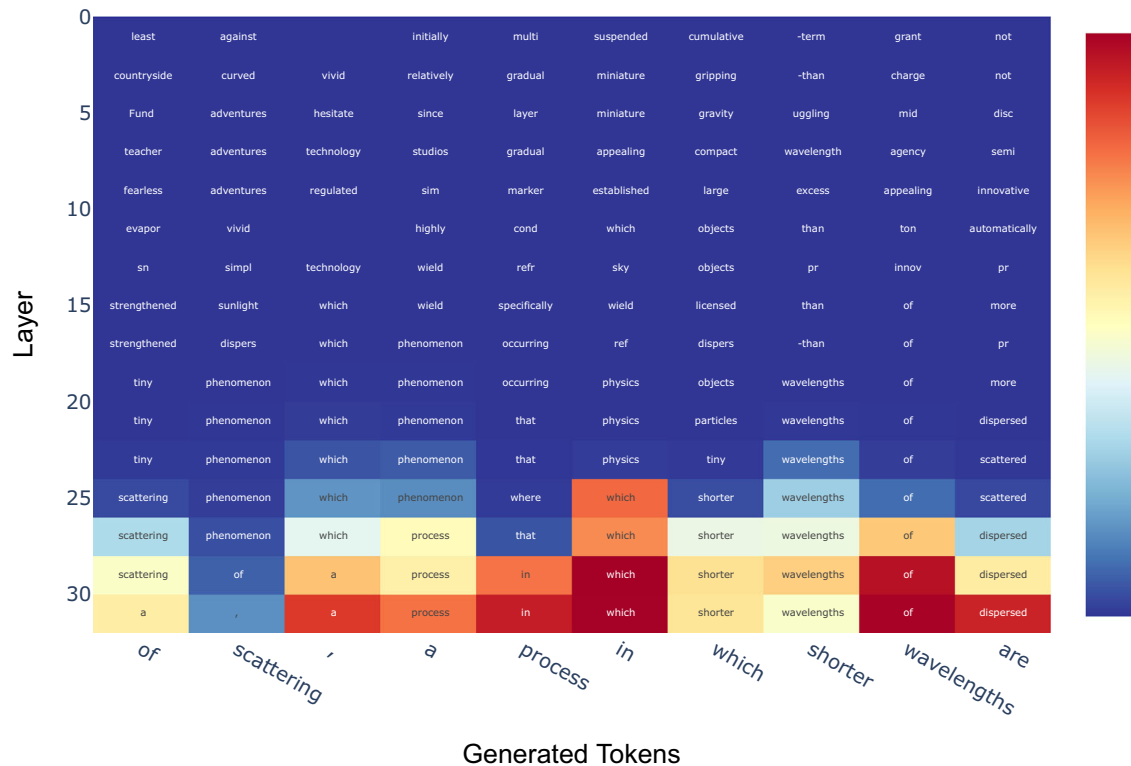
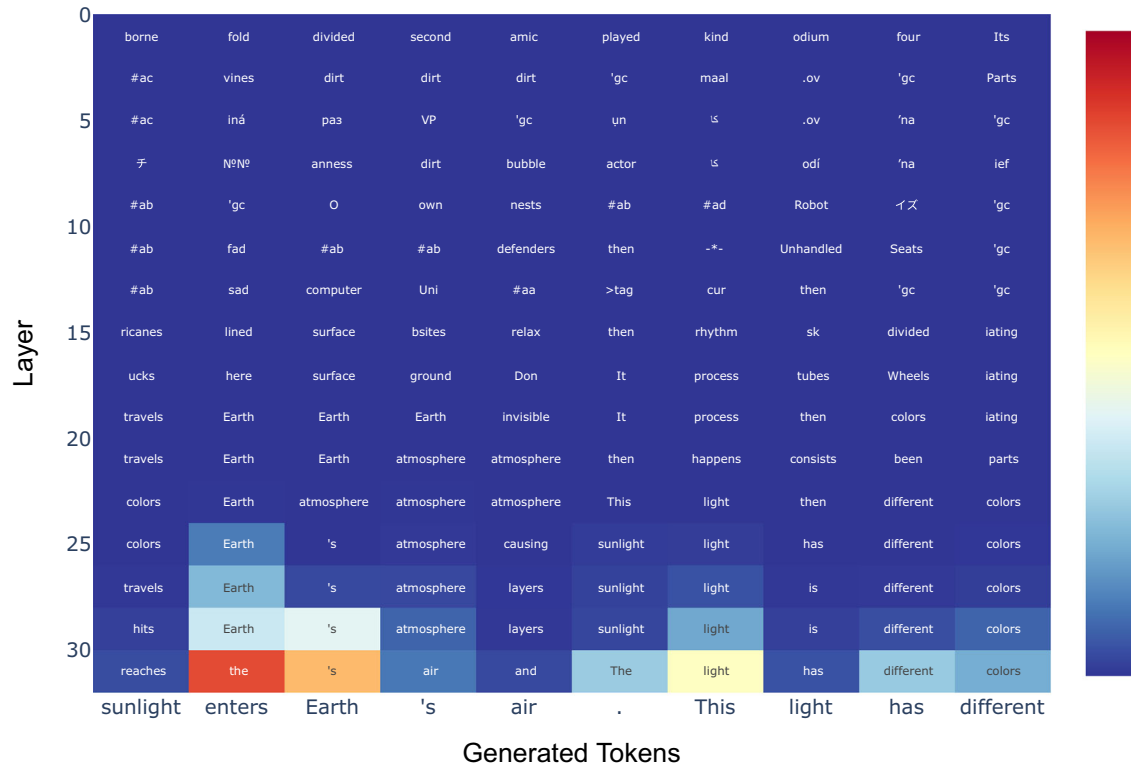


Fig. 3 | Logit-lens visualization for the lower-elementary (top) and adult (bottom) models of LLaMA3.1:8B on the prompt, “Why is the sky blue? The sky is blue because”. The bottom row for each figure shows the final output tokens, and each row above represents the top prediction at each transformer layer. Warmer colors (e.g., red) indicate higher confidence.

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 Fig. 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

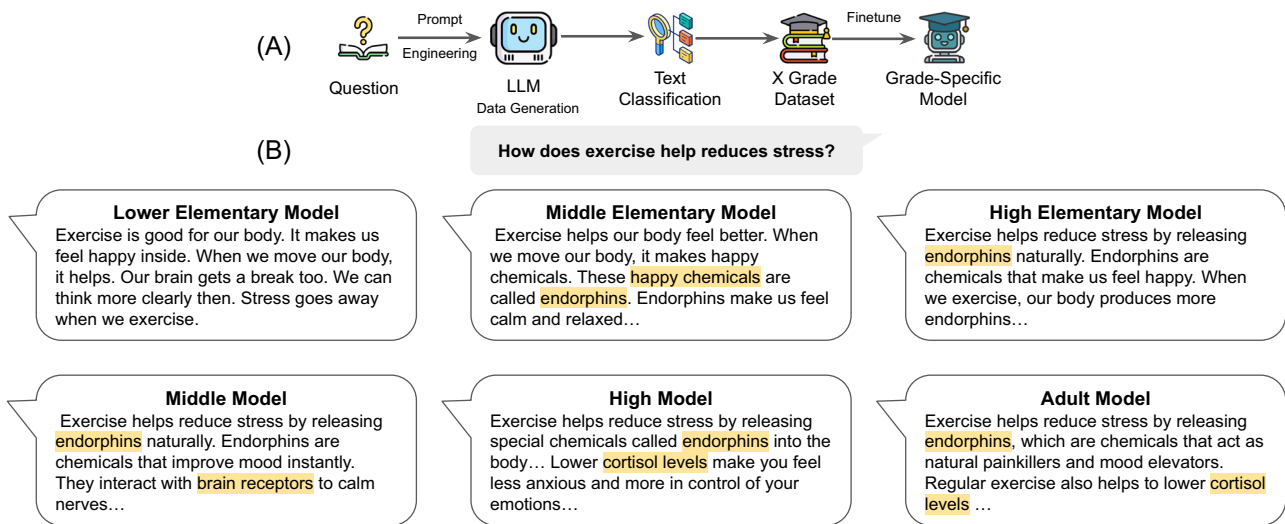


Fig. 4 | Grade-specific LLM finetuning framework and exemplar outputs.

A Overview of our pipeline: we generate open-ended educational questions, prompt an LLM with readability-aligned prompt variants to generate multiple candidate explanations, classify each explanation into one of six grade-levels using our integrated readability metric, and aggregate the labeled question-answer pairs into

grade-specific datasets used to finetune corresponding grade-specific models.

B Example outputs for the question, "How does exercise help reduce stress?" from the six grade-specific models (lower elementary to adult), illustrating progressively more complex wording and sentence structure at higher grade-levels.

Metric²⁸. 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:

- G₁ = Metrics utilizing predefined easy word lists (e.g., DC, Sp),
- G₂ = Metrics focusing on average sentence and word length (e.g., FRES, FKGL, CLI),
- G₃ = 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
            |  $\text{Vote}(G_i) \leftarrow L_{G_i}$ ;
        else
            |  $\text{Vote}(G_i) \leftarrow \min\{L_f \mid f \in G_i\}$ ;

// 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.

Data availability

The data used and generated in this study will be available upon publication at <https://github.com/harryoh99/ClassroomAI>.

Code availability

The Python code supporting the framework will be available upon publication at <https://github.com/harryoh99/ClassroomAI>.

Received: 17 June 2025; Accepted: 11 February 2026;

Published online: 03 March 2026

References

- 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 2023). <https://doi.org/10.1145/3604237.3626869>.
- Lee, J., Stevens, N. & Han, S. C. Large language models in finance (finllms). *Neural. Comput. Appl.* **37**, 24853–24867 (2025).
- Nazi, Z. A. & Peng, W. Large language models in healthcare and medical domain: A review. *Inform.* **11** (2024).
- Tam, TYC. et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ digital med.* **7**, 258 (2024).
- Huang, Y. et al. Social science meets llms: how reliable are large language models in social simulations? Preprint at <https://arxiv.org/abs/2410.23426> (2024).
- Wang, S. et al. Large language models for education: A survey and outlook. *IEEE Signal Processing Magazine* **42**, 51–63 (2026).
- Gan, W., Qi, Z., Wu, J. & Lin, J. C.-W. Large language models in education: Vision and opportunities. In *2023 IEEE international conference on big data (BigData)*, 4776–4785 (IEEE, 2023).
- OpenAI. Gpt-4 technical report. Preprint at <https://arxiv.org/abs/2303.08774> (2024).
- Team, G. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Preprint at <https://arxiv.org/abs/2403.05530> (2024).
- team, L. The llama 3 herd of models. Preprint at <https://arxiv.org/abs/2407.21783> (2024).
- UNESCO. 244m children won't start the new school year. <https://www.unesco.org/gem-report/en/articles/244m-children-wont-start-new-school-year> (2022).
- National Center for Education Statistics. Public school leaders reports suggest no overall change in the number of teaching positions since last school year. https://nces.ed.gov/whatsnew/press_releases/12_12_2024.asp (2024).
- 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).
- 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).
- Khan, M. J. & Ahmed, J. Child education in the time of pandemic: learning loss and dropout. *Children Youth Serv. Rev.* **127**, 106065 (2021).
- World Teachers' Day: The Teachers We Need for the Education We Want: The Global Imperative to Reverse the Teacher Shortage*. SAGE Publishing, <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> (2024).
- Roeein, D., Curry, A. C. & Hovy, D. Know your audience: Do llms adapt to different age and education levels? Preprint at <https://arxiv.org/abs/2312.02065> (2023).
- 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). <https://aclanthology.org/2023.gem-1.18/>.
- Gobara, S., Kamigaito, H. & Watanabe, T. Do llms implicitly determine the suitable text difficulty for users? In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, 940–960 (2024).
- Hsu, Y.-S., Feldhus, N. & Hakimov, S. Free-text rationale generation under readability level control. <https://api.semanticscholar.org/CorpusID:270870139> (2024).
- 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, 2024). <https://doi.org/10.1145/3627508.3638345>.
- 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, 2024). <https://aclanthology.org/2024.findings-acl.926/>.
- Tran, H., Yao, Z., Li, L. & Yu, H. Readctrl: Personalizing text generation with readability-controlled instruction learning. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, 19–36 (2025).
- 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*, <https://openreview.net/forum?id=IFNbElsnCi> (2023).
- Chall, J. S. *Stages of reading development* (McGraw-Hill New York, 1983).
- National Institute of Child Health and Human Development. *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. Preprint at <https://arxiv.org/abs/2412.19010> (2024).
 28. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **32**, 221 (1948).
 29. Kincaid, J. P., Fishburne, R. P., Jr, Rogers, R. L. & Chissom, B. S. *Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel* (US Naval Air Station, 1975).
 30. Coleman, M. & Liau, T. L. A computer readability formula designed for machine scoring. *J. Appl. Psychol.* **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* (McGraw-Hill, 1952).
 33. Chall, J. S., & Dale, E. *Readability Revisited: The New Dale-Chall Readability Formula* (Brookline Books, 1995).
 34. Spache, G. A new readability formula for primary-grade reading materials. *Elementary School J.* **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. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, 10443–10461 (2023).
 36. Long, L. et al. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*. 11065–11082 (2024).
 37. Peng, B., Li, C., He, P., Galley, M. & Gao, J. Instruction tuning with gpt-4. Preprint at <https://arxiv.org/abs/2304.03277> (2023).
 38. Chen, Z., Deng, Y., Yuan, H., Ji, K. & Gu, Q. Self-play fine-tuning converts weak language models to strong language models. In *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235, 6621–6642 (PMLR, 2024).
 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. Preprint at <https://arxiv.org/abs/2402.14830> (2024).
 41. Jiang, A. Q. et al. Mixtral of experts. Preprint at <https://arxiv.org/abs/2401.04088> (2024).
 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. Preprint at <https://arxiv.org/abs/2202.00132> (2022).
 44. Wang, Y. et al. Two-stage llm fine-tuning with less specialization and more generalization. In *International Conference on Learning Representations*. Vol. 2024, 20380–20398 (2024).
 45. Dodge, J. et al. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. Preprint at <https://arxiv.org/abs/2002.06305> (2020).
 46. Luo, Y. et al. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing* (2025).
 47. Zheng, L. et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*. Vol. 36, 46595–46623 (Curran Associates, Inc., 2023).
 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, 2024). <https://aclanthology.org/2024.emnlp-main.896/>.
 49. Belrose, N. et al. Eliciting latent predictions from transformers with the tuned lens. Preprint at <https://arxiv.org/abs/2303.08112> (2023).
 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). https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfbcb4967418bfb8ac142f64a-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. https://huggingface.co/datasets/rexarski/eli5_category (2021).
 53. Kwiatkowski, T. et al. Natural questions: a benchmark for question answering research. In *Transactions of the Association of Computational Linguistics* (MIT Press, 2019).

Acknowledgements

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. Jio Oh conducted this work during internships at Microsoft Research Asia and William & Mary. 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.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s44387-026-00081-7>.

Correspondence and requests for materials should be addressed to Jindong Wang.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026