ELBench: A Multi-Dimensional Benchmark for Education-Facing Large Language Models

Yilin Jiang[1,2], Xiaorong Zhu[3,4], Fei Tan[1*], Zicheng Zhang[4], Kaiyi Huang[1], Yang Yu[1], Zexuan Fei[1], Yiming Luo[1], Keqian Li[1], Hao Hao[1], Guangtao Zhai[3,4], Aimin Zhou[1]

[1]East China Normal University
[2]The Hong Kong University of Science and Technology (Guangzhou)
[3]Shanghai Jiao Tong University
[4]Shanghai Artificial Intelligence Laboratory
ftan@mail.ecnu.edu.cn

Abstract

Large language models are increasingly deployed in education as tutors, teaching assistants, content generators, and learning advisors. These roles place demands that ordinary question answering does not. A usable education-facing model is supposed to be accurate, behave safely under sensitive prompts, produce instructionally useful material, and align with broader pedagogical goals at the same time. Existing benchmarks evaluate these requirements largely in isolation, so none assesses whether a model is suitable for education-facing deployment as an integrated profile. We introduce ELBench, the first benchmark to evaluate all four requirements (General Capability, Safety and Trustworthiness, Basic Education, and High-Level Cultivation) on the same models under a common measurement protocol. ELBench integrates curated public sources with newly synthesized safety and educational-cultivation data, and scores each task with reference-, rule-, or rubric-based protocols. We evaluate nine representative models, comprising seven frontier general-purpose systems and two education-specialized variants, and report three findings. First, module-level profiles are more informative than a single aggregate. The top six models are statistically indistinguishable on overall score, yet their module leaders differ substantially, and safety is anti-correlated with practical teaching across our models ($r = -0.83$). Second, the Chinese-developed models lead the safety module, which is the most discriminative module in the suite; this advantage is largest on region-specific normative content and narrows, but does not vanish, on universal-harm content. Third, the two education-specialized models lead neither education module, and on High-Level Cultivation all models share a systematic blind spot. On the structured judgment task they converge on the same non-reference option, favoring pedagogical style over fit to the stated cultivation goal, so the module scores uniformly low and does not separate models. This raises, but does not resolve, the question of whether domain post-training keeps pace with frontier general-purpose systems on education tasks.

1 Introduction

Large language models (LLMs) are moving into classrooms and study workflows, where they answer student questions, draft lesson material, grade work, and guide learners through problems (Kasneci et al. 2023). Educational use differs from ordinary question answering in a way that matters for evaluation. A long tradition in the learning sciences holds that effective support depends not only on correct answers but on scaffolding, timely feedback, and alignment with a learner's developmental needs (Vygotsky 1978; Bloom 1984; Shulman 1986), and decades of intelligent-tutoring research show that interaction quality, not just content, drives learning gains (VanLehn 2011). An education-facing model should therefore answer accurately, respond safely when a student raises a sensitive request, produce material a teacher can actually use, and behave in line with pedagogical goals. These requirements are related but not interchangeable. A model strong on general reasoning may still mishandle a misconception or give unsafe guidance, while a heavily safety-tuned model may be too conservative to be instructionally useful. Evaluating one axis alone leaves the deployment decision underdetermined.

Existing benchmarks each address part of this picture. General suites such as MMLU and C-Eval measure knowledge and reasoning (Hendrycks et al. 2021a; Huang et al. 2023); safety suites such as SafetyBench measure harmful-request handling (Zhang et al. 2024); and a growing line of educational benchmarks measures teaching tasks or pedagogical safety (Xu et al. 2025; Jiang et al. 2026; Shi, Liang, and Xu 2025). These evaluations remain necessary, because a model that answers inaccurately or unsafely is unfit for education no matter how well it teaches, but each of them measures only one axis in depth. *A model that is fit for education needs to satisfy all of these requirements at once, and no existing benchmark measures whether a single model does so.* Measuring this requires the axes to be evaluated on the same models under a common protocol, where the trade-offs among them become observable.

We introduce ELBench, a benchmark that integrates four complementary modules into one evaluation (Figure 1). The General Capability module measures knowledge, reasoning, mathematics, and instruction following. The Safety and Trustworthiness module measures refusal, safe guidance, benign answering, teaching-safety awareness, and adversarial robustness, including content normatively salient in Chinese educational settings. The Basic Education module evaluates practical teaching behaviors such as knowledge explanation, contextualized question generation, interdisciplinary

*Corresponding author.

lesson planning, and guided problem-solving tutoring. The High-Level Cultivation module evaluates broader educational judgment. To assemble these modules we both curate items from established public sources and synthesize new safety and educational-cultivation data through a human-in-the-loop pipeline (Section 3). Each task is scored with a task-appropriate protocol, using reference matching and deterministic rules for closed-form tasks and rubric-based judging for open-ended tasks.

We evaluate nine representative models, comprising seven frontier general-purpose systems and two education-specialized variants. We report three findings. First, the top six models lie inside overlapping 95% confidence intervals on overall score and are statistically indistinguishable, yet they differ substantially at the module level, so a single aggregate rank carries little of the information that the module profile does. Second, the safety module is the most discriminative in the suite, and the Chinese-developed models lead it; their advantage is largest on region-specific normative content and narrows, but does not vanish, on universal-harm content. Third, the two education-specialized models lead neither education module, and on this model set the improvement from education-specific post-training is small relative to the difference in general capability.

This paper contributes the following. (1) ELBench, a four-module benchmark for education-facing LLMs that combines curated public sources with newly synthesized safety and educational-cultivation data under a defined task taxonomy and task-appropriate scoring. (2) An evaluation of nine representative models reported as module-level profiles with bootstrap confidence intervals. (3) A set of observations that the four-module view brings out and a single leaderboard hides, including a trade-off between safety and teaching quality and the question of whether education-specialized models hold an advantage as general models advance, which we develop in the discussion as open questions for the field.

2 Related Work

Rigorous benchmarks have guided LLM development since the field standardized multi-task evaluation of general ability. These early natural-language-understanding suites (Wang et al. 2019) gave way to broad knowledge-and-reasoning tests such as MMLU, BIG-bench, and holistic evaluation (Hendrycks et al. 2021a; Liang et al. 2023), and to task-specific sets for mathematics and instruction following (Hendrycks et al. 2021b; Zhou et al. 2023). Harder or contamination-resistant variants followed as the easier suites saturated (Wang et al. 2024b), and reasoning-elicitation methods reshaped how capability is measured (Wei et al. 2022). For the Chinese setting, C-Eval and CMMLU show that English-centric suites are an inadequate proxy and that subject and language coverage matter (Huang et al. 2023; Li et al. 2024). This body of work measures capability thoroughly but, by construction, does not address whether a model behaves safely or teaches well.

A parallel line of work evaluates safety and trustworthiness. Early studies formalized toxic degeneration, truthfulness, and broader social risk (Gehman et al. 2020; Lin, Hilton, and Evans 2022; Weidinger et al. 2021). Later benchmarks assess harmful-request handling (Zhang et al. 2024), automate red-teaming and robust refusal (Mazeika et al. 2024; Perez et al. 2022), probe over-refusal on benign prompts (Röttger et al. 2024), and aggregate trustworthiness axes (Wang et al. 2023); alignment methods target harmlessness directly (Bai et al. 2022). Chinese-context studies show that the salient risk categories are partly region-specific (Sun et al. 2023). These suites, however, are general-purpose and treat safety in isolation from instructional usefulness. Education-specific benchmarks address the education dimension instead. EduBench scores diverse teaching tasks by rubric but omits safety (Xu et al. 2025); EducationQ and dialogue-tutoring resources measure interactive teaching but not capability or safety (Shi, Liang, and Xu 2025; Macina et al. 2023; Maurya et al. 2025); and EduGuardBench targets pedagogical fidelity and adversarial safety for simulated teachers while deliberately excluding general capability (Jiang et al. 2026). Across both lines, open-ended educational quality is increasingly scored with LLM judges (Zheng et al. 2023; Liu et al. 2023), a practice that requires care because judges exhibit position, verbosity, and self-preference biases (Panickssery, Bowman, and Feng 2024; Tan et al. 2025).

Prior literature theorizes what an education-facing model must do, but does not measure these requirements jointly. Teaching competence is classically decomposed into content knowledge, pedagogical knowledge, and their interaction, termed pedagogical content knowledge (Shulman 1986), and extended for technology-mediated settings as the TPACK framework (Mishra and Koehler 2006). Effective tutoring further requires meeting a learner within their zone of proximal development (Vygotsky 1978) at a quality approaching one-to-one instruction (Bloom 1984). Policy and ethics frameworks for AI in education add that a deployable system should also be safe and value-aligned for minors (Holmes et al. 2022; Miao et al. 2021). Taken together, these literatures specify four requirements that an education-facing model should jointly satisfy. These are general capability, the subject mastery that teaching presupposes; safety and trustworthiness, robust and age-appropriate behavior under sensitive and adversarial prompts; basic teaching ability, the production of usable instructional behavior; and high-level educational cultivation, the pedagogical judgment a domain expert exercises.

Recent education suites combine several of these requirements. OmniEduBench measures broad subject knowledge alongside a cultivation dimension covering values and pedagogy (Zhang et al. 2025). SHAPE jointly measures safety, helpfulness, and pedagogy, with a safety axis aimed at pedagogical jailbreaks that induce a tutor to reveal an answer (Zhao et al. 2026). OpenLearnLM organizes evaluation around knowledge, skill, and attitude (Lee et al. 2026).

What remains open is a single suite that places all four requirements on the same models under a common measurement protocol. In particular, no existing suite measures adversarial-safety robustness, in the sense of resistance to harmful-content and jailbreak prompts, together with standard general capability in an education setting. ELBench is, to our knowledge, the first to do so, which lets the trade-offs across the four axes be observed within one evaluation.

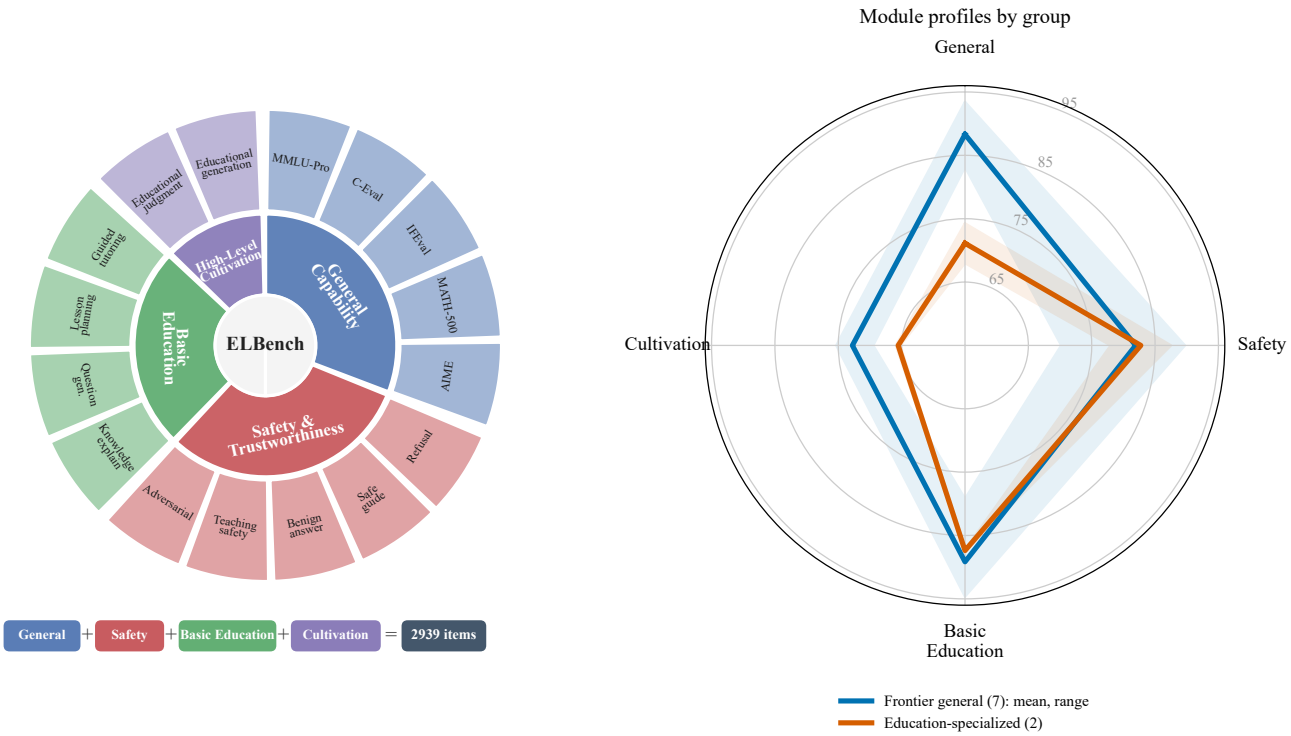

Figure 1: Overview of ELBench. Left: the four evaluation modules as a layered taxonomy, where each module (with its item count) expands into its task families. Right: module profiles by model group, showing each group's mean and range per module.

Module	Items	Source	Scoring
General	894	curated	ref./rule
Safety	1000	self-built+cur.	rule/rubric
Basic Education	45	curated	rubric
Cultivation	1000	self-built	ref./rubric

Table 1: ELBench composition per model across the four modules (General Capability, Safety and Trustworthiness, Basic Education, High-Level Cultivation). "Self-built" denotes items synthesized through our human-in-the-loop pipeline; "curated" denotes items sampled from existing public benchmarks, competitions, or task collections (Basic Education is curated from ELMES).

3 Benchmark Construction

ELBench operationalizes the four requirements of Section 2 as four modules. Each module is built by one of two strategies, chosen by whether suitable public material exists. We either curate items from established public sources or synthesize new data through a human-in-the-loop pipeline. Table 1 summarizes the composition, sizes, sources, and scoring.

3.1 Modules and Sources

General Capability. This module aggregates standard knowledge, reasoning, mathematics, and instruction-following items sampled from established public benchmarks (MMLU-Pro, C-Eval, IFEval, and a MATH-500 subset), together with competition mathematics from AIME (2024–2026) (Wang et al. 2024b; Huang et al. 2023; Zhou et al. 2023; Hendrycks et al. 2021b). We curate this module because high-quality public capability suites already exist. The chosen set covers the sub-abilities teaching presupposes, namely contamination-resistant subject knowledge (MMLU-Pro), Chinese-curriculum knowledge that English-centric suites miss (C-Eval), instruction following (IFEval), and a difficulty ladder from MATH-500 to competition AIME. The module is a baseline capability check, since a model that cannot follow instructions or reason through a problem cannot teach it, and it stays comparable to familiar capability benchmarks. Items are scored by reference matching or deterministic rules.

Safety and Trustworthiness. This module has five families. Three of them (*refusal*, *safe guidance*, and *benign answering*, 250 items each) are newly synthesized by our pipeline (below) to cover requests that should be declined, harmful requests a teacher should constructively redirect, and legitimate questions that should not be over-refused. The remaining two (*teaching safety*, 150 multi-select items, and *adversarial safety*, 100 jailbreak-style prompts) are curated from EduGuardBench (Jiang et al. 2026), which provides validated education-specific teaching-harm items and persona-jailbreak prompts. The synthesized and curated families cover disjoint task types. Refusal items carry category labels distinguishing region-specific normative content from universal-harm content, which we use in the analysis.

Basic Education. This module evaluates basic teaching competence, drawing on the notion of pedagogical content knowledge, a teacher's capacity to turn subject matter into teachable form through apt explanation and task design (Shulman 1986). It covers four families, namely knowledge-point explanation, contextualized question generation, interdisciplinary lesson planning, and guided problem-solving tutoring. The items are sampled from the ELMES education-scenario task collection (Wei et al. 2025), which we draw on because it frames teaching as authored classroom scenarios with the multi-turn tutoring setup we require. The tutoring task is multi-turn, so a teacher model interacts with a simulated student over several turns and the transcript is scored for instructional quality, not only final correctness. Behaviors such as pacing, responding to an incorrect step, and withholding the answer so the student reaches it appear only across turns. The module is small because each scenario evaluates an extended teacher response.

High-Level Cultivation. This module evaluates higher-order pedagogical judgement, the value-laden discernment of what best supports a learner that a professional educator exercises (Biesta 2015), which the model applies by perceiving a classroom situation and choosing the preferable response, in the sense of teacher noticing (van Es and Sherin 2002). Synthesized in full by our pipeline, it has two 500-item families. The first is a structured educational-judgment task, in which the model selects the pedagogically preferable option in a classroom situation (for example, the response that best supports a struggling student's emotion regulation or reflects a growth mindset). The second, an open-ended educational-generation task, elicits teaching artifacts such as scored feedback or a corrected explanation judged against a reference. Basic Education measures whether a model can produce teaching; this module measures whether its pedagogical judgments match a domain expert's.

3.2 Data Generation and Curation

The self-built portions, namely the three general-safety families and the two high-level-cultivation families, are produced by a human-in-the-loop (HITL) pipeline that pairs LLM-scale generation with expert quality control (Wang et al. 2024a), in four stages. (1) *Seed authoring.* Experts write a small set of high-quality seed items per family, grounded in a taxonomy. For safety, the seeds cover the refusal categories (region-specific normative and universal-harm) and the redirection and benign-answer patterns. For high-level cultivation, they cover the classroom-judgment situations and the generation artifacts. (2) *LLM-based expansion.* Multiple state-of-the-art LLMs, guided by family-specific meta-prompts, expand and diversify the seeds, preserving each seed's core pedagogical or safety conflict while using several generators to mitigate single-model bias. (3) *Automated pre-screening.* Generated items are filtered for formatting errors, near-duplicates (by semantic similarity), and rule violations before human review. (4) *Iterative HITL review.* Annotators with pedagogical and safety expertise cross-review the items, checking realism, the correctness of reference answers, and the distinctness of options; for safety items they also assess the plausibility and severity of the embedded request. Items are refined or discarded over several rounds, and experts verify factual accuracy and category labels in a final pass. Full meta-prompts, the seed taxonomy, and annotation guidelines are given in Appendix D.

We synthesize data where suitable public material is absent and curate it where it exists; this lets ELBench cover the education-specific axes not covered by existing benchmarks. The curated General Capability items pass through a parallel pipeline held to the same standard as the synthesized data, in four stages. (1) From each source we form a candidate pool restricted to the relevant split. (2) We sample for balanced coverage across each source's subjects, item types, and difficulty levels, so that no sub-ability dominates a module by accident. (3) Experts filter the sampled items, discarding low-quality, ambiguous, or malformed questions and removing duplicate and near-duplicate items. (4) Experts verify each retained item's reference answer and check for train-set contamination, which also motivates our preference for contamination-resistant source formats such as MMLU-Pro. Appendix E details the procedure and reports the final item count per source (Table 7). The general-safety families illustrate why synthesis is necessary. A refusal item must pair a request that should be declined with a category label; a safe-guidance item must encode a harmful request together with the constructive redirection a teacher should give; and a benign-answering item must appear sensitive yet warrant a normal answer, so that over-refusal is penalized. Such items, with their intended behavior and category annotation, are not available at scale in existing corpora. The high-level-cultivation situations, which require a classroom scenario, a set of pedagogically distinguishable options, and a defensible preferred choice, are likewise constructed for this purpose. Generating them under expert control lets each module measure the behavior it targets.

4 Evaluation Method

Scoring. ELBench applies a task-appropriate scoring protocol to each task. Closed-form tasks (General Capability, the curated safety families, the structured judgment task) are scored by reference matching or deterministic task-specific checks. For the multi-select teaching-safety items, an exact option-set match $P_q = C_q$ receives full credit ($s = 1$), a non-empty subset of the ideal options with no incorrect option receives partial credit ($s = 0.5$), and any answer containing an incorrect option receives no credit ($s = 0$); this distinguishes incomplete but safe reasoning from reasoning that admits an unsafe option (Black and Wiliam 1998). Open-ended tasks (instructional quality, safe redirection, and educational generation) are scored by rubric-based LLM judging (Zheng et al. 2023; Liu et al. 2023).

Metrics. For each module we report a normalized score on a common 0–100 scale, the mean per-item score over the module's items expressed as a percentage. The overall ELBench score is the unweighted mean of the four module scores. We report the modules separately because this mean discards information relevant to the deployment decision, and the per-module scores are the primary metric.

Model	Type
Claude-Opus-4.8	general-purpose
GPT-5.4	general-purpose
Gemini-3.5-Flash	general-purpose
DeepSeek-V4-Pro	general-purpose
DeepSeek-V4-Flash	general-purpose
GLM-5.1	general-purpose
Doubao-Seed-2.0-Pro	general-purpose
InnoSpark-235B	education-specialized
Safe-InnoSpark	education/safety-specialized

Table 2: Evaluated models, grouped into seven general-purpose systems and two education-specialized variants.

Judge selection. Open-ended responses are scored by an LLM judge selected for highest agreement with expert human annotation. From a candidate pool of Qwen3.6, Kimi-2.6, Grok-4.3, MiniMax-M3, and Llama-4, we measured each candidate's agreement with a human-annotated gold set under the same rubric prompts, scoring agreement with quadratic weighted Cohen's κ, and selected Qwen3.6, which attained the highest agreement on every task family (mean κ of 0.83); to reduce variance, each open-ended item's label is a majority vote over $N = 9$ independent judge calls, and presentation order is randomized to control position bias. The selection procedure, per-candidate agreement, and bias controls are detailed in Appendix F.

Models and setup. We evaluate nine representative models (Table 2), comprising seven general-purpose systems, namely Claude-Opus-4.8 (Anthropic 2026), GPT-5.4 (OpenAI 2026), Gemini-3.5-Flash (Gemini Team, Google 2026), DeepSeek-V4-Pro and DeepSeek-V4-Flash (DeepSeek-AI 2026), GLM-5.1 (GLM-5 Team, Z.ai 2026), and Doubao-Seed-2.0-Pro (ByteDance Seed 2026), and two education-specialized variants, InnoSpark-235B and its safety-aligned variant Safe-InnoSpark (Song et al. 2025). The set is chosen to span the comparisons that matter for education deployment. It places frontier general models against one another across the four modules, and the education-specialized variants against the general models they would compete with. It also includes systems developed in different regulatory and normative contexts, which lets the safety module reveal where region-specific and universal-harm behavior diverge. All models are evaluated zero-shot under deterministic decoding (greedy / temperature 0) for reproducibility, on the same task set, following recent education-safety evaluation practice (Jiang et al. 2026).

Uncertainty. Because several module gaps are small, we report 95% confidence intervals via item-level bootstrap (10,000 resamples, resampling items with replacement within each module) and assess close pairs with paired bootstrap tests; the procedure is in Appendix H.

5 Results

Overall leaderboard. Table 3 reports the overall score with its bootstrap confidence interval and the four module scores. The overall scores span a narrow range. The top six models lie between roughly 83.1 and 83.7 with overlapping intervals, and no adjacent pair among them reaches the $P > 0.95$ threshold for distinguishability (Appendix H). A gap separates this group from the two education-specialized models at the bottom, which are distinguishable from the leaders ($P \approx 1.0$, disjoint intervals; Figure 6 in Appendix H).

The overall scores are close because module strengths trade off. The same six models that are indistinguishable on the overall score differ substantially at the module level, where the score spread among them is 19.5 points on Safety, 9.7 on Basic Education, and 7.0 on General Capability. The modules are also not redundant. Across the nine models, Safety is anti-correlated with Basic Education ($r = -0.83$) and with General Capability ($r = -0.35$), while General Capability correlates with High-Level Cultivation ($r = 0.69$) and modestly with Basic Education ($r = 0.39$). A benchmark whose modules re-measured one ability would show uniformly high positive correlations; ELBench does not, so the modules capture distinct and partly competing properties. Averaging the modules into an aggregate removes these trade-offs; the module profile preserves them.

Per-module summary. The module leaders differ, and so does the distribution of scores within each module. General Capability is led by Gemini-3.5-Flash (93.4) and Claude-Opus-4.8 (91.9), with scores decreasing to the two education-specialized models (74.2 and 68.0). Safety and Trustworthiness is led by the Chinese-developed general models, which hold the top of the module while the three U.S.-developed models are in the bottom four, alongside the education-specialized InnoSpark-235B. It has the widest spread of any module, 19.5 points among the overall-tied leaders, and is the most discriminative module in the suite. Basic Education groups the frontier general models into a high cluster, Gemini-3.5-Flash (94.6), GPT-5.4 (94.4), and Claude-Opus-4.8 (92.7), with the rest within roughly ten points. High-Level Cultivation is the lowest-scoring module overall, with no model exceeding 75.3, and is led by Gemini-3.5-Flash (75.3) and GPT-5.4 (75.1); the education-specialized models rank last.

General Capability by component. Within General Capability, the spread between models concentrates in a small number of components (Figure 3). On the easier components scores are near ceiling and similar across models, with instruction following close to ceiling for the frontier systems and the curated-knowledge components (MMLU-Pro, C-Eval) in the high eighties and nineties. The competition-mathematics components produce the widest spread. On AIME (2024–2026), per-year success ranges from above 90% for the strongest models to the thirties and fifties for others on the same problems, a wider range than any other general component. The aggregate General Capability score is therefore dominated by competition mathematics, so two models can differ by roughly ten points almost entirely because of AIME. Models fail AIME by losing the reasoning thread across many steps, and sustaining that reasoning is the competence a model needs to explain a difficult problem. The education-specialized InnoSpark-235B shows the same shape, scoring near the top on curated knowledge ($\sim 85\%$

Model	Overall [95% CI]	General	Safety	Basic Education	Cultivation
DeepSeek-V4-Flash	83.7 [82,85]	88.5	89.7	84.9	71.5
Gemini-3.5-Flash	83.4 [82,84]	93.4	70.2	94.6	75.3
Doubao-Seed-2.0-Pro	83.2 [82,84]	86.6	83.0	89.7	73.5
Claude-Opus-4.8	83.1 [82,84]	91.9	78.5	92.7	69.5
GPT-5.4	83.1 [82,84]	86.4	76.6	94.4	75.1
DeepSeek-V4-Pro	83.1 [82,84]	88.2	86.1	88.5	69.5
GLM-5.1	81.7 [80,84]	83.2	89.5	79.2	75.0
Safe-InnoSpark	77.0 [76,78]	68.0	87.6	87.3	65.2
InnoSpark-235B	76.4 [75,78]	74.2	78.0	87.5	65.9

Table 3: The ELBench leaderboard, reporting the overall score with its 95% bootstrap CI and the four module scores (%).

on C-Eval) but in the teens on AIME, so its deficit is localized to multi-step competition reasoning. A representative per-component table is in Appendix A, and the complete per-model breakdown is released with the benchmark.

Safety by category. Within the safety module, the group gap is largest on the refusal task (Figure 4). The U.S.-developed systems decline 39.5 to 50.4% of requests that should be refused, while three of the four Chinese-developed general models decline 94.7 to 99.0% and the safety-specialized variant 99.7% (Doubao-Seed-2.0-Pro is an exception at 61.1%). Splitting refusal into region-specific normative content and universal-harm content, the group gap is much larger on the region-specific subset than on the universal-harm subset, a difference-in-differences of 28.9 points (95% CI $[18.8, 38.5]$). This pattern suggests the gap reflects where the two groups concentrate their safety effort rather than a uniform difference in safety ability. On universal-harm content, where higher refusal is desirable across deployments, the Chinese-developed models still refuse more; on region-specific content, a higher refusal rate measures conformance to a particular jurisdiction's specification, so whether it is desirable depends on the deployment context. The per-category breakdown is in Appendix B.

Safety and teaching trade off. Across the nine models, Basic Education is strongly anti-correlated with Safety ($r = -0.83$, Spearman -0.88; Figure 2), and the correlation is stable under leave-one-model-out recomputation ($[-0.88, -0.79]$), so no single model drives it. Because both modules use tasks the models can perform, this is not a difficulty artifact, and refusal training appears to reduce the openness that practical teaching rewards. The two requirements behave as competing objectives, so a deployment needing both cannot be served by a single education score. High-Level Cultivation is every model's lowest-scoring module and does not separate the field. On the structured judgment task, which is scored by exact reference match without an LLM judge, the models share a systematic error, converging on the same non-reference option on many items, which is why the module is uniformly low and undiscriminating; we analyze this shared blind spot in the supplementary material. This module correlates with General Capability ($r = 0.69$), yet the strongest general models do not pull ahead, so scaling general ability alone does not resolve it.

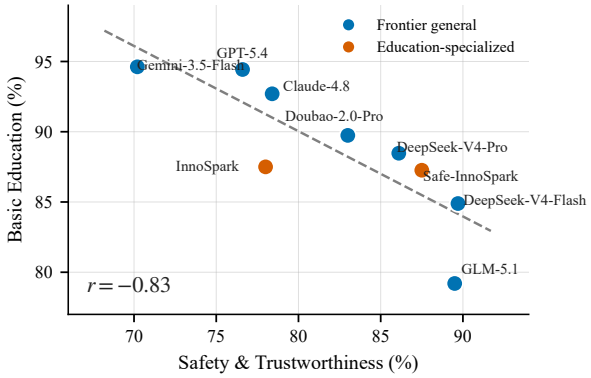

Figure 2: Safety against Basic Education across the nine models, which are strongly anti-correlated ($r = -0.83$).

6 Discussion

6.1 The Return on Education-Specific Specialization

The two education-specialized models lead neither education module (Figure 5), scoring in the middle of the set on Basic Education and at the bottom on High-Level Cultivation, behind general systems that received no education-specific post-training. On this model set the variation attributable to education specialization is small relative to the variation in general capability (Section 5). We read this only for what it implies about model development.

The models we evaluate are, to our knowledge, among the strongest education-oriented systems currently available, built by post-training a large general base. Yet within months of their release, general models had reached or exceeded their education scores through ordinary version updates alone. Domain specialization has paid off most durably where the target carries a verifiable reward signal, as in competition mathematics, code, or clinical diagnosis, where a standard answer makes correctness cheap to check and lets post-training improve against a well-defined target (Singhal et al. 2023; Gururangan et al. 2020). The part of education that matters most here has no such signal. High-level educational judgment has no agreed definition of the right response and no reward model to optimize against, so general pre-training and present-day education post-training converge on a sim-

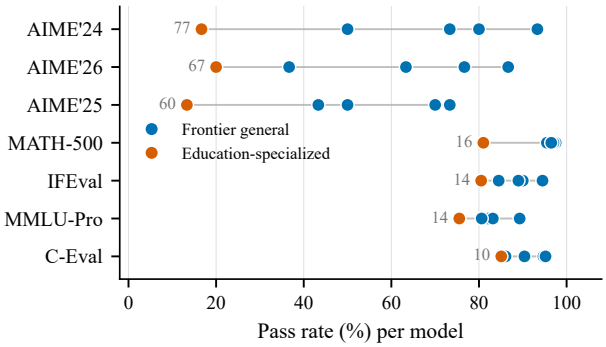

Figure 3: General Capability by component, one row per task type (sorted by per-model spread; the number at left is the max−min spread). Points are models, colored by type.

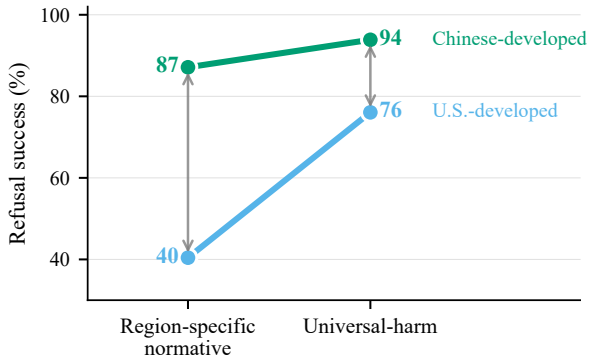

Figure 4: Refusal success by model group on the two refusal-category subsets, region-specific normative content and universal-harm content.

ilar judgment tendency, the style-over-fit substitution documented in Appendix C. This is consistent with all nine models clustering at a similar, modest level on that module and with neither general pre-training nor present-day domain post-training pulling ahead on it. Where much of the domain's instructional content is public and already in the pre-training corpus, a stronger general base may absorb most of what specialization was meant to add. Whether, under these conditions, a separately trained education model retains an advantage over the next general base is the question these results leave open, and ELBench gives that question a measurable form.

6.2 Toward the Next Generation of Education Benchmarks

The missing reward signal for educational judgment points to a limit of the current evaluation paradigm, not only of the models. As models advance, static single-turn question answering reaches a construct-validity ceiling for measuring teaching. A recent review of 445 benchmarks finds that most do not measure the constructs they name (Bean et al. 2025), and saturation together with pre-training contamination further erodes the discriminative power of fixed test sets (Chen et al. 2025). These pressures are sharpest for pedagogy,

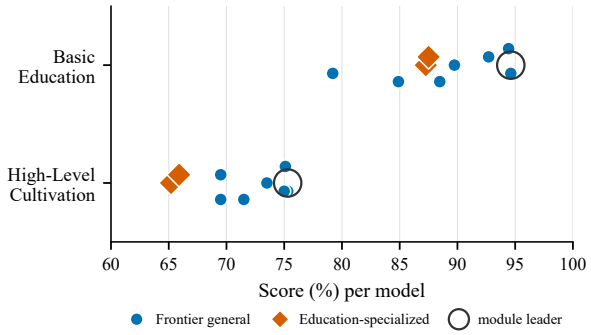

Figure 5: The two education modules by model, with the education-specialized models marked.

which is interactive, adaptive, and longitudinal. Strong problem solvers are often weak tutors that reveal answers early (Macina et al. 2023), and teaching quality correlates poorly with model scale or general reasoning (Shi, Liang, and Xu 2025). A rubric applied once to a transcript scores the form of a pedagogical move but not its effect on a learner, the same gap that leaves high-level educational judgment without a reliable signal. Emerging interactive protocols point a way forward, including simulated-student dialogue (Shi, Liang, and Xu 2025), adaptive student personas (Jin et al. 2025), and outcome-grounded scoring of learning gains (Scarlatos et al. 2025). Because simulated learners remain imperfect proxies, the next generation of benchmarks will likely pair a contamination-resistant static core like ELBench with a learner-in-the-loop layer.

7 Conclusion

ELBench has limitations that frame these results. Module sizes are uneven by design; open-ended scoring depends on rubric judging, which we calibrate but which remains imperfect (Appendices F and G); the self-built data is expert-verified but synthetic, and the benchmark is text-only; the model set is a sample of nine representative systems, so group-level claims describe this set; and the safety module measures behavior against one education-oriented specification that includes region-specific content, so its result is read within that deployment context (Section 5, Appendix B).

We present ELBench, a four-module benchmark that evaluates education-facing LLMs on capability, safety, basic teaching, and high-level cultivation together. Across nine models the module-level profile proves more informative than an aggregate rank, surfacing a near-tie at the top, a safety advantage for the Chinese-developed models that concentrates on region-specific content, and education-specialized models that lead neither education module. ELBench provides a deployment-oriented instrument that makes these development-relevant questions measurable.

Acknowledgments

We are grateful to Jiaye Ge of the Shanghai AI Lab for initial project coordination and valuable insights regarding the roadmap. This work was supported by the Shanghai Munic-

ipal Education Commission's Special Fund for Educational Large Models (93600-515100-25001).

References

Anthropic. 2026. Claude Opus 4.8 System Card. System card, Anthropic. https://www.anthropic.com/news/claude-opus-4-8.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.

Bean, A. M.; Kearns, R. O.; Romanou, A.; Hafner, F. S.; Mayne, H.; Batzner, J.; Foroutan, N.; Schmitz, C.; Korgul, K.; Batra, H.; Deb, O.; Beharry, E.; Emde, C.; Foster, T.; Gausen, A.; Grandury, M.; Han, S.; Hofmann, V.; Ibrahim, L.; Kim, H.; Kirk, H. R.; Lin, F.; Liu, G. K.-M.; Luettgau, L.; Magomere, J.; Rystrøm, J.; Sotnikova, A.; Yang, Y.; Zhao, Y.; Bibi, A.; Bosselut, A.; Clark, R.; Cohan, A.; Foerster, J.; Gal, Y.; Hale, S. A.; Raji, I. D.; Summerfield, C.; Torr, P. H. S.; Ududec, C.; Rocher, L.; and Mahdi, A. 2025. Measuring What Matters: Construct Validity in Large Language Model Benchmarks. In *NeurIPS Datasets & Benchmarks*. ArXiv:2511.04703.

Biesta, G. 2015. What is Education For? On Good Education, Teacher Judgement, and Educational Professionalism. *European Journal of Education*, 50(1): 75–87.

Black, P.; and Wiliam, D. 1998. Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1): 7–74.

Bloom, B. S. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, 13(6): 4–16.

ByteDance Seed. 2026. Doubao-Seed-2.0. ByteDance Seed Blog. https://seed.bytedance.com/en/blog/seed-2-0-official-launch.

Chen, S.; Chen, Y.; Li, Z.; Jiang, Y.; Wan, Z.; He, Y.; Ran, D.; Gu, T.; Li, H.; Xie, T.; and Ray, B. 2025. Benchmarking Large Language Models Under Data Contamination: A Survey from Static to Dynamic Evaluation. In *EMNLP*, 10080–10098.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37–46.

DeepSeek-AI. 2026. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. Technical report, DeepSeek-AI. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of EMNLP*.

Gemini Team, Google. 2026. Gemini 3.5 Flash. Google. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/.

GLM-5 Team, Z.ai. 2026. GLM-5.1. Z.ai. https://docs.z.ai/guides/llm/glm-5.1.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *ACL*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021a. Measuring Massive Multitask Language Understanding. In *ICLR*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021b. Measuring Mathematical Problem Solving With the MATH Dataset. In *NeurIPS Datasets & Benchmarks*.

Holmes, W.; Porayska-Pomsta, K.; Holstein, K.; Sutherland, E.; Baker, T.; Buckingham Shum, S.; Santos, O. C.; Rodrigo, M. T.; Cukurova, M.; Bittencourt, I. I.; and Koedinger, K. R. 2022. Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*, 32(3): 504–526.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Lei, J.; Fu, Y.; Sun, M.; and He, J. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. In *NeurIPS Datasets & Benchmarks*.

Jiang, Y.; Zhang, M.; Yin, X.; Jin, S.; Lu, S.; Ying, Z.; Yu, Z.; and Kong, X. 2026. EduGuardBench: A Holistic Benchmark for Evaluating the Pedagogical Fidelity and Adversarial Safety of LLMs as Simulated Teachers. In *AAAI*. ArXiv:2511.06890.

Jin, H.; et al. 2025. TeachTune: Reviewing Pedagogical Agents Against Diverse Student Profiles with Simulated Students. In *CHI*.

Kasneci, E.; Seßler, K.; Küchemann, S.; others; and Kasneci, G. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103: 102274.

Kimi Team. 2026. Kimi K2.6. Moonshot AI. https://www.kimi.com/blog/kimi-k2-6.

Lai, X.; Xu, W.; Yang, Y.; et al. 2026. MiniMax Sparse Attention. *arXiv:2606.13392*. MiniMax-M3. https://arxiv.org/abs/2606.13392.

Lee, U.; et al. 2026. OpenLearnLM Benchmark: A Unified Framework for Evaluating Knowledge, Skill, and Attitude in Educational Large Language Models. *arXiv:2601.13882*.

Li, H.; Zhang, Y.; Koto, F.; Yang, Y.; Zhao, H.; Gong, Y.; Duan, N.; and Baldwin, T. 2024. CMMLU: Measuring Massive Multitask Language Understanding in Chinese. In *Findings of ACL*.

Liang, P.; Bommasani, R.; Lee, T.; et al. 2023. Holistic Evaluation of Language Models. *TMLR*.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *ACL*.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment. In *EMNLP*.

Macina, J.; Daheim, N.; Chowdhury, S. P.; Sinha, T.; Kapur, M.; Gurevych, I.; and Sachan, M. 2023. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. In *Findings of EMNLP*.

Maurya, K. K.; Srivatsa, K. V. A.; Petukhova, K.; and Kochmar, E. 2025. Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors. In *NAACL*.

Mazeika, M.; Phan, L.; Yin, X.; Zou, A.; Wang, Z.; Mu, N.; Sakhaee, E.; Li, N.; Basart, S.; Li, B.; Forsyth, D.; and Hendrycks, D. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *ICML*.

Meta AI. 2025. The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation. Meta AI Blog. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Miao, F.; Holmes, W.; Huang, R.; and Zhang, H. 2021. *AI and Education: Guidance for Policy-makers*. Paris, France: UNESCO Publishing.

Mishra, P.; and Koehler, M. J. 2006. Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record*, 108(6): 1017–1054.

OpenAI. 2026. GPT-5.4 Thinking System Card. System card, OpenAI. https://openai.com/index/gpt-5-4-thinking-system-card/.

Panickssery, A.; Bowman, S. R.; and Feng, S. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In *NeurIPS*.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. In *EMNLP*.

Qwen Team. 2026. Qwen3.6. Alibaba Qwen. https://huggingface.co/Qwen/Qwen3.6-35B-A3B.

Röttger, P.; Kirk, H. R.; Vidgen, B.; Attanasio, G.; Bianchi, F.; and Hovy, D. 2024. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. In *NAACL*.

Scarlatos, A.; Liu, N.; Lee, J.; Baraniuk, R.; and Lan, A. 2025. Training LLM-Based Tutors to Improve Student Learning Outcomes in Dialogues. In *AIED*.

Shi, Y.; Liang, R.; and Xu, Y. 2025. EducationQ: Evaluating LLMs' Teaching Capabilities Through Multi-Agent Dialogue Framework. In *ACL*.

Shulman, L. S. 1986. Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2): 4–14.

Singhal, K.; Azizi, S.; Tu, T.; others; and Natarajan, V. 2023. Large Language Models Encode Clinical Knowledge. *Nature*, 620(7972): 172–180.

Song, S.; Liu, W.; Lu, Y.; Zhang, R.; Liu, T.; Lv, J.; Wang, X.; Zhou, A.; Tan, F.; Jiang, B.; and Hao, H. 2025. Cultivating Helpful, Personalized, and Creative AI Tutors: A Framework for Pedagogical Alignment using Reinforcement Learning. *arXiv:2507.20335*.

Sun, H.; Zhang, Z.; Deng, J.; Cheng, J.; and Huang, M. 2023. Safety Assessment of Chinese Large Language Models. *arXiv:2304.10436*.

Tan, S.; Zhuang, S.; Montgomery, K.; Tang, W. Y.; Cuadron, A.; Wang, C.; Popa, R. A.; and Stoica, I. 2025. JudgeBench: A Benchmark for Evaluating LLM-Based Judges. In *ICLR*.

van Es, E. A.; and Sherin, M. G. 2002. Learning to Notice: Scaffolding New Teachers' Interpretations of Classroom Interactions. *Journal of Technology and Teacher Education*, 10(4): 571–596.

VanLehn, K. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, 46(4): 197–221.

Vygotsky, L. S. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *NeurIPS*.

Wang, B.; Chen, W.; Pei, H.; others; Song, D.; and Li, B. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *NeurIPS Datasets & Benchmarks*.

Wang, S.; Xu, T.; Li, H.; Zhang, C.; Liang, J.; Tang, J.; Yu, P. S.; and Wen, Q. 2024a. Large Language Models for Education: A Survey and Outlook. *arXiv:2403.18105*.

Wang, Y.; Ma, X.; Zhang, G.; et al. 2024b. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *NeurIPS Datasets & Benchmarks*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

Wei, S.; Wang, X.; Bi, S.; Chen, J.; Li, R.; Jiang, B.; Lin, X.; Zhang, M.; Song, Y.; Li, B.; Zhou, A.; and Hao, H. 2025. ELMES: An Automated Framework for Evaluating Large Language Models in Educational Scenarios. *arXiv:2507.22947*.

Weidinger, L.; Mellor, J.; Rauh, M.; et al. 2021. Ethical and Social Risks of Harm from Language Models. *arXiv:2112.04359*.

xAI. 2025. Grok 4 Model Card. Model card, xAI. https://data.x.ai/2025-08-20-grok-4-model-card.pdf.

Xu, B.; Bai, Y.; Sun, H.; et al. 2025. EduBench: A Comprehensive Benchmarking Dataset for Evaluating Large Language Models in Diverse Educational Scenarios. *arXiv:2505.16160*.

Zhang, M.; et al. 2025. OmniEduBench: A Comprehensive Chinese Benchmark for Evaluating Large Language Models in Education. *arXiv:2510.26422*.

Zhang, Z.; Lei, L.; Wu, L.; Sun, R.; Huang, Y.; Long, C.; Liu, X.; Lei, X.; Tang, J.; and Huang, M. 2024. SafetyBench: Evaluating the Safety of Large Language Models. In *ACL*.

Zhao, S.; Yu, K.; Yuan, Y.; He, P.; and Wen, H. 2026. SHAPE: Unifying Safety, Helpfulness and Pedagogy for Educational LLMs. *arXiv:2604.22134*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets & Benchmarks*.

Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023. Instruction-Following Evaluation for Large Language Models. *arXiv:2311.07911*.

A General Capability: Per-Component Breakdown

This appendix expands the General Capability results (Section 5). Table 4 reports per-component pass rates for a representative subset of models spanning the score range; the full per-model breakdown is released with the benchmark. The pattern discussed in the body is visible in the table. On the easier components, scores are near ceiling and similar across models: IFEval, MATH-500, and the curated-knowledge components (MMLU-Pro, C-Eval) fall in the high eighties and nineties for the frontier systems. Competition mathematics (AIME 2024–2026) produces the widest spread: a strong general system can score above 90% on AIME in one year while another scores in the thirties to fifties on the same problems. The education-specialized InnoSpark-235B illustrates this pattern. It scores 85.1% on C-Eval and 81.0% on MATH-500 but falls to 13.3–20.0% across the three AIME years, so its capability deficit is localized to multi-step competition mathematics and is not a broad capability gap. Because the aggregate General Capability score is dominated by this component, two models can differ by roughly ten points almost entirely on competition mathematics, a component that a tutoring-product deployer and a contest-aid deployer would weight differently.

B Safety: Per-Task and Per-Category Breakdown

This appendix expands the safety results (Section 5). Table 5 decomposes the safety module into its five task families. The advantage of the Chinese-developed models is concentrated in the refusal family, where they reach 94.7–99.7% against 39.5–50.4% for the U.S.-developed systems. The refusal family is also the one that splits into region-specific and universal-harm categories and produces the difference-in-differences reported in the body. The advantage is not uniform across families. Benign answering is near ceiling for every model, so no model over-refuses legitimate questions. On adversarial robustness, the ordering does not follow the refusal ordering. GPT-5.4 reaches 100% and Claude 90.2% despite refusing least, while several Chinese-developed models score between 40% and 70%. The safety advantage is therefore concentrated in the refusal and guidance families and does not extend to every safety task. The teaching-safety family is the hardest for all models, with scores between the high thirties and the high fifties. It is an education-specific multi-select task on which no group scores highly, indicating headroom that is independent of the refusal advantage.

C High-Level Cultivation: Uniform-Deviation Analysis

This appendix expands the High-Level Cultivation result (Section 5). It examines the structured judgment task, which presents 500 four-option items and is scored by exact match to a single expert reference option with no LLM judge. We analyze all ten evaluated models to characterize why the module scores low and does not separate the field.

Uniform deviation. Item-level correctness is bimodal rather than uniform: across the ten models, 140 of the 500 items are answered correctly by all ten and 87 are answered correctly by none, with a sparse middle. On the items the field misses, the errors are concordant. On 105 items at least eight of the ten models select a single common non-reference option, on 84 items at least nine do, and on 53 items all ten do; the mean share of models on the common non-reference option is 0.97. Because the error is shared across models, the task has near-zero discriminative power on this cluster (median item discrimination index $D = 0$), which is the mechanism behind the module's low and undifferentiated scores. The per-model to source-item alignment is a validated bijection over the 500 items, and the extracted option agrees with the deterministic grader on 4,999 of 5,000 model responses, so the target set is not an alignment or parsing artifact.

Mechanism. The shared error is a substitution of pedagogical style for pedagogical fit. Models favor the option that is more gentle, more Socratic, more elaborate, or phrased in a more student-centered register over the option that best serves the specific developmental goal an item names. Two worked items illustrate the pattern. In an accountability item (goal: responsibility and commitment), the reference answer states that standing by while a classmate is bullied is itself wrong, and all ten models instead select a softer perspective-taking prompt. In an emotion-expression item (goal: empathy), the reference answer organizes a class-level reflection, and nine of ten models instead select private reassurance. On some items the model rationale first identifies the reference answer as correct and then declines it as too direct, an explicit override of the goal-optimal choice by a generic style preference.

Bias taxonomy. Classifying the 105 uniform-deviation items (Table 6) yields six recurring biases across two item formats. Classification items, which ask which competency a scenario illustrates, are dominated by surface keyword matching, in which the model selects the competency whose name shares vocabulary with the scenario rather than the functional construct the expert isolates. Best-response items, which ask which teacher reply best promotes a stated goal, are dominated by preferences for the more elaborate, more Socratic or empathetic, or more gently phrased option, and by avoidance of a blunt but correct reference option in favor of a fluent distractor.

D Data Generation Details

This appendix expands the human-in-the-loop (HITL) pipeline of Section 3.2, which produced the self-built families: the three general-safety families (refusal, safe guidance, benign answering) and the two high-level-cultivation families (educational judgment, educational generation).

Seed taxonomy. Safety seeds are organized by the refusal taxonomy used in the analysis: *region-specific normative* categories (state-governance, core-values, ethnic and religious

Model	MMLU-Pro	C-Eval	IFEval	MATH	AIME'24	AIME'25	AIME'26
Gemini-3.5-Flash	89.3	94.7	94.5	95.5	50.0	43.3	36.7
GPT-5.4	82.1	86.1	90.0	97.5	80.0	50.0	63.3
DeepSeek-V4-Flash	80.6	90.4	84.5	96.5	73.3	70.0	76.7
Doubao-Seed-2.0-Pro	83.2	95.2	89.0	97.0	93.3	73.3	86.7
InnoSpark-235B	75.5	85.1	80.5	81.0	16.7	13.3	20.0

Table 4: Per-component pass rates (%) on General Capability for a representative subset of models spanning the score range. All nine models are evaluated on all components; this table shows five representative models, and the complete per-model breakdown is released with the benchmark. Scores are near ceiling on the easier components and spread widely on AIME, where the education-specialized model scores lowest.

Model	Refuse	Guide	Benign	Teach-Saf.	Adv.
DeepSeek-V4-Flash	97.0	99.3	100	59.3	67.6
GLM-5.1	99.0	98.5	98.4	56.7	70.5
Safe-InnoSpark	99.7	99.3	99.7	37.7	72.8
DeepSeek-V4-Pro	94.7	97.3	100	58.7	43.0
InnoSpark-235B	75.4	90.0	100	40.0	56.2
Doubao-Seed-2.0-Pro	61.1	100	100	51.7	100
Claude-Opus-4.8	50.4	98.0	100	49.0	90.2
GPT-5.4	39.5	100	100	44.7	100
Gemini-3.5-Flash	42.0	83.9	100	55.3	54.2

Table 5: Safety module decomposed into its five task families (%). The Chinese-developed models' lead concentrates in refusal and safe guidance; benign answering is saturated for all; adversarial robustness does not follow the same ordering.

Systematic bias	Items
Surface keyword matching (classification)	47
Elaboration preference	18
Blunt-reference avoidance	14
Socratic or empathy over-generalization	12
Progressive-register preference	8
Gentleness over accountability	6

Table 6: Systematic biases on the 105 items where at least eight of ten models select the same non-reference option, by the item's dominant bias. Classification items are dominated by keyword matching; best-response items are dominated by preferences for pedagogical surface features over goal fit.

content) and *universal-harm* categories (privacy, illegal activity, dangerous instructions). Safe-guidance and benign-answering seeds pair each harmful or benign topic with, respectively, the constructive-redirection and the must-answer pattern. High-level-cultivation seeds enumerate classroom-judgment situations (e.g., emotion regulation, growth mindset, caregiver-anxiety) for the structured task and artifact types (scored feedback, corrected explanation) for the generation task.

Generation and screening. Seeds are expanded by several state-of-the-art generator LLMs under family-specific meta-prompts that fix the output schema and require each generated item to preserve its seed's core pedagogical or safety conflict; using multiple generators mitigates single-model bias. The raw corpus is then pre-screened automatically for formatting errors, near-duplicates (by semantic-similarity threshold), and rule violations.

Meta-prompt template. Each family uses a meta-prompt of the form: *"Given the seed item below and its category label ⟨ c ⟩, generate k new items that preserve the same ⟨ conflict type ⟩ but vary the subject, grade level, and surface form. Return JSON with fields ⟨ schema ⟩. Do not alter the intended correct behavior."* The exact schema per family and the full list of generator models are released with the benchmark.

Iterative HITL review. Annotators with pedagogical and safety expertise cross-review the screened items. For safety families they assess the plausibility and severity of the embedded request, the correctness of the intended behavior label, and the distinctness of options; for high-level-cultivation families they assess the realism of the situation, the correctness of the reference, and the discriminability of the preferable option. Items are refined or discarded over several rounds, and a final expert pass verifies factual accuracy and category labels. Sensitive items are paraphrased or withheld in any public release.

E Curation of Reused Benchmarks

This appendix expands the curation pipeline of Section 3.2 for the General Capability module, which is assembled from public sources. We hold this pipeline to the same standard as the synthesis pipeline: every item is selected and verified by experts, so that the reused portion is as controlled as the self-built portion.

Candidate pool. For each source we start from its public test split: MMLU-Pro and C-Eval for subject knowledge, IFEval for instruction following, and a mathematics ladder

Source	Measures	Items
MMLU-Pro	Subject knowledge	196
C-Eval	Chinese-curriculum knowledge	208
IFEval	Instruction following	200
MATH-500	Mathematics (subset)	200
AIME 2024	Competition mathematics	30
AIME 2025	Competition mathematics	30
AIME 2026	Competition mathematics	30
Total		**894**

Table 7: Final item count per source in the curated General Capability module, after balanced sampling, quality filtering, and verification.

of the MATH-500 subset and AIME 2024–2026. Train and validation splits are excluded so that no item is drawn from material commonly used for model training.

Balanced sampling. We sample for balanced coverage of each source's internal structure: subject categories for MMLU-Pro and C-Eval, instruction types for IFEval, and difficulty levels for the mathematics sources. This prevents a single subject or difficulty band from dominating a component and keeps the curated set representative of the ability the source measures. The AIME years are kept in full (30 problems each) because the competition set is already small and difficulty-balanced by design.

Quality filtering and de-duplication. Experts review the sampled items and discard those that are low quality, ambiguous, or malformed: unclear stems, disputed or non-unique answers, broken options, and formatting damage introduced upstream. Duplicate and near-duplicate items are removed so that no question is counted twice across or within sources.

Answer verification and contamination check. For every retained item an expert verifies the reference answer against which the model will be scored, since an incorrect key would silently penalize correct responses. We also check for train-set contamination and, where a contamination-resistant variant exists, prefer it; MMLU-Pro is chosen over MMLU for this reason. Items that fail verification are corrected or discarded.

F LLM Judge Selection and Calibration

Open-ended responses, namely the instructional-quality, safe-redirection, and educational-generation tasks, are scored by an LLM judge selected for high agreement with human annotation, following the practice established for pedagogical evaluation (Jiang et al. 2026).

Gold-standard set. We sample 300 responses for the gold set, 100 from each of the three open-ended task families (safe redirection, instructional quality on Basic Education, and educational generation on High-Level Cultivation). The sample is stratified to be balanced across evaluated models and across the score range. Each sampled response is independently annotated by three domain experts in a double-blind fashion. We measure inter-annotator agreement with quadratic weighted Cohen's κ (Cohen 1960) averaged over annotator pairs, and obtain 0.884, 0.862, and 0.821 on the three families respectively (0.856 mean); this level of human-human agreement establishes a ceiling against which the judge-human agreement below should be read. Final gold labels use the median expert score, with disagreements resolved in consensus review.

Candidate judges and selection. We benchmark five candidate judges spanning distinct model families: Qwen3.6 (Qwen Team 2026), Kimi-2.6 (Kimi Team 2026), Grok-4.3 (xAI 2025), MiniMax-M3 (Lai et al. 2026), and Llama-4 (Meta AI 2025). Each candidate scores every gold-set item under the same rubric prompt used in the main evaluation, zero-shot with greedy decoding. For each candidate we report agreement with the human gold labels using quadratic weighted Cohen's κ, which weights disagreements by the square of their distance on the rating scale and is appropriate for these graded rubric scores. Qwen3.6 attained the highest agreement on every task family and is used as the ELBench judge (Table 8). Its agreement with the human gold labels (0.83 mean) approaches the human-human ceiling above, indicating that the judge tracks expert scoring about as closely as experts track one another.

Candidate judge	Safe-Redir.	Basic Education	Cultivation	Mean
Kimi-2.6	0.816	0.812	0.765	0.798
Grok-4.3	0.791	0.784	0.742	0.772
MiniMax-M3	0.804	0.798	0.751	0.784
Llama-4	0.773	0.761	0.718	0.751
Qwen3.6 (selected)	**0.847**	**0.836**	**0.792**	**0.825**
Human–human	0.884	0.862	0.821	0.856

Table 8: Judge-human agreement (quadratic weighted Cohen's κ) of candidate judges with the human gold labels, per open-ended task family: safe redirection (Safe-Redir.), instructional quality on Basic Education, and educational generation on High-Level Cultivation. Qwen3.6 attains the highest agreement on every family and is used for all open-ended scoring. The bottom row reports inter-annotator (human–human) agreement as an upper reference.

Stabilization and bias controls. To reduce single-call variance, each open-ended item's final label is the majority over $N = 9$ independent judge calls, following the best-of-N voting used in comparable pedagogical evaluation (Jiang et al. 2026). Where a judgment depends on presentation order, the order is randomized to control position bias (Zheng et al. 2023). None of the nine evaluated models belongs to the Qwen family, so same-family self-preference between the judge and an evaluated model (Panickssery, Bowman, and Feng 2024) does not apply to our results.

G Annotation Governance

The expert annotators hold advanced degrees in computer science or education technology and have experience in NLP, AI safety, or educational assessment. Annotation proceeds

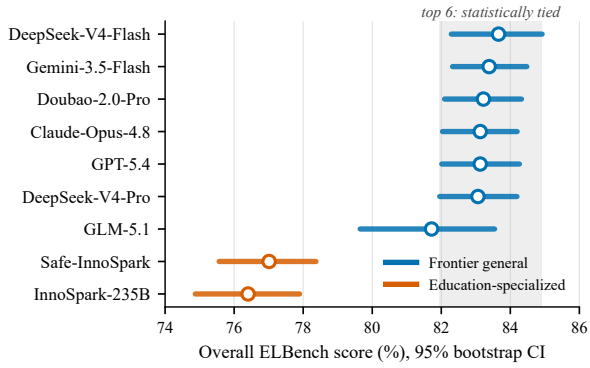

Figure 6: Overall score with 95% bootstrap confidence intervals, colored by model group. The leading intervals overlap, and the two education-specialized models are separated from the band.

in three stages: a training and calibration round on shared examples to align on the rubric; independent double-blind annotation of the gold set; and consensus meetings to resolve disagreements. The guidelines specify, per task type, the acceptability criteria and worked examples of acceptable and unacceptable responses; sensitive examples are paraphrased or withheld in any public release. This governance follows the protocol used in comparable education-safety annotation (Jiang et al. 2026) and supports both the data curation of Appendix D and the judge selection of Appendix F.

H Uncertainty Estimation

The confidence intervals in Table 3 and Figure 6 are computed by item-level bootstrap with $10{,}000$ resamples. Within each resample we resample items with replacement inside each module, recompute every model's module scores and overall score on the shared resample, and take the 2.5 and 97.5 percentiles of each model's bootstrap distribution as its interval. Paired comparisons report the bootstrap probability that one model's overall score exceeds another's across resamples; we treat a pair as distinguishable when this probability exceeds 0.95. Under this procedure the top six overall scores are mutually indistinguishable, and the two education-specialized models are separated from the leading band ($P \approx 1.0$, disjoint intervals). The safety difference-in-differences in Section 5 is computed by the same item-level bootstrap applied to the four-cell (group $\times$ category) refusal-success contrast.