OMNIEDU: OPEN FOUNDATION MODELS FOR LEARNING AND TEACHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Educational foundation models must do more than produce correct answers: they must understand where a problem sits in a curriculum, diagnose why a learner is struggling, and choose an appropriate instructional response. Existing educational language models often specialize in either subject problem solving or tutoring, while their training mixtures are commonly organized by source or task and do not explicitly balance these capabilities. We present OMNIEDU, an open family of foundation models for K–12 learning and teaching, trained with a capability-oriented instruction-tuning corpus. The corpus combines more than 100 educational resources and general instruction sources and organizes supervision around four complementary capabilities: subject competence, curriculum grounding, diagnostic reasoning, and pedagogical action and scaffolding. A multi-stage pipeline performs deterministic cleaning, semantic auditing and rewriting, task-specific quality scoring, token-budgeted diversity selection, and pedagogical instruction assignment, yielding 69,999 examples and 15.96M supervised response tokens, including 60,951 education-specific examples. We fine-tune 4B, 9B, and 27B models and evaluate them on curriculum-grounding, K–12 problem-solving, and pedagogical-tutoring benchmarks, with auxiliary tests of general capability. Across model scales, education-oriented tuning consistently improves all three educational capability groups. In particular, OMNIEDU-27B reaches 63.12% EM / 76.69% F1 on K12-Bench, 85.89% on MathFish, 86.95% on EDUMATH, 78.74% on MathTutorBench's Scaffold setting, and the best Teaching average of 3.02 on LongTutor among the evaluated models. These results show that carefully curated, capability-balanced supervision can turn a general language model into a stronger educational system that not only solves problems, but also understands curriculum structure and supports effective teaching interactions.

1 INTRODUCTION

Large language models are increasingly used in education, but educational usefulness extends beyond answer accuracy. A capable learning and teaching assistant must solve a student's problem, connect it to the relevant knowledge point and prerequisite structure, diagnose misconceptions, and choose an intervention that advances learning. It must therefore coordinate *what to teach*, *where the learner is*, and *how to respond*, rather than treating education as ordinary question answering.

Recent educational models have advanced individual aspects of this problem, including examination-style problem solving, tutoring dialogue, Socratic questioning, and curriculum alignment (Dan et al., 2023; MuduoLLM Contributors from BNU and TAL, 2025; Wu et al., 2025b; LearnLM Team et al., 2024). However, these capabilities are often developed or evaluated separately. A model may answer correctly while failing to identify the relevant curriculum concept, explain an error, or provide an appropriate scaffold. The field thus lacks an open model family explicitly organized around the full learning–teaching loop.

We argue that this gap is partly a data-design problem. Educational instruction data are heterogeneous: short answers, curriculum relations, misconception diagnoses, and tutoring exchanges supervise different behaviors. Mixing them only by source or subject can obscure the intended response policy and overrepresent redundant examples. The challenge is therefore to construct a training

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mixture in which each example has a clear capability target and pedagogically appropriate response behavior.

To address this, we present OMNIEDU, an open family of foundation models for K–12 learning and teaching, organized around four complementary capabilities. *Subject competence* covers K–12 problem solving and explanation; *curriculum grounding* covers grade level, knowledge points, prerequisites, difficulty, and curriculum localization; *diagnostic reasoning* identifies errors, misconceptions, and missing prerequisites from learner work; and *pedagogical action and scaffolding* selects and executes interventions such as questions, hints, prerequisite review, or direct explanation. This taxonomy provides a common structure for data construction and evaluation.

Starting from more than 100 datasets and educational resources, our data-centric pipeline combines deterministic cleaning and decontamination, LLM-assisted semantic auditing, task-specific quality scoring, diversity-aware selection, and pedagogical instruction assignment. It reduces an initial education-specific pool of approximately 1.34M examples to 60,951 high-quality examples with about 12.0M supervised response tokens. We then add 9,048 general-purpose instruction examples and assign one of 19 task-specific pedagogical system instructions to each education-specific example, yielding 69,999 examples and 15.96M supervised response tokens with provenance and audit metadata retained throughout.

We fine-tune 4B, 9B, and 27B models and evaluate curriculum grounding, K–12 problem solving, and pedagogical tutoring on K12-Bench, MathFish, EDUMATH, GAOKAO-Bench, EXAMS-V, MDK12-Bench, MathTutorBench, TutorBench, and LongTutor, together with auxiliary general-capability evaluations. Education-oriented tuning improves all three educational dimensions across model scales. OMNIEDU-27B reaches 63.12% EM / 76.69% F1 on K12-Bench, 85.89% on MathFish, 86.95% on EDUMATH, 78.74% Scaffold win rate on MathTutorBench, and a Teaching average of 3.02 on LongTutor. These gains extend beyond answer correctness to curriculum reasoning and pedagogical interaction.

Our contributions are:

- We introduce OMNIEDU, an open family of K–12 foundation models organized around four capabilities that connect subject knowledge, curriculum structure, learner diagnosis, and pedagogical action.
- We develop a reproducible capability-oriented data construction pipeline combining semantic quality control, task-specific filtering, token-budgeted diversity selection, and explicit pedagogical instructions, producing a 69,999-example training mixture with 15.96M supervised response tokens.
- We provide a cross-capability evaluation showing consistent gains across curriculum grounding, K–12 problem solving, and pedagogical tutoring at 4B, 9B, and 27B scales, together with auxiliary tests of general model capability.
- We provide public project and dataset entry points for reproducibility through the project repository and the Hugging Face dataset page.

2 RELATED WORK

2.1 EDUCATIONAL FOUNDATION MODELS AND AI TUTORS

Recent work increasingly adapts general-purpose LLMs to educational scenarios. EduChat (Dan et al., 2023) supports broad learning interactions such as question answering, assessment, and Socratic teaching, while MuduoLLM (MuduoLLM Contributors from BNU and TAL, 2025) targets the Chinese K–12 curriculum and Confucius3-Math (Wu et al., 2025b) specializes in mathematical reasoning. Frontier systems also incorporate explicit pedagogical behaviors: LearnLM (LearnLM Team et al., 2024) formulates educational adaptation as *pedagogical instruction following*, while ChatGPT Study Mode (OpenAI, 2025), Khanmigo (Khan Academy, 2023), and recent teacher-facing systems (Anthropic, 2026) emphasize guided reasoning, scaffolding, or curriculum-aware workflows. These efforts demonstrate the value of educational adaptation, but subject knowledge, curriculum understanding, diagnosis, and pedagogy are often developed separately or embedded in closed systems. OmniEdu instead jointly targets these capabilities in an open model family.

2.2 Educational Data and Data-Centric Post-Training

Instruction-tuning research shows that performance depends strongly on training-data quality and composition. Recent data-centric frameworks and benchmarks further extend this view to reproducible data preparation, quality evaluation, and dynamic data optimization (Liang et al., 2025; 2026c;a; He et al., 2026). LIMA (Zhou et al., 2023a) demonstrates the effectiveness of carefully curated supervision, while AlpaGasus (Chen et al., 2023), Deita (Liu et al., 2024b), and LESS (Xia et al., 2024) study LLM-based filtering, quality and diversity, and targeted data selection. Educational data introduce additional curriculum structure. Human Curriculum (Lee et al., 2024) organizes instructions by subject hierarchy and difficulty, K12-KGraph (Liang et al., 2026b) derives curriculum-structured supervision from textbook knowledge graphs, and EDUMATH (Christ et al., 2025) constructs standards-aligned math-generation data. OmniEdu extends these data-centric ideas to a unified mixture organized by educational capability and pedagogical behavior.

2.3 Evaluation of Educational LLMs

Educational evaluation has expanded from final-answer accuracy toward curriculum and pedagogical competence. Learner-oriented benchmarks assess school-level knowledge and reasoning across text, multilingual, and multimodal settings (Lu et al., 2022; Zhang et al., 2023b; Hou et al., 2024; Zhang et al., 2023a; He et al., 2024; Li et al., 2025b; Das et al., 2024; Zhou et al., 2025; Ye et al., 2025; Lei et al., 2024; Li et al., 2025a). Curriculum-oriented benchmarks evaluate standards alignment, grade level, knowledge points, prerequisites, and curriculum-conditioned generation (Li et al., 2024; Liang et al., 2026b; Zhang et al., 2024b; Naeem et al., 2025; Wu et al., 2025a; Christ et al., 2025), while pedagogy-oriented benchmarks cover explanation, scaffolding, feedback, diagnosis, adaptive tutoring, multimodal interaction, and long-term student modeling (Macina et al., 2025; Srinivasa et al., 2025; Li et al., 2026; Maurya et al., 2025; Yang et al., 2026; Shi et al., 2026; Hazra et al., 2026). Our evaluation follows this broader learning–teaching view.

3 Data Preparation

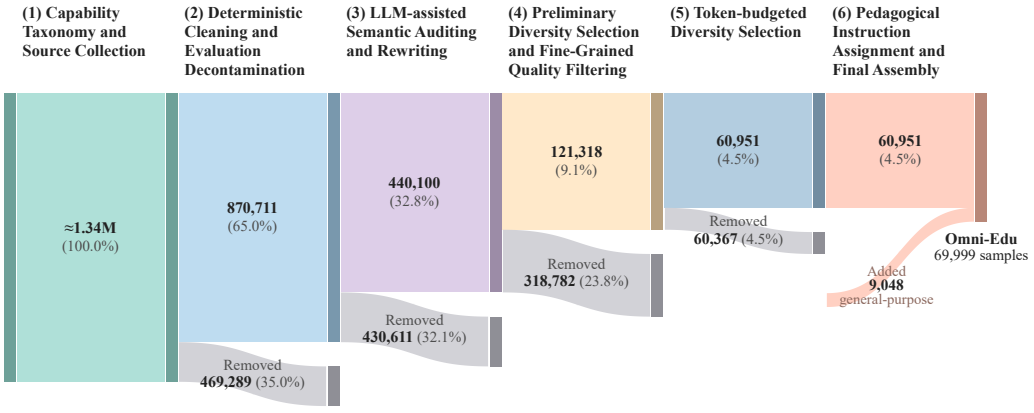


Figure 1: Overview of the OmniEdu data preparation pipeline. The education-specific pool is progressively curated from approximately 1.34M candidates to 60,951 high-quality examples through six stages. Gray branches indicate filtered-out examples.

We construct a large-scale instruction-tuning corpus for K–12 education by integrating heterogeneous sources and curating them according to the capabilities required of an educational language model. Our pipeline consists of six stages: (1) capability taxonomy and source collection, (2) deterministic cleaning and evaluation decontamination, (3) LLM-assisted semantic auditing and rewriting, (4) preliminary diversity selection and fine-grained quality filtering, (5) token-budgeted diversity selection, and (6) pedagogical instruction assignment and final assembly. Example-level provenance and audit metadata are retained throughout. Starting from more than 100 datasets and

educational resources, the education-specific pool contains approximately 1.34M examples before filtering.

3.1 Capability Taxonomy and Source Collection

Our corpus contains two complementary components: **general-purpose instruction data** and **education-specific data**. The former contains 9,048 examples from DataFlow-Instruct-10K (Liang et al., 2025) (7,431), the Tulu-3-SFT-mixture (Lambert et al., 2024) (1,495), and MathV360K (Shi et al., 2024) (122), preserving general instruction following, multilingual coverage, refusal behaviour, and diagram-based reasoning (see Appendix A.1 for details). The latter targets the knowledge, reasoning, and pedagogical capabilities required in K–12 education.

Rather than organizing education-specific data solely by subject or source, we define four capability categories according to the behaviors to be learned. *Subject competence* covers solving K–12 problems and producing correct answers and explanations across mathematics, science, reading comprehension, and writing. *Curriculum grounding* captures grade-level alignment, knowledge-point identification, prerequisite relationships, difficulty, and curriculum placement. *Diagnostic reasoning* focuses on identifying errors and misconceptions in student solutions and inferring missing prerequisite knowledge. *Pedagogical action and scaffolding* covers selecting and executing an appropriate instructional action, such as asking a question, providing a hint, revisiting a prerequisite, or giving a direct explanation.

Each example is assigned exactly one primary capability according to its main supervision objective, while secondary properties are retained as metadata. Within each capability, we further partition examples into **fine-grained task buckets** by factors such as task form, modality, and supervision type. A complete source inventory, licensing information, and source-to-category mapping are provided in Appendix A.1.

3.2 Deterministic Cleaning and Evaluation Decontamination

We first standardize heterogeneous sources into a unified representation, remove exact duplicates, and discard examples with missing or malformed required fields. Where answers can be deterministically verified, we check their consistency with the corresponding references or structured annotations. For mixed-domain sources, we retain only K–12-relevant examples and remove unrelated content such as finance or general encyclopedic knowledge. For multimodal examples, we additionally verify image availability and alignment with textual inputs. Source and license information is retained for provenance tracking, and data with unclear usage conditions are excluded.

To prevent evaluation contamination, we remove any example overlapping with the final evaluation set. After this stage, 870,711 education-specific examples remain. Detailed cleaning and decontamination rules are provided in Appendix A.2.1.

3.3 LLM-assisted Semantic Auditing and Rewriting

We use Qwen3.5-122B-A10B-FP8 (Qwen Team, 2026a) for semantic quality control that cannot be reliably handled by deterministic rules. For each example, the auditor produces task-specific checklist judgments together with an overall usability score, which are then mapped by predefined decision rules to one of three actions: *keep*, *rewrite*, or *remove*. The detailed auditing criteria and decision rules are provided in Appendix A.2.2.

The criteria depend on the supervision target. Subject-competence data are checked for answer correctness, answer–explanation consistency, and grounding in the input; curriculum examples for valid curriculum relations and sufficiently specific knowledge-point alignment; diagnostic data for evidence-supported error identification and consistent corrective responses; and pedagogical data for instructional relevance, coherence, scaffolding quality, and premature answer leakage.

For *rewrite* cases, we keep the original input fixed and repair only the defective supervision. The rewritten example is audited again using the same criteria and retained only if classified as *keep*. Full auditing prompts and criteria are retained with the release metadata. This stage leaves 440,100 education-specific examples.

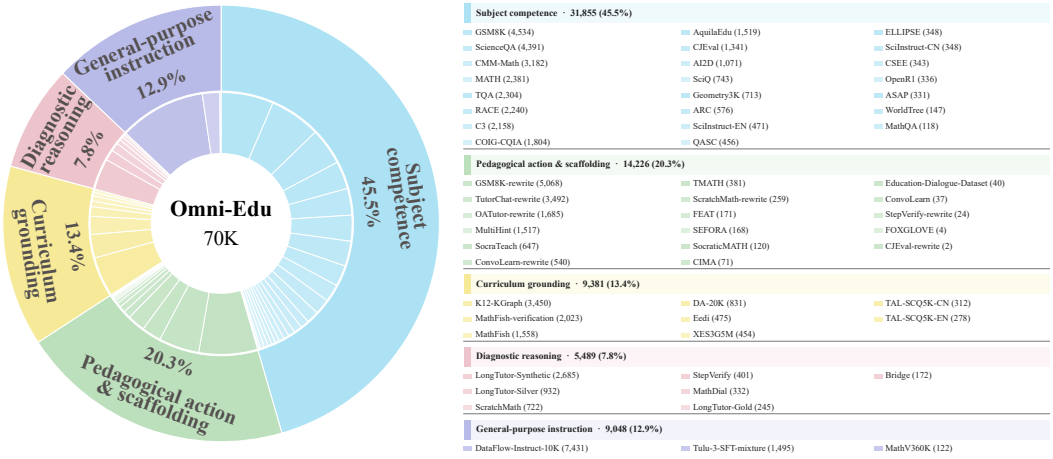


Figure 2: Composition of the final OMNIEDU (69,999 examples / 15.96M supervised response tokens). Left: Data Distribution within Each Category. The outer circle shows the distribution of the four education-specific capability categories and the general-purpose instruction data, and the inner circle shows the distribution of source subsets. Right: The detailed quantities of datasets.

3.4 PRELIMINARY DIVERSITY SELECTION AND FINE-GRAINED QUALITY FILTERING

Before applying the more expensive quality scorer, we reduce a small number of highly overrepresented sources whose candidate sizes substantially exceed their intended contributions. We use k-center greedy over BGE-M3 embeddings (Multi-Granularity, 2024) to select semantically diverse subsets rather than random samples. For example, RACE is reduced from 60,180 to 5,000 examples, AquilaEdu from 23,226 to 2,000, and CJEval from 17,178 to 2,000; sources without substantial redundancy are retained without this preliminary compression.

We then use GPT-5.6-Terra for fine-grained quality scoring. The rubrics refine the criteria in Section 3.3 into task-specific dimensions scored on a 1–5 scale. For example, problem-solving data are evaluated for problem validity, answer correctness, reasoning correctness, completeness, relevance, and clarity (see Appendix A.2.3 for details). An example is retained only if all applicable dimensions score at least 3, with a threshold of 4 for task-critical dimensions such as correctness, validity, and grounding. After this stage, 121,318 education-specific examples remain.

3.5 TOKEN-BUDGETED DIVERSITY SELECTION

The remaining candidates are high-quality but still unevenly distributed across tasks and sources. We therefore perform final selection independently within the fine-grained task buckets, again using k-center greedy over BGE-M3 embeddings to maximize semantic coverage.

Because response lengths vary substantially across tasks, we allocate each bucket a budget in supervised response tokens rather than example count. Examples are selected greedily within each bucket until its token budget is reached (see Appendix A.2.4 for details). This reduces the education-specific pool to **60,951 examples** containing approximately **12.0M supervised response tokens**.

3.6 PEDAGOGICAL INSTRUCTION ASSIGNMENT AND FINAL ASSEMBLY

Different educational tasks require different response behaviors. A problem-solving response should generally provide and justify the final answer, whereas a scaffolding response may deliberately withhold it and guide the student through intermediate reasoning. A single generic system instruction would therefore conflate potentially conflicting behaviors.

Following *pedagogical instruction following* in LearnLM (LearnLM Team et al., 2024), we associate each selected example with a task-specific system instruction determined by its predefined

task bucket. We use 19 templates covering behaviors such as reasoned problem solving, reading comprehension, diagnosis and correction, Socratic questioning, and tutoring dialogue; the complete templates are provided in Appendix A.3. Each example receives exactly one system instruction, while its original user and assistant content remains unchanged.

Finally, all examples are converted to a unified message-based format, referenced images are materialized and aligned with their placeholders, and we check message structure, media availability, duplicate identifiers, and cross-category duplicates.

3.7 Final Data Statistics

Figure 2 summarizes the final mixture. The education-specific component contains 60,951 examples and approximately 12.0M supervised response tokens; together with 9,048 general-purpose examples, the complete corpus contains **69,999 examples and 15.96M supervised response tokens**.

4 Experiments

We evaluate whether education-oriented instruction tuning improves three capabilities required by a practical K–12 educational assistant: (1) *curriculum grounding*, measuring knowledge points and their structural relations within a curriculum; (2) *K–12 problem solving*, covering authentic school-level problems across subjects and modalities; and (3) *pedagogical tutoring*, assessing explanations, feedback, and instructional guidance beyond answer correctness. We additionally evaluate general-purpose benchmarks to examine whether educational specialization degrades broader capabilities.

4.1 Experimental Setup

Models. We fine-tune three backbone models, Qwen3.5-4B-Base (Qwen Team, 2026a), Qwen3.5-9B-Base (Qwen Team, 2026a), and Qwen3.8-27B (Qwen Team, 2026b), using the corpus in Section 3, obtaining OmniEdu-4B, OmniEdu-9B, and OmniEdu-27B. Unless otherwise specified, all models use the same prompting and decoding protocol. Training uses Llama-Factory (Zheng et al., 2024) with a learning rate of 5×10^{-6}, a maximum sequence length of 32,768, and 3 epochs. Additional training details are provided in Appendix A.4.

Baselines. We compare against two groups of baselines. First, we include **open-weight educational LLMs**, including Confucius3-Math, MuduoLLM, and multiple variants of EduChat. These models cover different education-oriented training paradigms, including domain-specific pretraining, supervised fine-tuning, preference optimization, and reinforcement learning. Second, we compare against **frontier proprietary models**, including GPT-5.4, GPT-5.6-Sol, Claude-Opus-5, GLM-5.3, and Kimi-K3. We additionally compare each fine-tuned model with its corresponding base model to isolate the effect of our education-oriented training. The five open-weight educational LLMs and GLM-5.3 do not support image inputs, they therefore receive only the textual component for image-dependent examples. Since EXAMS-V items encode essential information in the image, such models may degenerate to constant or unparseable predictions on this benchmark.

Benchmarks. For *curriculum grounding*, we evaluate on K12-Bench (Liang et al., 2026b), MathFish (Li et al., 2024), and EDUMATH (Christ et al., 2025). K12-Bench evaluates curriculum cognition through graph-derived tasks involving concept grounding, prerequisite relations, neighborhood relations, supporting evidence, and curriculum localization. MathFish evaluates whether a model can determine the alignment between a mathematical problem and fine-grained curriculum standards. EDUMATH further examines curriculum-grounded generation by requiring models to generate high-quality mathematical word problems conditioned on educational standards and student interests. For *K–12 problem solving*, we primarily report results on GAOKAO-Bench (Zhang et al., 2023b), EXAMS-V (Das et al., 2024), and MDK12-Bench (Zhou et al., 2025). These benchmarks cover authentic K–12 examination problems, including both open-ended and objective questions as well as multimodal inputs. For *pedagogical capability*, we evaluate on MathTutorBench (Macina et al., 2025), TutorBench (Srinivasa et al., 2025), and LongTutor (Li et al., 2026). Together, they cover scaffolded instruction, pedagogical response quality, adaptive explanation, feedback, active-learning guidance, and long-term personalized tutoring. We follow the official evaluation protocol

Table 1: Curriculum-grounding results across all evaluated models. All values are percentages. Top-three values are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	K12-Bench EM	K12-Bench F1	MathFish Acc.	EDUMATH MaC
Open-weight Models					
Qwen3.5-4B-Base	4B	42.72%	69.20%	80.33%	47.60%
OmniEdu-4B (ours)	4B	54.25%	71.75%	83.19%	68.40%
Qwen3.5-9B-Base	9B	48.52%	71.99%	79.66%	58.60%
OmniEdu-9B (ours)	9B	**55.46%**	**73.68%**	83.70%	74.00%
Qwen3.8-27B	27B	52.11%	73.48%	83.54%	70.60%
OmniEdu-27B (ours)	27B	**63.12%**	**76.69%**	**85.89%**	**86.95%**
Confucius3-Math	14B	5.23%	19.23%	0.00%	3.00%
MuduoLLM	14B	48.24%	70.46%	82.85%	57.20%
EduChat-SFT-Qwen2.5-7B	7B	45.20%	67.58%	78.23%	28.00%
EduChat-R1-Qwen3-8B	8B	25.25%	49.67%	19.49%	47.00%
EduChat-R1-Qwen3-32B	32B	49.42%	71.21%	83.71%	61.20%
Proprietary Models					
GPT-5.4	–	43.23%	67.80%	80.65%	80.50%
GPT-5.6-Sol	–	48.63%	71.03%	83.68%	**84.92%**
Claude-Opus-5	–	48.48%	71.34%	83.52%	75.00%
GLM-5.3	–	43.38%	57.66%	**85.23%**	62.78%
Kimi-K3	–	**55.03%**	**74.71%**	**85.40%**	**90.00%**

and metric of each benchmark. MathTutorBench evaluates the pedagogical quality of mathematical tutoring, focusing on scaffolding and pedagogical instruction following. TutorBench evaluates tutoring quality across adaptive explanation, student feedback, and active-learning hints. LongTutor evaluates long-term personalized tutoring with an emphasis on history-grounded diagnosis and adaptive teaching. Finally, we use IFEval (Zhou et al., 2023b), GPQA (Rein et al., 2023), and MMMU-Pro (Yue et al., 2024) as auxiliary diagnostics of how education-oriented fine-tuning affects general instruction following, scientific reasoning, and multimodal reasoning. Detailed inference, judging, multimodal, and answer-parsing protocols are provided in Appendix A.5.

4.2 Curriculum Grounding

K12-Bench. As summarized in Table 1, instruction tuning substantially improves curriculum understanding across all model scales. OmniEdu-27B increases the overall EM from 52.11% to **63.12%** and F1 from 73.48% to **76.69%**, achieving the best overall performance among the evaluated models. The gains are consistent across all K12-Bench subtasks, showing improvements in both knowledge grounding and structural curriculum reasoning.

MathFish and EDUMATH. On MathFish, OmniEdu-27B improves overall accuracy from 83.54% to **85.89%**, achieving the best overall performance, with a large gain on positive alignment cases. On EDUMATH, the MaC scores of the 4B, 9B, and 27B models improve by 20.80, 15.40, and 16.35 points, respectively, with OmniEdu-27B reaching **86.95%**, second only to Kimi-K3. Together, these results demonstrate consistent gains in both directions of curriculum alignment: mapping problems to their underlying standards and generating problems from specified standards.

Overall findings. Across all three curriculum benchmarks, education-oriented tuning consistently improves performance at every model scale. The 27B model consistently ranks among the strongest models, while the 4B and 9B variants also show substantial and consistent gains over their corresponding base models. This cross-benchmark consistency indicates that the improvements in curriculum competence generalize across different task formulations and model scales.

4.3 K–12 Problem Solving

Table 2 reports the headline results on GAOKAO-Bench, EXAMS-V, and MDK12-Bench; complete breakdowns are deferred to Appendix B. On GAOKAO-Bench, OmniEdu-27B improves the

Table 2: K–12 problem-solving results across all evaluated models. All values are percentages. Top-three values are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	GAOKAO-Bench Full	EXAMS-V Overall	MDK12-Bench Full
Open-weight Models				
Qwen3.5-4B-Base	4B	88.89%	44.69%	35.43%
OmniEdu-4B (ours)	4B	90.45%	57.62%	46.36%
Qwen3.5-9B-Base	9B	92.95%	63.00%	44.50%
OmniEdu-9B (ours)	9B	93.67%	66.40%	50.80%
Qwen3.8-27B	27B	91.55%	**68.65%**	46.04%
OmniEdu-27B (ours)	27B	**94.87%**	**69.52%**	**57.76%**
Confucius3-Math	14B	84.51%	21.95%	39.13%
MuduoLLM	14B	89.66%	21.95%	45.44%
EduChat-SFT-Qwen2.5-7B	7B	70.74%	21.95%	32.50%
EduChat-R1-Qwen3-8B	8B	76.06%	0.69%	41.29%
EduChat-R1-Qwen3-32B	32B	83.87%	21.95%	41.84%
Proprietary Models				
GPT-5.4	–	93.44%	35.66%	54.77%
GPT-5.6-Sol	–	**97.44%**	24.82%	54.67%
Claude-Opus-5	–	**98.12%**	64.34%	**57.46%**
GLM-5.3	–	94.87%	22.29%	55.33%
Kimi-K3	–	94.87%	**87.29%**	**63.60%**

Table 3: Pedagogical-tutoring results across all evaluated models. 'MathTutor-S' and 'MathTutor-SH' denote Scaffold and Scaffold-hard win rates; 'LongTutor-E' and 'LongTutor-T' denote Evidence and Teaching averages. Win rates and Evidence are percentages; Teaching is on the benchmark's original scale. Top-three values are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	MathTutor-S	MathTutor-SH	TutorBench	LongTutor-E	LongTutor-T
Open-weight Models						
Qwen3.5-4B-Base	4B	20.42%	18.36%	45.52%	25.67%	1.60
OmniEdu-4B (ours)	4B	75.79%	**84.77%**	46.67%	65.88%	2.29
Qwen3.5-9B-Base	9B	14.00%	13.67%	45.38%	5.81%	1.48
OmniEdu-9B (ours)	9B	75.26%	81.64%	48.16%	66.63%	2.66
Qwen3.8-27B	27B	57.16%	55.86%	**58.58%**	36.80%	**2.74**
OmniEdu-27B (ours)	27B	**78.74%**	**83.59%**	**59.42%**	**78.20%**	**3.02**
Confucius3-Math	14B	27.68%	20.70%	34.72%	23.07%	1.36
MuduoLLM	14B	30.42%	23.44%	35.98%	24.31%	1.28
EduChat-SFT-Qwen2.5-7B	7B	18.21%	22.66%	21.48%	26.55%	1.21
EduChat-R1-Qwen3-8B	8B	11.58%	7.42%	23.71%	33.75%	1.14
EduChat-R1-Qwen3-32B	32B	22.21%	16.80%	25.92%	52.03%	1.56
Proprietary Models						
GPT-5.4	–	6.32%	1.96%	46.34%	75.50%	1.62
GPT-5.6-Sol	–	10.53%	17.65%	50.97%	**77.50%**	1.89
Claude-Opus-5	–	**87.89%**	**84.31%**	54.01%	**84.17%**	**2.72**
GLM-5.3	–	**77.89%**	76.92%	38.62%	77.42%	2.66
Kimi-K3	–	20.00%	15.38%	**63.65%**	74.73%	2.17

full-score rate from 91.55% to **94.87%**, with gains on both objective and subjective questions. On EXAMS-V, the improvement is particularly pronounced for OmniEdu-4B, whose overall accuracy increases from 44.69% to 57.62%. On MDK12-Bench, OmniEdu-27B improves the overall score from 46.04% to **57.76%**, with open-question correctness rising substantially from 62.08% to **78.77%**.

Overall findings. Across all three K–12 problem-solving benchmarks, each tuned model outperforms its corresponding base model on the overall metric. OmniEdu-27B consistently outperforms the evaluated open-weight educational baselines and ranks within the top three overall on all three benchmarks. Although frontier proprietary models retain an advantage on some exam-style evaluations, OmniEdu-27B remains highly competitive despite its substantially smaller scale.

4.4 PEDAGOGICAL TUTORING

MathTutorBench. As shown in Table 3, education-oriented tuning produces particularly large improvements in mathematical tutoring. For OmniEdu-27B, Scaffold and Scaffold-hard win rates increase from 57.16% and 55.86% to **78.74%** and **83.59%**, respectively. The effect is even more pronounced for the smaller models, with Scaffold win rate improving by 55.37 points for OmniEdu-4B and 61.26 points for OmniEdu-9B.

TutorBench. All three tuned models improve in overall performance, with OmniEdu-9B showing the largest gain among the smaller models from 45.38% to 48.16%. OmniEdu-27B achieves the strongest overall score of **59.42%**, ranking second behind Kimi-K3, while obtaining the best aggregate performance on text-only examples.

LongTutor. LongTutor shows particularly strong gains in the use of student learning history. The Evidence average increases from 25.67% to 65.88% for OmniEdu-4B, from 5.81% to 66.63% for OmniEdu-9B, and from 36.80% to **78.20%** for OmniEdu-27B. While Claude-Opus-5 achieves a higher Evidence score, OmniEdu-27B obtains the best Teaching average of **3.02**. OmniEdu also performs particularly well on knowledge-state diagnosis, with OmniEdu-27B and OmniEdu-9B ranking first and second in diagnosis accuracy at **54.04%** and **53.55%**, respectively. However, the relatively low accuracies indicate that knowledge-state diagnosis remains comparatively challenging.

Overall findings. Across all three tutoring benchmarks, education-oriented tuning improves pedagogical performance at every model scale. Taken together, these results demonstrate that the benefit of our corpus extends beyond answer correctness to broader tutoring capabilities. The largest gains appear in scaffolded instruction and the use of longitudinal student evidence.

4.5 GENERAL CAPABILITY RETENTION

In addition to the above education domain benchmarks, we verify that specializing the models for K–12 education does not come at the expense of their general capabilities. We evaluate the base and tuned models on IFEval, GPQA, and MMMU-Pro, covering instruction following, challenging knowledge-intensive reasoning, and multimodal general understanding. Complete results are reported in Appendix B.4. Across the three scales, tuning generally improves IFEval metrics and improves both GPQA Main and Diamond accuracy. MMMU-Pro overall accuracy increases from 50.46% to 52.60% for 4B, from 58.38% to 60.75% for 9B, and from 64.97% to 67.98% for 27B. These results indicate that the educational specialization gains do not require a broad loss of general capability, although changes vary by MMMU-Pro discipline.

5 CONCLUSION

In this work, we present OmniEdu, a family of open foundation models for K–12 learning and teaching, trained with a capability-oriented instruction-tuning corpus. Rather than organizing training data solely by source or subject, we structure supervision around four complementary educational capabilities: *subject competence*, *curriculum grounding*, *diagnostic reasoning*, and *pedagogical action and scaffolding*. We construct a compact, high-quality corpus through a multi-stage pipeline that combines deterministic cleaning, semantic auditing and revision, task-specific quality assessment, and diversity-aware sampling. Experiments across 4B, 9B, and 27B models show consistent improvements across model scales. In particular, the OmniEdu family achieves the strongest open-weight results across the curriculum-grounding and pedagogical-tutoring benchmarks, while OmniEdu-27B remains competitive with frontier proprietary models on K–12 problem solving and substantially outperforms existing open-weight educational models. We additionally use general instruction-following, reasoning, and multimodal evaluations to quantify the broader effects of education-oriented specialization. Overall, our results suggest that carefully curated and capability-balanced supervision provides an effective path toward educational foundation models that go beyond problem solving, enabling stronger understanding of curriculum structure and more effective pedagogical interaction.

AI USE STATEMENT

In this work, we used generative AI tools for several research tasks for which disclosure is required. During data construction, we used Qwen3.5-122B-A10B-FP8 for task-specific semantic auditing of candidate training examples and for targeted revision of defective supervision. The original input was kept fixed during rewriting, and revised examples were re-audited before being retained. We further used GPT-5.6-Terra to produce fine-grained, multidimensional quality scores that informed subsequent rule-based filtering and diversity-aware data selection. Generative AI tools were also used to assist with the implementation of our training pipeline, evaluation pipeline, and data-processing code. All AI-assisted code was reviewed and tested by the authors before being used in the reported experiments.

We did not use generative AI tools for qualitative or thematic data analysis, interpretation of experimental results, the design of or feedback on research methodology or experiments, translation, or the development of conceptual frameworks or research hypotheses. Tasks involving mathematical claims, formal proofs, or assistance with proof construction are not applicable to this work.

Additionally, we used generative AI tools for several tasks for which disclosure is recommended, including language polishing of the manuscript, improving clarity and readability, assisting with literature search and identifying relevant literature, formatting LaTeX and tables, and modifying scientific figures and refining plotting code. All AI-assisted text, code, figures, and other artifacts were reviewed and revised by the authors. We take responsibility for the final content of this work, including its methodology, experiments, results, claims, text, and code, including any content produced or modified with the aid of generative AI.

ETHICS STATEMENT

This work studies foundation models for K–12 learning and teaching. All experiments are conducted using existing datasets, educational resources, and evaluation benchmarks, without direct interaction with students or teachers. During data construction, we retain source and licensing information for provenance tracking, exclude data with unclear usage conditions, and perform data cleaning and evaluation decontamination before training.

Educational language models may produce incorrect answers, misdiagnose student misconceptions, or provide inappropriate pedagogical guidance. Moreover, our evaluation covers only a limited set of benchmarks, subjects, languages, curricula, and educational settings; the reported results therefore should not be assumed to generalize uniformly across all learner populations or deployment contexts. Residual biases or errors in the training data may also propagate to the resulting models. Accordingly, the reported benchmark performance does not establish that these models are suitable for autonomous high-stakes educational decision making or as substitutes for qualified teachers. Appropriate human oversight is recommended when the models are used for consequential assessment, feedback, or personalized instruction.

REPRODUCIBILITY STATEMENT

We provide detailed information to support the reproducibility of our data construction, model training, and evaluation. Section 3 describes the overall data preparation pipeline, including capability-based categorization, deterministic cleaning, LLM-assisted auditing and rewriting, quality filtering, diversity-aware selection, and pedagogical instruction assignment. Section 4.1 describes the experimental setup and the benchmarks used in our evaluation. Appendix A provides additional implementation details, including the data sources and task buckets, auditing and quality-scoring procedures, diversity-selection settings, pedagogical system instructions, training configuration, and evaluation protocols. Complete model-by-model and fine-grained benchmark results are reported in Appendix B.

REFERENCES

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2357–2367, 2019.

Anthropic. Claude for teachers, 2026. URL https://www.anthropic.com/news/claude-for-teachers.

Yuelin Bai, Xeron Du, Yiming Liang, Leo Jin, Junting Zhou, Ziqiang Liu, Feiteng Fang, Mingshan Chang, Tianyu Zheng, Xincheng Zhang, et al. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 8190–8205, 2025.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a better alpaca with fewer data, 2023. URL https://arxiv.org/abs/2307.08701.

Bryan R. Christ, Penelope Molitz, Jonathan Kropko, and Thomas Hartvigsen. Edumath: Generating standards-aligned educational math word problems, 2025. URL https://arxiv.org/abs/2510.06965.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. The english language learner insight, proficiency and skills evaluation (ellipse) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269, 2023.

Scott A Crossley, Perpetual Baffour, L Burleigh, and Jules King. A large-scale corpus for assessing source-based writing quality: Asap 2.0. *Assessing Writing*, 65:100954, 2025.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Stepwise verification and remediation of student reasoning errors with large language model tutors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8386–8411, 2024.

Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, Aimin Zhou, Ze Zhou, Qin Chen, Jie Zhou, Liang He, and Xipeng Qiu. Educhat: A large-scale language model-based chatbot system for intelligent education, 2023. URL https://arxiv.org/abs/2308.02773.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 7768–7791, 2024. URL https://aclanthology.org/2024.acl-long.420/.

Yuyang Ding, Hanglei Hu, Jie Zhou, Qin Chen, Bo Jiang, and Liang He. Boosting large language models with socratic method for conversational mathematics teaching. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3730–3735, 2024.

Rima Hazra, Bikram Ghuku, Ilona Marchenko, Yaroslava Tokarieva, Sayan Layek, Somnath Banerjee, Julia Stoyanovich, and Mykola Pechenizkiy. Safetutors: Benchmarking pedagogical safety in ai tutoring systems, 2026. URL https://arxiv.org/abs/2603.17373.

Runming He, Zhen Hao Wong, Hao Liang, Zimo Meng, Chengyu Shen, Xiaochen Ma, and Wentao Zhang. Dataflow-harness: A grounded code-agent platform for constructing editable llm data pipelines. *arXiv preprint arXiv:2607.16617*, 2026.

Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. Cmmu: A benchmark for chinese multi-modal multi-type question understanding and reasoning, 2024. URL https://arxiv.org/abs/2401.14011.

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

Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, and Min Yang. E-eval: A comprehensive chinese k-12 education evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7753–7774, 2024. URL https://aclanthology.org/2024.findings-acl.462/.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T Morrison. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5376–5384. IEEE, 2017.

Khan Academy. Khanmigo: Ai-powered teaching assistant and tutor, 2023. URL https://khanmigo.ai/.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8082–8090, 2020.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 785–794, 2017.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, et al. Learnlm: Improving gemini for learning, 2024. URL https://arxiv.org/abs/2412.16429.

Bruce W. Lee, Hyunsoo Cho, and Kang Min Yoo. Instruction tuning with human curriculum. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1281–1309, 2024. URL https://aclanthology.org/2024.findings-naacl.82/.

Zhikai Lei, Tianyi Liang, Hanglei Hu, Jin Zhang, Yunhua Zhou, Yunfan Shao, Linyang Li, Chenchui Li, Changbo Wang, Hang Yan, and Qipeng Guo. Gaokao-eval: Does high scores truly reflect strong capabilities in llms?, 2024. URL https://arxiv.org/abs/2412.10056.

Chong Li, Chenglin Zhu, Tao Zhang, Mingan Lin, Zenan Zhou, and Jian Xie. K12-pebench: A process evaluation benchmark for k-12 multimodal reasoning, 2025a. URL https://github.com/lichongod/K12Vista. Released with K12Vista.

Chong Li, Chenglin Zhu, Tao Zhang, Mingan Lin, Zenan Zhou, and Jian Xie. K12vista: Exploring the boundaries of mllms in k-12 education, 2025b. URL https://arxiv.org/abs/2506.01676.

Lucy Li, Tal August, Rose E. Wang, Luca Soldaini, Courtney Allison, and Kyle Lo. Evaluating language model math reasoning via grounding in educational curricula, 2024. URL https://arxiv.org/abs/2408.04226.

Ning Li, Zheng Zhang, Zhenya Huang, Rui Li, Yi Zhan, Yinbo Luo, Qi Liu, and Enhong Chen. Longtutor: Benchmarking large language models for long-term personalized tutoring. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026. URL https://aclanthology.org/2026.acl-long.1371/.

Hao Liang, Xiaochen Ma, Zhou Liu, Zhen Hao Wong, Zhengyang Zhao, Zimo Meng, Runming He, Chengyu Shen, Qifeng Cai, Zhaoyang Han, et al. Dataflow: An llm-driven framework for unified data preparation and workflow automation in the era of data-centric ai. *arXiv preprint arXiv:2512.16676*, 2025.

Hao Liang, Qifeng Cai, Yibo Lin, Jianzhuo Du, Qifeng Xia, Sizhe Qiu, Linzhuang Sun, Meiyi Qiang, Zhaoyang Han, Xiaochen Ma, et al. Dataprep-bench: Benchmarking llms as training data preparators. *arXiv preprint arXiv:2607.20465*, 2026a.

Hao Liang, Qihan Lin, Zhaoyang Han, Xiaochen Ma, Zhen Hao Wong, Meiyi Qiang, Linzhuang Sun, and Wentao Zhang. K12-kgraph: A curriculum-aligned knowledge graph for benchmarking and training educational llms, 2026b. URL https://arxiv.org/abs/2605.09635.

Hao Liang, Zhengyang Zhao, Meiyi Qiang, Mingrui Chen, Lu Ma, Rongyi Yu, Hengyi Feng, Shixuan Sun, Zimo Meng, Xiaochen Ma, et al. Dataflex: A unified framework for data-centric dynamic training of large language models. *arXiv preprint arXiv:2603.26164*, 2026c.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. Socraticlm: Exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37:85693–85721, 2024a.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *International Conference on Learning Representations*, 2024b. URL https://arxiv.org/abs/2312.15685.

Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12585–12591, 2025.

Yijun Liu, Yifan Song, John Gallagher, Sarah Sterman, and Tal August. Foxglove: Understanding goal-oriented and anchored writing feedback from experts and llms on argumentative essays. *arXiv preprint arXiv:2606.06271*, 2026.

Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. *Advances in Neural Information Processing Systems*, 36:32958–32970, 2023.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786, 2021.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022. URL https://scienceqa.github.io/.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5602–5621, 2023.

K. Macina et al. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors, 2025. URL https://eth-lre.github.io/mathtutorbench/.

Kaushal Kumar Maurya, K. V. Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors, 2025. URL https://arxiv.org/abs/2412.09416.

MuduoLLM Contributors from BNU and TAL. Muduollm: A high-performance llm for intelligent education solutions, 2025. URL https://huggingface.co/ERC-ITEA/MuduoLLM.

Multi-Linguality Multi-Functionality Multi-Granularity. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

Numaan Naeem, Abdellah El Mekki, and Muhammad Abdul-Mageed. Eduadapt: A question answer benchmark dataset for evaluating grade-level adaptability in llms, 2025. URL https://arxiv.org/abs/2510.17389.

OpenAI. Introducing study mode, 2025. URL https://openai.com/index/chatgpt-study-mode/.

Shayan Peyghambari Oskoui, Norah Almousa, Zhaoyi Joey Hou, Carolina Gustafson, Gayle Rogers, Raquel Coelho, Diane Litman, and Xiang Lorraine Li. Sefora: Student essays with feedback corpus and llm feedback evaluation framework. *arXiv preprint arXiv:2607.00274*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Hyein Seo, Taewook Hwang, Yohan Lee, and Sangkeun Jung. Feat: A preference feedback dataset through a cost-effective auto-generation and labeling framework for english ai tutoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 575–589, 2025.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, et al. Multi-turn reinforcement learning from preference human feedback, 2024. *URL https://arxiv. org/abs/2405*, 14655.

Mayank Sharma, Roy Pea, and Hari Subramonyam. Convolearn: A dataset of constructivist tutor-student dialogue. *arXiv e-prints*, pp. arXiv–2601, 2026.

Weikang Shi, Houxing Ren, Junting Pan, Aojun Zhou, Ke Wang, Zimu Lu, Yunqiao Yang, Yuxuan Hu, Linda Wei, Mingjie Zhan, and Hongsheng Li. From solver to tutor: Evaluating the pedagogical intelligence of llms with kmp-bench, 2026. URL https://arxiv.org/abs/2603.02775.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4663–4680, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.268. URL https://aclanthology.org/2024.findings-emnlp.268/.

Dingjie Song, Tianlong Xu, Yi-Fan Zhang, Hang Li, Zhiling Yan, Xing Fan, Haoyang Li, Lichao Sun, and Qingsong Wen. Can mllms read students' minds? unpacking multimodal error analysis in handwritten math. In *International Conference on Artificial Intelligence in Education*, pp. 110–124. Springer, 2026.

Rakshith S. Srinivasa, Zora Che, Chen Bo Calvin Zhang, Diego Mares, Ernesto Hernandez, Jayeon Park, Dean Lee, Guillermo Mangialardi, Charmaine Ng, et al. Tutorbench: A benchmark to assess tutoring capabilities of large language models, 2025. URL https://arxiv.org/abs/2510.02663.

Katherine Stasaski, Kimberly Kao, and Marti A Hearst. Cima: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 52–64, 2020.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Investigating prior knowledge for challenging chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155, 2020.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2174–2199, 2024.

Xiaolu Wang, Ziqi Ding, and Liangyu Chen. Automatic tagging of knowledge points for k12 math problems. *arXiv preprint arXiv:2208.09867*, 2022.

Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Jordan Zaykov, Jose Miguel Hernandez-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al. Results and insights from diagnostic questions: The neurips 2020 education challenge. In *NeurIPS 2020 Competition and Demonstration Track*, pp. 191–205. PMLR, 2021.

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

Jing Wu et al. Knowshiftqa: How robust are rag systems when textbook knowledge shifts in k-12 education?, 2025a. URL https://github.com/HKUST-KnowComp/KnowShiftQA.

Lixin Wu, Na Cai, Qiao Cheng, Jiachen Wang, and Yitao Duan. Confucius3-math: A lightweight high-performance reasoning llm for chinese k-12 mathematics learning, 2025b. URL https://arxiv.org/abs/2506.18330.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. Human-ai collaborative essay scoring: A dual-process framework with llms. In *Proceedings of the 15th international learning analytics and knowledge conference*, pp. 293–305, 2025.

Tang Yang et al. Mmtutorbench: The first multimodal benchmark for ai math tutoring, 2026. URL https://github.com/TangciuYueng/MMTutorBench.

Yuqing Ye et al. K-12edubench: A benchmark for evaluating large language models' knowledge, problem-solving, and educational goal cognition in k-12 education, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/40744.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

Kamyar Zeinalipour, Amir Sadeghi, Giovanni Angelini, Leonardo Rigutini, Marco Gori, Marco Maggini, et al. Multihint-9k: Orchestrating human and ai agency for scalable pedagogical hint generation. In *CEUR WORKSHOP PROCEEDINGS*, volume 4248, pp. 144–150. CEUR-WS, 2026.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. Sciinstruct: a self-reflective instruction annotated dataset for training scientific language models. *Advances in Neural Information Processing Systems*, 37:1443–1473, 2024a.

Qian-Wen Zhang, Haochen Wang, Fang Li, Siyu An, Lingfeng Qiao, Liangcai Gao, Di Yin, and Xing Sun. Cjeval: A benchmark for assessing large language models using chinese junior high school exam data, 2024b. URL https://arxiv.org/abs/2409.16202.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems*, 2023a. URL https://arxiv.org/abs/2306.05179.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark, 2023b. URL https://arxiv.org/abs/2305.12474.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=KBMOKmX2he.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023b.

Pengfei Zhou, Xiaopeng Peng, Fanrui Zhang, Zhaopan Xu, Jiaxin Ai, Yansheng Qiu, Chuanhao Li, Zhen Li, Ming Li, Yukang Feng, Jianwen Sun, Haoquan Zhang, Zizhen Li, Xiaofeng Mao, Zekai Li, Wangbo Zhao, Kai Wang, Xiaojun Chang, Wenqi Shao, Yang You, and Kaipeng Zhang. Mdk12-bench: A comprehensive evaluation of multimodal large language models on multidisciplinary exams, 2025. URL https://arxiv.org/abs/2508.06851.

A Reproducibility Details

This appendix provides implementation details that complement the methodology described in the main text. We focus on information needed to reproduce the construction of the final training mixture, model training, and evaluation, while avoiding repetition of the high-level pipeline already described in Section 3.

A.1 Final Data Sources and Task Buckets

Table 4 lists every source that contributes at least one example to the final OmniEdu training corpus, together with its fine-grained task bucket, modality, final size, scoring rubric, system instruction, and license information.

For the general-purpose component, the 9,048 examples are retained as a frozen subset of the final mixture. From DataFlow-Instruct-10K, we retain 7,431 examples after removing code-oriented instances and mathematics problems beyond the K–12 scope, including competition-level problems. The 1,495 examples from the Tulu-3-SFT mixture are drawn from five subsets: Tulu-PersonaHub-IF-V3 (572), Tulu-Aya-General-V3 (348), Tulu-TableGPT-V3 (208), Tulu-CoCoNot-V3 (189), and Tulu-SciRIFF-V3 (178), providing complementary coverage of instruction following, multilingual interaction, table reasoning, refusal and safety behavior, and scientific tasks. The 122 MathV360K examples are drawn from three subsets—GeoQA+ (88), TabMWP (19), and GEOS (15)—to retain diagram- and geometry-based reasoning.

Sources removed during curation are not listed in Table 4 and fall into one of four recorded categories. *Unavailable*: the training split was never published, the official release requires an application we did not hold, or the download link is dead. *License or version conflict*: the license is unknown or non-redistributable, or the source duplicates another source. *Misaligned supervision*: the target is a one- or two-token classification label rather than the rich pedagogical or diagnostic text we require. *Covered elsewhere*: the capability is already represented by a retained source at the same granularity. For each excluded source, the removal reason is recorded in the release manifest.

A.2 Data Curation Implementation

This section documents implementation details of the filtering and selection procedures that are not specified in the main text.

A.2.1 Deterministic Cleaning and Evaluation Decontamination

Normalization and duplicate detection. Cleaning is performed in two passes, within source and then across sources, using only deterministic string operations. Before hashing, a record's supervision-bearing text is NFKC-normalized, all whitespace is collapsed, and case is folded. Duplicate keys are task-family aware: mathematical items are keyed by $(\text{problem}, \text{options})$, short-answer items by $(\text{problem}, \text{student answer}, \text{label})$, and dialogue items by the normalized JSON of the message list. Within a source, the first occurrence of a key is kept and later occurrences are dropped. Across sources, the same key is used, so an item that appears in two datasets is retained once.

Deterministic answer and relevance validation. Deterministic verification is applied wherever the answer can be checked without a model. For multiple-choice items we verify that the option structure is complete and that the annotated answer refers to an existing option. For numerically answerable items we compare the provided answer against the reference or structured annotation after numeric normalization. For structured-label items we verify that the label lies in the declared value domain. Records that fail these checks are discarded before semantic auditing. For mixed-domain sources such as COIG-CQIA and AquilaEdu we retain only the K–12-relevant subset; this filtering uses the source's own domain metadata and task-type fields.

Multimodal validation. For every example that references an image we check that the image can actually be opened and decoded, and that the number of `<image>` placeholders in the user turn equals the number of referenced images. An example with a broken or unopenable image, or an example whose placeholder count and image count disagree is discarded. Multi-image examples are supported and keep their original image order; the released mixture contains up to sixteen images per row.

Evaluation decontamination. Decontamination is defined against the final evaluation set. The check compares the normalized question text of each candidate against the normalized question text of every evaluation item, using exact match after the same normalization as the deduplication pass, and removes matches before the audit queue is built. The check is text-based and is applied to the text component of multimodal items.

Table 4: Data sources contributing to the final OmniEdu training corpus. SR ID refers to the scoring rubric in Table 6, and SP ID refers to the system instruction in Table 8. Entries with multiple SC or SP IDs indicate that examples from the same source are routed to different scoring rubrics or system instructions.

Source	Task Bucket	Modality	SR ID	SP ID	#Samples	Tokens	License
Subject competence							
GSM8K (Cobbe et al., 2021)	math/text/cot	Text	sr_1	sp_2	4,534	560,303	MIT
ScienceQA (text) (Lu et al., 2022)	science/text/cot	Text	sr_1	sp_2	2,124	482,091	CC BY-NC-SA 4.0
ScienceQA (image)	science/multimodal/cot	MM	sr_1	sp_2	2,267	360,171	CC BY-NC-SA 4.0
CMM-Math (text cot) (Liu et al., 2025)	math/text/cot	Text	sr_1	sp_2	1,376	196,204	research use
CMM-Math (text answer)	math/text/answer_only	Text	sr_1	sp_1	159	538	research use
CMM-Math (image)	math/multimodal/cot	MM	sr_1	sp_2	1,647	320,858	research use
MATH (Hendrycks et al., 2021)	math/text/cot	Text	sr_1	sp_2	2,381	467,803	MIT
TQA (text) (Kembhavi et al., 2017)	science/text/cot	Text	sr_1	sp_2	404	56,700	per-row metadata
TQA (diagram)	science/multimodal/cot	MM	sr_1	sp_2	1,900	258,537	per-row metadata
RACE (Lai et al., 2017)	reading/reading-race	Text	sr_1	sp_3	2,240	396,097	research use
C3 (Sun et al., 2020)	reading/reading-c3	Text	sr_1	sp_3	2,158	396,160	CC BY-SA 4.0
COIG-CQIA (Bai et al., 2025)	science/text/cot	Text	sr_1	sp_2	1,804	279,435	Apache-2.0
AquilaEdu[1]	science/text/cot	Text	sr_1	sp_2	1,519	164,922	per-row metadata
CJEval (Zhang et al., 2024b)	science/text/cot	Text	sr_1	sp_2	1,341	191,493	CC BY-NC-SA 4.0
AI2D (Kembhavi et al., 2016)	science/multimodal/cot	MM	sr_1	sp_2	1,071	146,466	CC BY-SA 4.0
SciQ (Welbl et al., 2017)	science/text/cot	Text	sr_1	sp_2	743	90,094	CC BY-NC 3.0
Geometry3K (Lu et al., 2021)	math/multimodal/cot	MM	sr_1	sp_1, sp_2	713	279,401	CC BY 4.0
ARC (Clark et al., 2018)	science/text/cot	Text	sr_1	sp_2	576	108,509	CC BY-SA 4.0
SciInstruct (EN) (Zhang et al., 2024a)	science/text/cot	Text	sr_1	sp_2	471	305,184	research use
QASC (Khot et al., 2020)	science/text/cot	Text	sr_1	sp_2	456	45,410	CC BY 4.0
ELLIPSE (Crossley et al., 2023)	writing_feedback/en	Text	sr_2	sp_4	348	337,856	research use
SciInstruct (CN math)	math/text/cot	Text	sr_1	sp_2	348	111,453	research use
CSEE (Xiao et al., 2025)	writing_feedback/cn	Text	sr_2	sp_4	343	325,334	research use
OpenR1[2]	math/text/cot	Text	sr_1	sp_2	336	533,048	Apache-2.0
ASAP 2.0 (Crossley et al., 2025)	writing_feedback/en	Text	sr_2	sp_4	331	224,083	research use
WorldTree (Jansen et al., 2018)	science/text/cot	Text	sr_1	sp_2	147	18,380	CC BY-NC 4.0
MathQA (Amini et al., 2019)	math/text/cot	Text	sr_1	sp_2	118	9,549	Apache-2.0
Curriculum grounding							
K12-KGraph (text) (Liang et al., 2026b)	knowledge_graph/text	Text	sr_3	sp_6	1,411	74,376	CC BY-NC-SA 4.0
K12-KGraph (image)	knowledge_graph/multimodal/	MM	sr_3	sp_6	2,039	181,141	CC BY-NC-SA 4.0
MathFish-train (strict) (Li et al., 2024)	exercise_standard_mapping/strict	Text	sr_4	sp_7	2,023	160,012	research use
MathFish-train (multi-relation)	exercise_standard_mapping/multirelation	MM	sr_3	sp_8	1,558	225,049	research use
DA-20K (Wang et al., 2022)	general_mapping/mixed_or_text	Text	sr_3	sp_6	831	28,377	CC BY 4.0
Eedi (Wang et al., 2021)	general_mapping/multimodal	MM	sr_3	sp_6	475	54,884	CC BY-NC 4.0
XES3G5M (Liu et al., 2023)	general_mapping/mixed_or_text	MM	sr_3	sp_6	454	15,785	research use
TAL-SCQ5K-CN[3]	general_mapping/mixed_or_text	Text	sr_3	sp_6	312	11,922	research use
TAL-SCQ5K-EN	general_mapping/mixed_or_text	Text	sr_3	sp_6	278	7,700	research use
Diagnostic reasoning							
LongTutor-Synthetic	dialogue/long_conversation_history/synthetic	Text	sr_6	sp_19	2,685	1,203,200	generated
LongTutor-Silver (Li et al., 2026)	dialogue/long_conversation_history/official	Text	sr_6	sp_19	932	1,042,802	research use
ScratchMath (Song et al., 2026)	process/multimodal	MM	sr_5	sp_5	722	119,031	CC BY 4.0
StepVerify (Daheim et al., 2024)	process/text	Text	sr_5	sp_5	401	48,207	MIT
MathDial (Macina et al., 2023)	dialogue/no_conversation_history	Text	sr_5	sp_5	332	152,468	CC BY 4.0
LongTutor-Gold	dialogue/long_conversation_history/official	Text	sr_6	sp_19	245	282,035	research use
Bridge (Wang et al., 2024)	dialogue/short_conversation_history	Text	sr_5	sp_5	172	8,645	research use
Pedagogical action and scaffolding							
GSM8K-rewrite	scaffolding/minimal_scaffold	Text	sr_8	sp_10, sp_9	5,068	265,005	CC BY 4.0 / generated
TutorChat-rewrite	dialogue/multi_turn	Text	sr_9	sp_11	3,492	931,118	generated
OATutor-rewrite	scaffolding/hint_framed_as_a_question	Text	sr_7, sr_10	sp_12, sp_17	1,685	77,443	CC BY 4.0 / generated
MultiHint (Zeinalipour et al., 2026)	scaffolding/progressive_hints_no_question	Text	sr_7	sp_15	1,517	37,177	MIT
SocraTeach (Liu et al., 2024a)	questioning/one_question_per_turn	Text	sr_7	sp_16	647	149,909	CC BY-NC 4.0
ConvoLearn-rewrite	direct_explanation	Text	sr_7	sp_14	540	84,993	MIT
TMATH[4]	questioning/ordered_question_chain	Text	sr_7	sp_18	381	74,152	CC BY 4.0
ScratchMath-rewrite	assessment/feedback_on_a_solution	Text	sr_7	sp_13	259	36,345	CC BY 4.0 / generated
FEAT (Seo et al., 2025)	assessment/feedback_on_writing	Text	sr_7	sp_4	171	5,227	CC BY 4.0
SEFORA (Oskoui et al., 2026)	assessment/feedback_on_writing	Text	sr_7	sp_4	168	6,458	CC BY 4.0
SocraticMATH (Ding et al., 2024)	questioning/one_question_per_turn	Text	sr_7	sp_16	120	24,210	CC BY-NC 4.0
CIMA (Stasaski et al., 2020)	direct_explanation	MM	sr_7	sp_14	71	1,874	CC BY 2.5
Education-Dialogue-Dataset (Shani et al.)	direct_explanation	Text	sr_7	sp_14	40	6,548	generated
ConvoLearn (Sharma et al., 2026)	direct_explanation	Text	sr_7	sp_14	37	10,020	MIT
StepVerify-rewrite	assessment/feedback_on_a_solution	Text	sr_7	sp_13	24	3,821	MIT / generated
FOXGLOVE (Liu et al., 2026)	assessment/feedback_on_writing	Text	sr_7	sp_4	4	283	CC BY-NC-SA 4.0
CJEval-rewrite	dialogue/long_history	Text	sr_7	sp_19	2	1,270	generated
General-purpose retention							
DataFlow-Instruct-10K (Liang et al., 2025)	–	Text	–	–	7,431	3,249,865	per-row metadata
Tulu-3 SFT mixture (Lambert et al., 2024)	–	Text	–	–	1,495	654,482	ODC-BY
MathV360K (Shi et al., 2024)	–	MM	–	–	122	53,856	Apache-2.0
Total	–	–	–	–	**69,999**	**15,955,769**	–

A.2.2 Semantic Auditing and Rewriting

Auditing and decision rules. Semantic auditing is performed using Qwen3.5-122B-A10B-FP8 with task-specific prompts. For each of the four capability categories, the auditor evaluates a set of task-specific validity checks and assigns an overall usability score from 0 to 100. The checklist captures whether critical properties of the supervision are satisfied, while the overall score reflects its general quality and usability for training.

The checklist differs across capability categories. Subject-competence examples are checked for problem validity, answer correctness, and consistency of the explanation; curriculum-grounding examples are checked for label relevance, hierarchy and framework consistency, granularity, and coverage; diagnostic-reasoning examples are checked for whether the identified learner error is supported by the observed work and whether the proposed correction is consistent with that diagnosis; and pedagogical examples are checked for instructional appropriateness, factual correctness, and avoidance of inappropriate answer leakage.

The final action is determined jointly by the checklist and the usability score. An example is retained only if all critical checks pass and its score is at least 85. Examples with a valid underlying input and reference but repairable defects in the supervision, or with scores between 50 and 84, are routed to rewriting. Examples with failed critical validity checks or scores below 50 are removed.

Rewriting and re-auditing. For records routed to rewriting, the original input and any authoritative reference information—such as the correct answer, official score, or curriculum annotation—are kept fixed, and only the defective supervision field is regenerated. Task-specific constraints are imposed during rewriting: explanations must remain grounded in the supplied problem or passage, curriculum repair may modify only the annotation side, and pedagogical responses must realize the intended instructional action without unnecessary answer leakage.

Each rewritten record is subsequently audited again using the same task-specific criteria and hard checks. Only records receiving a *keep* verdict in the second pass are retained; a second *rewrite* or *remove* verdict causes the record to be discarded.

Inference configuration. Auditing and rewriting use vLLM (Kwon et al., 2023) with thinking disabled (`enable_thinking=False`). Auditing uses greedy decoding for stable quality assessment, while rewriting uses task-specific non-zero temperatures to allow flexible regeneration of supervision. The complete generation settings are summarized in Table 5.

A.2.3 Fine-grained Quality Scoring

Scoring setup. Fine-grained quality scoring is performed using GPT-5.6-Terra with `reasoning_effort=none`, `temperature=0.0`, a maximum output length of 2,048 tokens. Each request contains all fields required by the corresponding task, including associated images for multimodal examples. The scorer is instructed to independently verify the item rather than assume that the provided reference answer or annotation is correct. Each applicable quality dimension is scored from 1 to 5, while non-applicable dimensions may be returned as `null`.

Task-specific scoring rubrics. Table 6 summarizes the dimensions used for each task family. Unless otherwise specified by a dimension-specific anchor, scores of 5, 4, 3, 2, and 1 respectively indicate excellent quality, a minor issue, a meaningful but usable weakness, a major weakness, and invalid or fundamentally incorrect supervision.

Filtering policy. We first apply a hard dimensional gate: every applicable dimension must score at least 3, and every task-critical dimension must score at least 4. For buckets in which this gate retains more than 80% of the candidates, we additionally choose a bucket-specific threshold on the mean quality score to obtain a retention rate between 60% and 80%. No additional mean-score threshold is applied when the hard gate already retains at most 80% of the bucket. If multiple examples share the cutoff score, we retain the entire tie group rather than arbitrarily selecting a subset of tied examples.

Output format. Each task-specific prompt requests a JSON object containing exactly the dimensions defined for that rubric, with `null` used only for non-applicable dimensions. The prompts explicitly distinguish answer correctness from reasoning correctness, avoid penalizing answer-only examples for missing reasoning when such reasoning is not required, and instruct the scorer not to reward unnecessary verbosity.

A.2.4 Diversity Selection and Mixture Construction

Embedding and k-center configuration. Each candidate is embedded using BGE-M3 with `max_seq_length=8,192`. The embedding input consists of the normalized record content, excluding provenance and pipeline-tracking metadata, with the task bucket and source prepended to the input. For multi-turn records, turn texts are concatenated in order. Embeddings are L2-normalized, and cosine distance is used for diversity selection.

Within each fine-grained task bucket, we initialize k-center greedy with the example closest to the embedding centroid and iteratively select the example whose nearest selected center is farthest away. For final selection, the procedure terminates when the cumulative number of supervised response

Table 5: Inference configuration for semantic auditing and rewriting.

Stage	Task type	Temp.	Max tokens	Thinking
Audit	All	0.0	8,192	Off
Rewrite	Subject competence	0.3	8,192	Off
	Curriculum grounding	0.3	8,192	Off
	Diagnostic reasoning	0.7	12,288	Off
	Pedagogical action and scaffolding	0.7	12,288	Off

Table 6: Task-specific quality-scoring rubrics. **Bold** dimensions are critical and require a score of at least 4; all other applicable dimensions require a score of at least 3.

ID	Scoring Rubric	Quality Dimensions
sr_1	Problem solving	**problem validity**; **answer correctness**; reasoning correctness; completeness; relevance; clarity.
sr_2	Writing feedback	**task and score validity**; **score alignment**; **feedback grounding**; diagnostic specificity; actionability and completeness; relevance and clarity.
sr_3	Curriculum Label	**input validity**; **label or relation correctness**; **hierarchy and framework consistency**; granularity and coverage; grounding and explanation quality; instruction compliance and clarity.
sr_4	Strict standard alignment	**pair validity**; **binary-label correctness**; direct-requirement precision; standard specificity; instruction and format compliance.
sr_5	Error diagnosis	**input validity**; **diagnosis correctness**; **evidence grounding**; **correction and content correctness**; pedagogical quality; context consistency and completeness; clarity and efficiency.
sr_6	Longitudinal tutoring	**history and input validity**; **memory correctness**; **diagnosis evidence**; **strategy alignment**; **tutoring correctness**; personalization and pedagogy; answer leakage and efficiency; output consistency and clarity.
sr_7	General pedagogical action	**input and task validity**; **correctness and grounding**; **pedagogical appropriateness**; learner-state adaptation; guidance and engagement; completeness and instruction compliance; clarity and efficiency.
sr_8	Minimal scaffold	**problem and attempt validity**; error diagnosis; **mathematical correctness**; scaffolding quality; **non-leakage**; targeting and relevance; style and efficiency.
sr_9	Multi-turn dialogue	**context and input validity**; **correctness**; relevance; **diagnosis and feedback**; **pedagogy and clarity**; personalization and continuity; engagement and efficiency.
sr_10	Hint-based guidance	**input validity**; **correctness**; **hint grounding**; **next-step quality**; **non-leakage**; pedagogical actionability; clarity and efficiency.

tokens reaches the bucket-specific budget. The final selected example is retained even when it causes a small budget overshoot. Algorithm 1 summarizes the procedure.

The preliminary source-level compression uses the same k-center procedure, replacing the token-budget stopping criterion with a fixed target number of examples. In addition to the examples reported in the main text, OpenR1 is reduced from 25,201 to 1,000 examples, SciInstruct-EN from 29,010 to 1,000, SciQ from 11,052 to 1,000, and MathQA from 16,866 to 1,000.

Bucket-level token budgets. Table 7 reports the candidate size, supervised-response-token budget, and final composition of each fine-grained task bucket.

A.3 Pedagogical System Instructions

Each education-specific example is assigned one task-specific system instruction according to its fine-grained task bucket. We use 19 instruction templates in the final education-specific corpus, covering problem solving, curriculum grounding, diagnosis, and different forms of pedagogical interaction. The 9,048 frozen general-purpose examples retain their original system prompts and are not remapped to these templates.

Table 8 lists the instruction templates and their frequencies in the final corpus. The exact mapping from fine-grained task buckets to instruction IDs is deterministic and is provided with the released configuration.

A.4 Training Configuration

Table 9 summarizes the training configuration used for the OmniEdu models.

Algorithm 1: Token-budgeted k-center greedy selection

Input: Candidate set $\mathcal{D} = \{x_i\}_{i=1}^{N}$; normalized embeddings $\{z_i\}_{i=1}^{N}$; supervised-token counts $\{t_i\}_{i=1}^{N}$; token budget B
Output: Selected subset $\mathcal{S}$

$\mu \leftarrow \frac{1}{N}\sum_{i=1}^{N} z_i$;
$\hat{\mu} \leftarrow \mu / \|\mu\|_2$;
$c \leftarrow \arg\min_i \left(1 - z_i^\top \hat{\mu}\right)$;
$\mathcal{S} \leftarrow \{c\}$;
$T \leftarrow t_c$;
foreach $i \neq c$ **do**
  $d_i \leftarrow 1 - z_i^\top z_c$;
while $T < B$ ***and*** $|\mathcal{S}| < N$ **do**
  $j \leftarrow \arg\max_{i \notin \mathcal{S}} d_i$;
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{j\}$;
  $T \leftarrow T + t_j$;
  foreach $i \notin \mathcal{S}$ **do**
    $d_i \leftarrow \min\left(d_i,\ 1 - z_i^\top z_j\right)$;
return $\mathcal{S}$;

Multimodal rows are rendered by the release assembly described in Section 3: every referenced image is materialised into the release directory, and the user turn carries an `<image>` placeholder at the position where the image belongs. Images are passed to the training template at their native resolution through the LLaMA-Factory image pipeline, with no additional resizing, cropping, or augmentation.

A.5 Evaluation Protocol

General inference configuration. We follow the official evaluation split, prompting protocol, output parsing, and scoring implementation of each benchmark. Unless otherwise specified, we use `temperature=0.0` and generate one response per example. Open-weight models are evaluated with vLLM using the chat template distributed with each checkpoint, while proprietary models are queried through their corresponding API endpoints. The pedagogical system instructions used during training are not added at evaluation time.

LLM-based evaluation. For benchmarks requiring model-based grading, model responses are generated first and scored separately using benchmark-specific judge prompts and aggregation rules. We use *GPT-4.1-mini* as the judge with `temperature=0.0`, and each response is judged once without voting. The maximum judge output length is 1,024 tokens for GAOKAO-Bench subjective questions, MDK12-Bench open-ended questions, and LongTutor, 2,048 tokens for EDUMATH, and 6,144 tokens for TutorBench. MathTutorBench instead uses the officially released *Qwen2.5-1.5B-pedagogical-rewardmodel* and the benchmark's official scoring script.

Multimodal evaluation. For multimodal-capable models, benchmark images are provided directly together with the associated textual input using the model's native multimodal interface. We apply no additional image augmentation beyond the model's standard processor. For models without image-input support, we provide only the textual component of image-dependent examples, as described in the main text.

Proprietary-model evaluation. Proprietary baselines are evaluated on the same benchmark examples and with the same benchmark-specific prompts as the open-weight models. Table 10 reports the exact API identifiers and generation settings used in our experiments. For comparability, we use temperature 0, top-p 1.0, and disabled reasoning whenever supported; *GLM-5.3* and *Kimi-K3* instead use the lowest available reasoning effort, while *Kimi-K3* also uses the gateway-supported

Table 7: Candidate sizes, supervised-response-token budgets, and final composition of the fine-grained task buckets.

Task Bucket	Candidates	Token Budget	Final Examples	Final Tokens
Subject competence				
math/text/answer_only	159	538	159	538
math/text/cot	19,765	1,878,360	9,093	1,878,360
math/multimodal/answer_only	2	24	2	24
math/multimodal/cot	2,979	600,000	2,358	600,235
reading/reading-c3	5,172	396,000	2,158	396,160
reading/reading-race	2,427	396,000	2,240	396,097
science/text/cot	23,207	1,742,000	9,585	1,742,218
science/multimodal/cot	11,894	765,000	5,238	765,174
writing_feedback/en	679	561,939	679	561,939
writing_feedback/cn	343	325,334	343	325,334
Curriculum grounding				
general_mapping/mixed_or_text	1,875	63,784	1,875	63,784
general_mapping/multimodal	475	54,884	475	54,884
knowledge_graph/text	1,814	74,376	1,411	74,376
knowledge_graph/multimodal/	5,068	181,000	2,039	181,141
exercise_standard_mapping/multirelation	5,515	225,000	1,558	225,049
exercise_standard_mapping/strict	8,799	160,000	2,023	160,012
Diagnostic reasoning				
process/text	401	48,207	401	48,207
process/multimodal	999	119,000	722	119,031
dialogue/no_conversation_history	332	152,468	332	152,468
dialogue/short_conversation_history	172	8,645	172	8,645
dialogue/long_conversation_history/official	1,674	1,324,000	1,177	1,324,837
dialogue/long_conversation_history/synthetic	3,025	1,203,000	2,685	1,203,200
Pedagogical action and scaffolding				
scaffolding/minimal_scaffold	8,571	265,000	5,068	265,005
scaffolding/hint_framed_as_a_question	2,535	77,443	1,685	77,443
scaffolding/progressive_hints_no_question	3,382	37,175	1,517	37,177
questioning/one_question_per_turn	1,536	174,000	767	174,119
questioning/ordered_question_chain	492	74,000	381	74,152
multi_turn_dialogue	6,690	931,000	3,492	931,118
assessment/feedback_on_a_solution	303	40,000	283	40,166
assessment/feedback_on_writing	343	11,968	343	11,968
direct_explanation	688	103,435	688	103,435
dialogue/long_conversation_history	2	1,270	2	1,270
Education-specific total	**121,318**	–	**60,951**	**11,997,566**
General-purpose total	9,048	–	9,048	3,958,203
Combined	–	–	**69,999**	**15,955,769**

temperature 1.0 and top-p 0.95. These differences reflect API constraints rather than model-specific tuning.

Answer parsing. We follow the official parsing and scoring implementation of each benchmark. Structured or multiple-choice outputs are parsed using the benchmark's required answer format, whereas free-form responses are passed to the corresponding official scorer or judge. We do not re-prompt, repair, or manually modify model outputs before scoring.

B Detailed Evaluation Results

This appendix reports the complete model-by-model results underlying the compact summary tables in the main text. The original table structure is retained to make every metric and baseline directly auditable.

B.1 Curriculum Grounding

Tables 11, 12, and 13 provide detailed results on the three curriculum-grounding benchmarks. K12-Bench reports performance across different types of curriculum relations, MathFish separates positive and negative curriculum-alignment cases, and EDUMATH reports complementary metrics for

curriculum-grounded problem generation. Overall, the detailed results show that the improvements after education-oriented tuning extend across multiple forms of curriculum understanding rather than being driven by a single evaluation setting.

B.2 K–12 Problem Solving

Tables 14, 15, and 16 provide the complete breakdowns for the three K–12 problem-solving benchmarks. GAOKAO-Bench reports results by question type and school subject, EXAMS-V further distinguishes modalities, subject groups, and languages, and MDK12-Bench separates objective and open-ended questions. The breakdowns show that the gains from education-oriented tuning generally extend across different subjects, modalities, and problem formats, although the magnitude of improvement varies across individual subsets.

B.3 Pedagogical Tutoring

Tables 17, 18, and 19 provide fine-grained results for the three pedagogical-tutoring benchmarks. MathTutorBench evaluates scaffolded and pedagogy-oriented tutoring under standard and harder settings, TutorBench covers multiple tutoring use cases in both text and multimodal settings, and LongTutor evaluates evidence use, knowledge-state diagnosis, and teaching quality over longer interaction histories. Across these complementary settings, education-oriented tuning consistently strengthens the models' ability to provide pedagogically appropriate and context-sensitive tutoring responses.

B.4 General Capability Retention

In Tables 20, 21, and 22, we report the complete general-capability diagnostics used to measure whether education-oriented tuning changes instruction following, scientific reasoning, or multimodal reasoning. The paired base/SFT presentation makes the effect of tuning directly auditable.

Table 8: Task-specific system instructions used in the final education-specific OmniEdu corpus.

ID	Behavior	Examples	System Instruction
sp_1	Answer-only solving	161	Solve the problem independently and accurately using all provided text and visual evidence. Return only the final answer in the exact format requested, without explanation.
sp_2	Reasoned solving	26,274	Solve the problem independently and accurately using all provided text and visual evidence. Give a clear, self-contained derivation and finish with one unambiguous final answer in the requested format.
sp_3	Reading comprehension	4,398	Answer the reading-comprehension question from the supplied passage and options. Base the answer only on the passage, reason carefully when needed, and follow the requested answer format.
sp_4	Writing feedback	1,365	Evaluate the student's writing against the supplied task and rubric. Give specific, evidence-based feedback that identifies strengths and the most useful improvements without inventing unsupported details.
sp_5	Error diagnosis and correction	1,627	Diagnose the learner's actual error from the supplied work or dialogue. Preserve correct work, identify the first material issue, explain why it matters, and provide the most useful next correction or resolution.
sp_6	Curriculum grounding	5,800	This is a curriculum and knowledge-point task. Identify or judge the relevant concepts, standards, prerequisites, or relations from the provided information, including images when present, and follow the requested output format exactly.
sp_7	Strict standard alignment	2,023	Judge whether the proposed curriculum alignment is valid under the requested MathFish verification criterion. Answer Yes or No and give one brief, concrete reason.
sp_8	Multi-relation standard alignment	1,558	Analyze the curriculum relationship requested by the MathFish task. Distinguish direct alignment from prerequisites, neighboring skills, and later applications, and return only the requested relation format.
sp_9	Minimal scaffold	100	You are an experienced math teacher responding to a student in a useful and caring way. Give the smallest helpful next scaffold and keep the response to at most two sentences.
sp_10	One-question guided scaffold	4,968	Be a friendly, supportive tutor. Guide the student through incremental steps with guiding and probing questions. Pose just one question in this turn, avoid overwhelming the student, and keep the response to at most two sentences.
sp_11	Multi-turn dialogue	3,492	You are a highly capable, supportive tutor. Respond directly to the student. Be accurate, pedagogically helpful, appropriately calibrated to the student, and avoid unnecessary verbosity.
sp_12	Active probing	242	Facilitate active learning with one actionable question that asks the learner to calculate, compare, retrieve, test, explain, or inspect something relevant. Do not reveal the final answer.
sp_13	Answer assessment	283	Assess the learner's answer accurately. State what is correct, identify the first important error or omission, and give focused feedback or the next correction without unnecessary re-solving.
sp_14	Direct explanation	688	Give a clear, accurate explanation adapted to the learner's request. Cover the essential idea and reasoning directly, using a useful example or representation when appropriate.
sp_15	Progressive hints	1,517	Provide a short sequence of progressively stronger hints. Begin with the least revealing useful hint, keep each hint actionable, and do not disclose the final answer unless explicitly requested.
sp_16	Multi-turn Socratic tutoring	767	Tutor through a coherent Socratic dialogue. Use the conversation history, ask focused questions that build on the learner's responses, and advance one conceptual step at a time.
sp_17	Next-step guidance	1,443	Guide the learner with a targeted next step based on the current problem and work. Prefer a concise hint or question that makes the learner reason instead of giving away the final answer.
sp_18	Socratic question chain	381	Use a concise chain of logically ordered Socratic questions that leads the learner toward the key idea. Each question should have a clear purpose and must not reveal the final answer.
sp_19	Longitudinal tutoring	3,864	Uses the native LongTutor instruction, which conditions tutoring on longitudinal learner history and requires structured memory, diagnosis, strategy, and teaching outputs. The complete instruction is provided verbatim in the released system-prompt registry.

Table 9: Training configuration for OmniEdu.

Component	Setting
Backbones	Qwen3.5-4B-Base; Qwen3.5-9B-Base; Qwen3.8-27B
Framework	LLaMA-Factory
Training Stage	Supervised fine-tuning (SFT)
Fine-tuning Type	Full-parameter fine-tuning
Template	`qwen3_5_nothink`
Max Sequence Length	32,768
Packing	False
Per-device Batch Size	1
Gradient Accumulation Steps	8
Effective Global Batch Size	64 (4B, 9B); 128 (27B)
Accelerators	8 (4B, 9B); 16 across two nodes (27B)
Optimizer	AdamW (fused)
Adam β_1, β_2	0.9, 0.999
Adam ϵ	10^{-8}
Learning Rate	5×10^{-6}
LR Scheduler	Cosine
Warmup Ratio	0.1
Weight Decay	0.0
Gradient Clipping	1.0
Random Seed	42
Epochs	3
Precision	bf16
Gradient Checkpointing	Enabled
DeepSpeed	ZeRO-3
Preprocessing Workers	1
Dataloader Workers	4
Environment	Transformers 5.2.0; PyTorch 2.10.0; Datasets 4.0.0; Tokenizers 0.22.2

Table 10: API identifiers and generation settings for the proprietary baselines.

Model	Provider	API Identifier	Generation Protocol	Modality
GPT-5.4	OpenAI	`gpt-5.4`	temp. 0.0; top-p 1.0; reasoning off	Text + Image
GPT-5.6-Sol	OpenAI	`gpt-5.6-sol`	temp. 0.0; top-p 1.0; thinking off	Text + Image
Claude-Opus-5	Anthropic	`claude-opus-5`	temp. 0.0; top-p 1.0; thinking off	Text + Image
GLM-5.3	Zhipu AI	`glm-5.3`	temp. 0.0; top-p 1.0; reasoning effort low	Text
Kimi-K3	Moonshot AI	`kimi-k3`	temp. 1.0; top-p 0.95; reasoning effort low	Text + Image

Table 11: Detailed K12-Bench results across curriculum-grounding subtasks. Top-three values are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	Ground		Prereq		Neighbor		Evidence		Locate		Overall	
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Open-weight Models													
Qwen3.5-4B-Base	4B	47.16%	78.65%	41.53%	73.44%	66.83%	71.52%	28.48%	69.61%	15.53%	59.73%	42.72%	69.20%
OmniEdu-4B (ours)	4B	59.74%	82.59%	60.55%	79.26%	70.51%	71.13%	37.95%	72.71%	35.59%	63.86%	54.25%	71.75%
Qwen3.5-9B-Base	9B	55.93%	81.86%	50.03%	77.65%	70.44%	73.94%	31.51%	70.91%	24.33%	63.21%	48.52%	71.99%
OmniEdu-9B (ours)	9B	60.46%	83.34%	62.92%	82.26%	72.68%	74.13%	38.00%	72.94%	35.50%	64.98%	**55.46%**	**73.68%**
Qwen3.8-27B	27B	54.35%	81.23%	60.32%	83.56%	74.13%	74.90%	28.22%	70.37%	29.85%	64.59%	52.11%	73.48%
OmniEdu-27B (ours)	27B	70.80%	87.65%	73.67%	87.54%	75.85%	76.37%	46.25%	76.57%	48.77%	67.82%	**63.12%**	**76.69%**
Confucius3-Math	14B	5.68%	22.46%	2.99%	11.79%	7.12%	24.99%	1.47%	5.11%	6.74%	25.86%	5.23%	19.23%
MuduoLLM	14B	47.95%	77.66%	54.29%	78.53%	70.29%	72.01%	29.00%	69.04%	24.37%	61.54%	48.24%	70.46%
EduChat-SFT-Qwen2.5-7B	7B	50.40%	78.34%	52.78%	78.19%	63.89%	66.90%	33.50%	69.59%	18.13%	56.73%	45.20%	67.58%
EduChat-R1-Qwen3-8B	8B	30.55%	59.12%	23.83%	58.45%	40.39%	55.09%	15.14%	52.56%	8.93%	30.03%	25.25%	49.67%
EduChat-R1-Qwen3-32B	32B	56.29%	81.58%	58.00%	81.44%	71.39%	72.39%	29.05%	69.12%	22.96%	61.08%	49.42%	71.21%
Proprietary Models													
GPT-5.4	–	43.17%	74.42%	48.31%	77.71%	70.14%	71.16%	11.35%	59.59%	22.44%	60.31%	43.23%	67.80%
GPT-5.6-Sol	–	43.53%	74.83%	50.00%	74.81%	78.78%	79.72%	11.24%	58.59%	31.02%	63.54%	48.63%	71.03%
Claude-Opus-5	–	43.17%	75.69%	53.25%	81.03%	75.65%	76.63%	14.68%	61.17%	30.38%	63.09%	48.48%	71.34%
GLM-5.3	–	43.53%	62.29%	55.32%	72.02%	64.81%	65.31%	12.50%	44.14%	25.58%	44.94%	43.38%	57.66%
Kimi-K3	–	57.19%	80.79%	63.25%	82.21%	81.97%	81.99%	15.25%	61.76%	37.77%	66.63%	**55.03%**	**74.71%**

Table 12: Detailed MathFish results across positive and negative curriculum-alignment cases. Top-three values are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	Pos: Should Yes	Neg: ATC Neighbor	Neg: Same-domain		Neg: Diff-domain		Overall Acc
				Same-grade	Diff-grade	Same-grade	Diff-grade	
Open-weight Models								
Qwen3.5-4B-Base	4B	51.20%	77.70%	86.40%	90.70%	97.10%	98.60%	80.33%
OmniEdu-4B (ours)	4B	63.74%	73.22%	87.91%	89.50%	98.25%	99.64%	83.19%
Qwen3.5-9B-Base	9B	33.40%	88.10%	93.40%	95.90%	99.10%	99.50%	79.66%
OmniEdu-9B (ours)	9B	62.00%	76.50%	89.80%	91.10%	98.10%	99.50%	83.70%
Qwen3.8-27B	27B	45.95%	88.31%	95.86%	97.01%	99.64%	99.95%	83.54%
OmniEdu-27B (ours)	27B	67.60%	79.07%	89.50%	92.67%	98.97%	99.74%	**85.89%**
Confucius3-Math	14B	0.00%	0.20%	0.10%	0.10%	0.10%	0.10%	0.00%
MuduoLLM	14B	66.10%	71.00%	84.40%	89.70%	98.00%	99.20%	82.85%
EduChat-SFT-Qwen2.5-7B	7B	42.00%	78.20%	87.00%	90.40%	97.50%	98.70%	78.23%
EduChat-R1-Qwen3-8B	8B	12.30%	22.30%	32.00%	38.20%	49.70%	55.30%	19.49%
EduChat-R1-Qwen3-32B	32B	74.70%	66.50%	82.00%	88.80%	97.50%	99.00%	83.71%
Proprietary Models								
GPT-5.4	–	73.49%	61.76%	76.90%	81.68%	96.40%	98.20%	80.65%
GPT-5.6-Sol	–	65.27%	75.97%	86.88%	90.58%	96.92%	98.71%	83.68%
Claude-Opus-5	–	52.71%	83.98%	93.70%	93.46%	98.71%	99.23%	83.52%
GLM-5.3	–	66.67%	77.78%	89.50%	91.62%	98.97%	99.50%	**85.23%**
Kimi-K3	–	68.60%	75.50%	89.40%	92.60%	98.30%	99.50%	**85.40%**

Table 13: Detailed EDUMATH results on curriculum-grounded problem generation. Top-three values are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	Format Pass	Classifier GOOD	MaC
Open-weight Models				
Qwen3.5-4B-Base	4B	95.90%	86.90%	47.60%
OmniEdu-4B (ours)	4B	99.20%	90.20%	68.40%
Qwen3.5-9B-Base	9B	98.10%	89.20%	58.60%
OmniEdu-9B (ours)	9B	99.80%	91.70%	74.00%
Qwen3.8-27B	27B	100.00%	91.00%	70.60%
OmniEdu-27B (ours)	27B	98.80%	92.60%	**86.95%**
Confucius3-Math	14B	11.70%	91.90%	3.00%
MuduoLLM	14B	92.70%	90.60%	57.20%
EduChat-SFT-Qwen2.5-7B	7B	15.60%	96.10%	28.00%
EduChat-R1-Qwen3-8B	8B	36.60%	93.40%	47.00%
EduChat-R1-Qwen3-32B	32B	67.60%	91.80%	61.20%
Proprietary Models				
GPT-5.4	–	100.00%	94.50%	80.50%
GPT-5.6-Sol	–	99.50%	89.45%	**84.92%**
Claude-Opus-5	–	100.00%	75.50%	75.00%
GLM-5.3	–	96.77%	87.78%	62.78%
Kimi-K3	–	100.00%	84.00%	**90.00%**

Table 14: Detailed GAOKAO-Bench results by question type and school subject. Top-three values across all models are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	Obj Rate	Subj Rate	Full Rate	Chinese	English	Math	Physics	Chemistry	Biology	Politics	History	Geography
Open-weight Models													
Qwen3.5-4B-Base	4B	85.00%	92.07%	88.89%	90.81%	86.52%	89.39%	88.55%	86.05%	89.27%	88.44%	89.49%	88.00%
OmniEdu-4B (ours)	4B	87.10%	93.19%	90.45%	92.10%	88.41%	90.89%	90.16%	88.01%	90.78%	90.07%	90.97%	89.68%
Qwen3.5-9B-Base	9B	90.70%	94.79%	92.95%	94.06%	91.58%	93.24%	92.76%	91.31%	93.17%	92.69%	93.30%	92.44%
OmniEdu-9B (ours)	9B	91.50%	95.44%	93.67%	94.74%	92.35%	93.95%	93.48%	92.09%	93.88%	93.42%	94.00%	93.17%
Qwen3.8-27B	27B	84.00%	97.73%	91.55%	95.27%	86.96%	92.53%	90.90%	86.04%	92.29%	90.69%	92.73%	89.82%
OmniEdu-27B (ours)	27B	91.03%	**98.02%**	**94.87%**	96.77%	92.53%	95.37%	94.54%	92.07%	95.25%	94.44%	95.47%	94.00%
Confucius3-Math	14B	70.50%	96.00%	84.51%	91.44%	75.99%	86.35%	83.32%	74.29%	85.89%	82.92%	86.71%	81.32%
MuduoLLM	14B	87.30%	91.60%	89.66%	90.83%	88.23%	89.97%	89.46%	87.94%	89.90%	89.39%	90.03%	89.12%
EduChat-SFT-Qwen2.5-7B	7B	66.60%	74.13%	70.74%	72.78%	68.22%	71.28%	70.38%	67.72%	71.14%	70.27%	71.39%	69.79%
EduChat-R1-Qwen3-8B	8B	73.70%	78.00%	76.06%	77.23%	74.63%	76.37%	75.86%	74.34%	76.30%	75.79%	76.43%	75.52%
EduChat-R1-Qwen3-32B	32B	87.20%	81.15%	83.87%	82.23%	85.90%	83.44%	84.16%	86.30%	83.55%	84.25%	83.35%	84.63%
Proprietary Models													
GPT-5.4	–	88.09%	97.83%	93.44%	96.09%	90.19%	94.14%	92.99%	89.54%	93.97%	92.84%	94.28%	92.22%
GPT-5.6-Sol	–	**95.67%**	**98.89%**	**97.44%**	98.31%	96.36%	97.67%	97.29%	96.15%	97.61%	97.24%	97.72%	97.04%
Claude-Opus-5	–	**96.96%**	**99.07%**	**98.12%**	98.69%	97.41%	98.27%	98.02%	97.27%	98.23%	97.99%	98.30%	97.86%
GLM-5.3	–	93.33%	96.29%	94.87%	95.76%	93.00%	95.18%	94.83%	93.79%	95.12%	94.78%	95.22%	94.60%
Kimi-K3	–	**96.16%**	93.81%	94.87%	94.23%	95.65%	94.70%	94.98%	95.81%	94.74%	95.02%	94.67%	95.16%

Table 15: Detailed EXAMS-V results across modalities, subject groups, and languages. Top-three values in Overall, Text, and Image+Text are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	Overall	Text	Image+Text	Natural Science	Social Sciences	Other	Chinese	English
Open-weight Models									
Qwen3.5-4B-Base	4B	44.69%	48.57%	33.33%	41.31%	53.03%	36.68%	21.67%	23.05%
OmniEdu-4B (ours)	4B	57.62%	59.00%	53.56%	56.16%	64.46%	44.71%	50.17%	41.21%
Qwen3.5-9B-Base	9B	63.00%	64.68%	58.07%	60.65%	71.58%	49.27%	58.00%	40.35%
OmniEdu-9B (ours)	9B	66.40%	68.04%	61.59%	64.73%	73.58%	53.47%	59.33%	48.99%
Qwen3.8-27B	27B	**68.65%**	**69.02%**	**67.57%**	69.22%	72.83%	53.65%	65.50%	42.94%
OmniEdu-27B (ours)	27B	**69.52%**	**69.97%**	**68.22%**	70.62%	73.52%	52.55%	67.17%	45.24%
Confucius3-Math	14B	21.95%	22.26%	21.05%	21.37%	23.24%	20.99%	21.33%	25.94%
MuduoLLM	14B	21.95%	22.26%	21.05%	21.37%	23.24%	20.99%	21.33%	25.94%
EduChat-SFT-Qwen2.5-7B	7B	21.95%	22.26%	21.05%	21.37%	23.24%	20.99%	21.33%	25.94%
EduChat-R1-Qwen3-8B	8B	0.69%	0.92%	0.00%	0.72%	0.37%	1.46%	0.00%	0.00%
EduChat-R1-Qwen3-32B	32B	21.95%	22.26%	21.05%	21.37%	23.24%	20.99%	21.33%	25.94%
Proprietary Models									
GPT-5.4	–	35.66%	35.89%	34.98%	36.11%	37.93%	27.03%	44.17%	23.19%
GPT-5.6-Sol	–	24.82%	24.86%	24.69%	24.76%	26.02%	21.62%	31.67%	18.84%
Claude-Opus-5	–	64.34%	64.94%	62.55%	68.05%	63.01%	50.45%	64.17%	59.42%
GLM-5.3	–	22.29%	22.83%	20.90%	23.88%	21.52%	16.67%	21.74%	27.27%
Kimi-K3	–	**87.29%**	**86.99%**	**88.06%**	90.30%	85.44%	77.78%	92.75%	81.82%

Table 16: Detailed MDK12-Bench results across objective and open-ended question types. Top-three values in Objective Avg, Open Correctness, and Full Overall are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	Single-choice	Multiple-choice	True/False	Objective Avg	Open Correctness	Full Overall
Open-weight Models							
Qwen3.5-4B-Base	4B	30.99%	9.79%	28.43%	23.17%	49.54%	35.43%
OmniEdu-4B (ours)	4B	39.90%	7.60%	38.46%	28.66%	66.74%	46.36%
Qwen3.5-9B-Base	9B	49.92%	10.11%	18.52%	27.94%	63.56%	44.50%
OmniEdu-9B (ours)	9B	49.60%	15.98%	38.86%	35.37%	68.55%	50.80%
Qwen3.8-27B	27B	56.34%	13.42%	17.32%	32.10%	62.08%	46.04%
OmniEdu-27B (ours)	27B	53.97%	25.02%	31.74%	38.15%	**78.77%**	**57.76%**
Confucius3-Math	14B	34.16%	4.19%	14.53%	18.72%	62.63%	39.13%
MuduoLLM	14B	55.57%	11.60%	20.23%	31.12%	61.92%	45.44%
EduChat-SFT-Qwen2.5-7B	7B	37.45%	7.97%	10.77%	20.23%	46.62%	32.50%
EduChat-R1-Qwen3-8B	8B	52.65%	4.94%	25.36%	29.15%	55.27%	41.29%
EduChat-R1-Qwen3-32B	32B	52.53%	12.26%	7.58%	26.68%	59.30%	41.84%
Proprietary Models							
GPT-5.4	–	44.44%	35.20%	32.47%	38.05%	73.80%	54.77%
GPT-5.6-Sol	–	51.11%	41.26%	29.02%	**41.75%**	69.82%	54.67%
Claude-Opus-5	–	61.82%	23.08%	22.13%	37.89%	**78.43%**	**57.46%**
GLM-5.3	–	57.08%	31.03%	30.86%	**41.10%**	70.57%	55.33%
Kimi-K3	–	66.67%	29.06%	32.57%	**44.66%**	**83.89%**	**63.60%**

Table 17: Detailed MathTutorBench results across scaffolding and pedagogy settings. Top-three values are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	Scaffold WR	Scaffold-hard WR	Pedagogy WR	Pedagogy-hard WR
Open-weight Models					
Qwen3.5-4B-Base	4B	20.42%	18.36%	37.05%	50.78%
OmniEdu-4B (ours)	4B	75.79%	**84.77%**	76.53%	**86.33%**
Qwen3.5-9B-Base	9B	14.00%	13.67%	55.47%	56.64%
OmniEdu-9B (ours)	9B	75.26%	81.64%	78.00%	80.47%
Qwen3.8-27B	27B	57.16%	55.86%	74.74%	73.83%
OmniEdu-27B (ours)	27B	**78.74%**	**83.59%**	**79.79%**	82.81%
Confucius3-Math	14B	27.68%	20.70%	28.74%	25.39%
MuduoLLM	14B	30.42%	23.44%	50.95%	37.50%
EduChat-SFT-Qwen2.5-7B	7B	18.21%	22.66%	20.21%	28.52%
EduChat-R1-Qwen3-8B	8B	11.58%	7.42%	33.58%	20.70%
EduChat-R1-Qwen3-32B	32B	22.21%	16.80%	51.68%	33.20%
Proprietary Models					
GPT-5.4	–	6.32%	1.96%	33.16%	41.18%
GPT-5.6-Sol	–	10.53%	17.65%	60.00%	56.86%
Claude-Opus-5	–	**87.89%**	**84.31%**	**78.95%**	**88.24%**
GLM-5.3	–	**77.89%**	76.92%	**83.16%**	**84.62%**
Kimi-K3	–	20.00%	15.38%	69.47%	69.23%

Table 18: Detailed TutorBench results across tutoring use cases and input modalities. Top-three values are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	UC1		UC2		UC3		Text	Multimodal	Overall
		Text	Multimodal	Text	Multimodal	Text	Multimodal			
Open-weight Models										
Qwen3.5-4B-Base	4B	55.42%	52.75%	62.50%	38.10%	48.49%	31.01%	55.72%	37.91%	45.52%
OmniEdu-4B (ours)	4B	49.78%	53.66%	61.07%	36.69%	54.22%	40.47%	53.90%	41.05%	46.67%
Qwen3.5-9B-Base	9B	56.26%	56.89%	53.43%	38.35%	40.47%	35.05%	51.88%	40.46%	45.38%
OmniEdu-9B (ours)	9B	54.49%	54.88%	54.09%	41.31%	56.71%	40.24%	54.92%	43.06%	48.16%
Qwen3.8-27B	27B	69.68%	70.02%	70.68%	54.00%	54.72%	43.14%	**66.42%**	**52.64%**	**58.58%**
OmniEdu-27B (ours)	27B	65.67%	65.49%	74.31%	54.74%	61.38%	47.87%	**67.00%**	**53.79%**	**59.42%**
Confucius3-Math	14B	47.64%	49.67%	50.65%	5.99%	53.48%	30.10%	49.80%	23.34%	34.72%
MuduoLLM	14B	50.56%	51.37%	64.25%	9.88%	51.52%	23.16%	54.32%	22.48%	35.98%
EduChat-SFT-Qwen2.5-7B	7B	28.62%	27.04%	28.47%	5.44%	41.95%	17.95%	31.56%	14.19%	21.48%
EduChat-R1-Qwen3-8B	8B	37.55%	35.26%	33.10%	5.43%	29.46%	18.15%	34.52%	15.71%	23.71%
EduChat-R1-Qwen3-32B	32B	37.25%	36.97%	43.33%	5.23%	38.93%	17.95%	39.31%	15.70%	25.92%
Proprietary Models										
GPT-5.4	–	58.36%	65.74%	37.90%	51.71%	37.45%	28.20%	48.26%	44.92%	46.34%
GPT-5.6-Sol	–	62.30%	68.59%	54.81%	54.39%	37.09%	32.80%	54.66%	48.26%	50.97%
Claude-Opus-5	–	50.87%	59.68%	67.64%	55.70%	55.27%	45.09%	56.42%	52.22%	54.01%
GLM-5.3	–	63.20%	58.46%	75.32%	10.90%	58.78%	22.55%	**65.40%**	22.39%	38.62%
Kimi-K3	–	69.18%	75.00%	66.09%	68.17%	52.03%	53.76%	64.06%	**63.39%**	**63.65%**

Table 19: Detailed LongTutor results on evidence use, diagnosis, and teaching quality. Top-three values are shown in **red**, **orange**, and **yellow**, respectively.

Model	Size	Evidence				Diagnosis		Teaching					
		IE	MR	HC	Avg.	Acc	F1	R-L	Hist.	Strat.	Coher.	Appr.	Avg.
Open-weight Models													
Qwen3.5-4B-Base	4B	49.84%	14.42%	12.76%	25.67%	37.89%	24.71%	0.2953	1.36	1.54	2.19	2.86	1.60
OmniEdu-4B (ours)	4B	80.84%	27.45%	89.34%	65.88%	51.51%	29.08%	0.3376	1.73	2.23	2.83	2.68	2.29
Qwen3.5-9B-Base	9B	10.12%	0.48%	6.84%	5.81%	23.71%	17.96%	0.3140	1.28	1.44	1.86	2.18	1.48
OmniEdu-9B (ours)	9B	84.61%	28.58%	86.71%	66.63%	**53.55%**	26.13%	0.3469	2.04	2.60	3.18	3.00	2.66
Qwen3.8-27B	27B	34.31%	3.90%	72.17%	36.80%	**51.78%**	**38.41%**	0.3304	2.26	2.65	3.35	3.68	**2.74**
OmniEdu-27B (ours)	27B	92.87%	37.69%	68.24%	**78.20%**	**54.04%**	30.58%	0.3457	2.56	2.98	3.50	3.95	**3.02**
Confucius3-Math	14B	41.23%	11.95%	16.04%	23.07%	13.78%	10.16%	0.1653	1.25	1.34	1.72	1.98	1.36
MuduoLLM	14B	42.79%	13.94%	16.20%	24.31%	12.00%	6.46%	0.2352	1.19	1.27	1.65	2.09	1.28
EduChat-SFT-Qwen2.5-7B	7B	42.73%	5.65%	31.27%	26.55%	48.87%	19.69%	0.1028	1.14	1.18	1.66	2.05	1.21
EduChat-R1-Qwen3-8B	8B	64.91%	19.70%	16.63%	33.75%	18.41%	9.62%	0.0878	1.12	1.12	1.30	1.88	1.14
EduChat-R1-Qwen3-32B	32B	84.61%	26.48%	44.99%	52.03%	39.72%	24.47%	0.2480	1.44	1.50	2.09	2.82	1.56
Proprietary Models													
GPT-5.4	–	92.73%	59.89%	73.89%	75.50%	44.15%	30.94%	0.3525	1.37	1.58	2.08	1.98	1.62
GPT-5.6-Sol	–	94.05%	43.08%	95.35%	**77.50%**	45.09%	**38.07%**	0.3651	1.59	1.79	2.31	2.21	1.89
Claude-Opus-5	–	96.33%	68.16%	88.03%	**84.17%**	28.57%	27.13%	0.2691	2.03	2.69	3.12	2.78	**2.72**
GLM-5.3	–	94.09%	59.95%	78.23%	77.42%	40.59%	**34.50%**	0.3403	2.01	2.60	3.18	2.92	2.66
Kimi-K3	–	94.09%	52.96%	77.15%	74.73%	38.44%	30.97%	0.3520	1.75	2.07	2.63	2.52	2.17

Table 20: IFEval results for the base and education-tuned models. Prompt-level metrics require all verifiable instructions in a prompt to be satisfied, whereas instruction-level metrics average over individual instructions; strict and loose use the benchmark's two matching criteria (Zhou et al., 2023b). For each model scale, the better result between the base and SFT variants is shown in **bold**.

Model	Size	Prompt Strict	Inst. Strict	Prompt Loose	Inst. Loose
Qwen3.5-4B-Base	4B	64.88%	73.50%	67.84%	**76.50%**
OmniEdu-4B (ours)	4B	**65.06%**	**73.62%**	**68.21%**	76.38%
Qwen3.5-9B-Base	9B	69.69%	78.18%	74.31%	81.41%
OmniEdu-9B (ours)	9B	**73.01%**	**80.34%**	**75.05%**	**81.77%**
Qwen3.8-27B	27B	80.59%	86.33%	84.84%	**89.93%**
OmniEdu-27B (ours)	27B	**82.07%**	**87.89%**	**85.40%**	**89.93%**

Table 21: GPQA accuracy on the Main and Diamond subsets for the base and education-tuned models (Rein et al., 2023). For each model scale, the better result between the base and SFT variants is shown in **bold**.

Model	Size	Main Acc.	Diamond Acc.
Qwen3.5-4B-Base	4B	55.80%	58.08%
OmniEdu-4B (ours)	4B	**56.03%**	**60.61%**
Qwen3.5-9B-Base	9B	57.14%	62.12%
OmniEdu-9B (ours)	9B	**61.61%**	**63.64%**
Qwen3.8-27B	27B	70.76%	74.24%
OmniEdu-27B (ours)	27B	**72.32%**	**77.78%**

Table 22: MMMU-Pro accuracy across disciplines for the base and education-tuned models (Yue et al., 2024). The abbreviated discipline headers correspond to Art and Design, Business, Science, Health and Medicine, Humanities, and Technology and Engineering. For each model scale, the better result between the base and SFT variants is shown in **bold**.

Model	Size	Overall	Art	Business	Science	Health	Humanities	Tech.
Qwen3.5-4B-Base	4B	50.46%	59.21%	63.99%	49.14%	48.60%	56.75%	35.25%
OmniEdu-4B (ours)	4B	**52.60%**	55.26%	64.34%	51.55%	53.50%	56.31%	41.25%
Qwen3.5-9B-Base	9B	58.38%	58.77%	76.92%	55.33%	56.99%	61.71%	46.76%
OmniEdu-9B (ours)	9B	**60.75%**	56.58%	74.48%	61.51%	58.39%	66.21%	51.80%
Qwen3.8-27B	27B	64.97%	59.65%	82.52%	63.23%	62.59%	65.77%	58.27%
OmniEdu-27B (ours)	27B	**67.98%**	60.09%	84.27%	69.07%	65.04%	67.12%	62.83%