BANDURABENCH: THE MECHANISM MIRAGE IN THEORY-GROUNDED EDUCATIONAL LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are often evaluated by whether they produce the correct final answer, but a correct decision does not necessarily imply that the model identified the right evidence or recovered the mechanism supporting it. We introduce **BanduraBench**, a 4,530-item benchmark grounded in Bandura's Social Cognitive Theory of Self-Efficacy for evaluating the gap between **decision accuracy** and **mechanism recovery** in educational reasoning. BanduraBench combines foundational construct evaluation with contextualized instructor- and researcher-facing reasoning tasks constructed from an expert-reviewed ontology, human-verified seeds, gold-first blueprints, and human-adjudicated admissible pathways. Models are evaluated on explicit, externally specified mechanism components—Bandura source, task-specific online-learning self-efficacy, course-design interpretation, and causal-claim strength. Across 15 LLMs, DesignReasoning answer accuracy is nearly saturated at 96.6% on average (95.3–98.2%), while normalized four-component recovery averages only 65.5% and ranges from 45.1–84.7%. Strict four-component recovery averages 21.4%, showing that much of the strict-score reduction follows from conjunctive accumulation of component errors. BanduraBench therefore provides a conjunction-aware framework for evaluating whether similar educational decisions are supported by similarly coherent theory-grounded representations.

1 INTRODUCTION

Large language models (LLMs) are increasingly evaluated on tasks that require multi-step reasoning, yet performance is often summarized by whether the model reaches the correct final answer (Wei et al., 2022; Lightman et al., 2024). In educational decision making, this can be insufficient. Several interventions may appear pedagogically reasonable while reflecting different interpretations of the learner evidence and different theoretical mechanisms. A model may therefore select an appropriate action without correctly identifying why that action is supported.

Bandura's Social Cognitive Theory of Self-Efficacy provides a useful setting for studying this distinction. Self-efficacy is informed by mastery experiences, vicarious experiences, verbal or social persuasion, and physiological or affective states (Bandura, 1997; Usher & Pajares, 2008), while online-learning self-efficacy is task-specific rather than a single general trait (Shen et al., 2013; Zimmerman & Kulikowich, 2016). These constructs induce an observable reasoning structure connecting course conditions and learner evidence to a self-efficacy source, an online-learning capability, and an instructional or research response. This allows us to evaluate not only whether an LLM selects the benchmark-preferred decision, but whether it recovers the theory-grounded structure associated with that decision. We refer to this relationship as *decision–mechanism consistency*. Importantly, this is an evaluation of explicit model outputs against an externally specified benchmark representation; it does not claim to recover a model's hidden chain-of-thought.

We introduce **BanduraBench**, a 4,530-item benchmark for evaluating theory-grounded educational reasoning. It contains two complementary components: **BanduraBench-Core**, which evaluates foundational construct knowledge and local theoretical relations, and **BanduraBench-DesignReasoning**, which evaluates contextualized instructor- and researcher-facing reasoning

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.

across ten task families. BanduraBench uses a gold-first construction process in which expert-reviewed theoretical structures are fixed before model-facing scenarios and response options are realized. Directed acyclic graphs and ordered mechanism representations encode the intended relations, while a ContextMatrix systematically varies educational context without changing the target mechanism. The DesignReasoning items are human adjudicated, and the evaluation set is additionally audited for simple answer-position, option-length, and option-only lexical shortcuts.

Across 15 closed- and open-weight LLMs, the DesignReasoning evaluation reveals a substantially different picture from final-answer accuracy alone. Decision accuracy is tightly compressed between 95.3% and 98.2%, with a mean of 96.6%. In contrast, normalized recovery of four structured mechanism components—the Bandura source, OLSE indicator, course-design cluster, and the strength of the causal claim—ranges from 45.1% to 84.7% (mean 65.5%). Exact recovery of all four components simultaneously ranges from 5.8% to 50.4%.

Across models, the observed exact score largely tracks the conjunction expected from component-level performance, while course-design attribution emerges as the weakest component on average. Thus, near-ceiling decision accuracy can mask large differences in the quality of the explicit theoretical structure models recover.

Contributions. Our contributions are threefold:

- **An observable formulation of decision–mechanism consistency.** We separate final-decision correctness from recovery of explicit, externally specified theory-grounded components, without making claims about latent chain-of-thought.
- **A theory-grounded benchmark for educational LLM reasoning.** We develop BanduraBench using an expert-reviewed ontology, human-verified seeds, gold-first construction, DAG-based mechanism representations, and ContextMatrix-controlled variation across realistic STEM learning contexts.
- **A conjunction-aware evaluation of mechanism recovery.** Across 15 LLMs, we show that near-saturated decision accuracy obscures substantial variation in component-level theory recovery. We decompose this variation across Bandura source, online-learning self-efficacy, course-design, and causal-claim components, and distinguish accumulated component error from additional joint-recovery effects using explicit null models.

Figure 1 summarizes the complete BanduraBench construction and evaluation pipeline.

2 Related Work

Educational LLM benchmarks increasingly evaluate capabilities beyond subject-matter accuracy. K-12EduBench measures knowledge and educational-goal cognition (Ye et al., 2026); MathTutorBench and KMP-Bench evaluate open-ended tutoring and pedagogical reasoning (Macina et al., 2025; Shi et al., 2026); EduGuardBench evaluates professional teaching fidelity and education-specific safety (Jiang et al., 2026); and Teach2Eval evaluates LLMs through teaching effectiveness rather than static answer correctness (Zhou et al., 2026).

Recent benchmarks also show why final-answer correctness alone can be insufficient. CogBench evaluates cognitive alignment in educational question answering and reports that correct answers can coexist with explanations that do not match the intended cognitive level (Lu et al., 2026). FineReason decomposes logic-puzzle reasoning into atomic state-checking and state-transition tasks rather than evaluating only the final answer (Chen et al., 2025); PRISM-Physics similarly evaluates structured dependencies in physics reasoning (Zhao et al., 2026). BanduraBench is complementary to these approaches: its target is recovery of an externally specified, theory-grounded educational mechanism linking course conditions and learner evidence to Bandura efficacy-information sources, task-specific OLSE, course-design interpretation, causal qualification, and aligned educational responses. It additionally evaluates human-adjudicated alternative pathways and uses conjunction-aware controls so that strict joint recovery is not conflated with a separate reasoning failure.

Open-ended educational evaluation also requires reliable semantic assessment. MathTutorBench uses a pedagogical reward model (Macina et al., 2025), MRBench evaluates automated judges

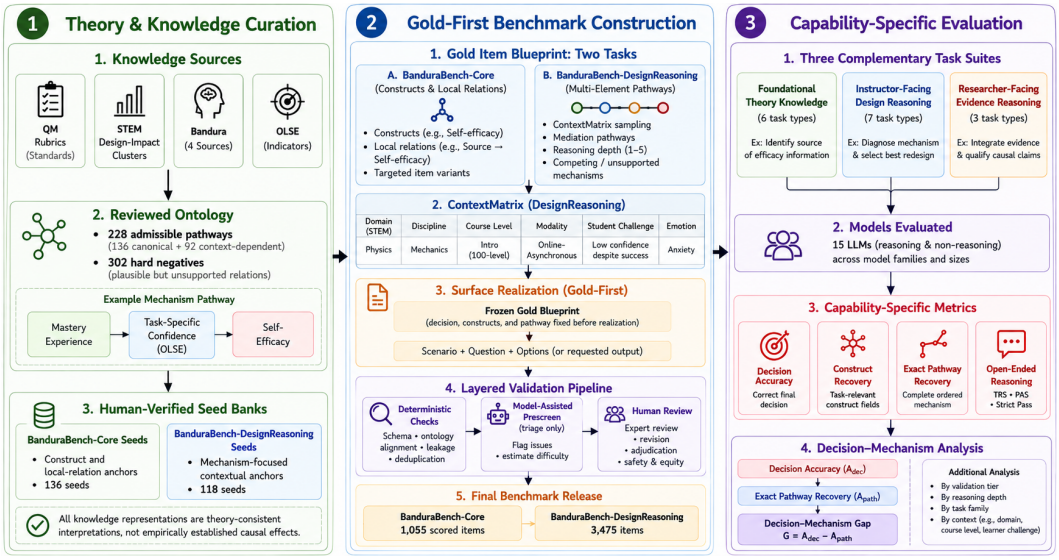


Figure 1: Overview of the BanduraBench pipeline, from theory-grounded curation and gold-first construction to capability-specific LLM evaluation.

against human annotations (Maurya et al., 2025), and Prometheus demonstrates fine-grained rubric-based LLM evaluation (Kim et al., 2024). BanduraBench therefore uses a hybrid protocol: decisions and controlled mechanism fields are scored deterministically, while open-ended explanations and pedagogical responses are evaluated with human-calibrated judges. The central structured-recovery conclusions do not depend on LLM-as-a-judge scoring.

3 Problem Formulation and Mechanism Representation

3.1 Observable Decision and Mechanism Targets

Consider an educational reasoning item i with observed scenario x_i and benchmark-preferred decision y_i^*. Given x_i, a model produces a decision $\hat{y}_i$ together with explicit structured outputs describing the mechanism supporting that decision.

For DesignReasoning, we distinguish the final decision from a controlled mechanism signature

$$\mathbf{z}_i^* = (b_i^*, o_i^*, c_i^*, q_i^*), \tag{1}$$

where b_i^* is the Bandura self-efficacy source, o_i^* is the task-specific online-learning self-efficacy (OLSE) indicator, c_i^* is the course-design cluster, and q_i^* is the appropriate causal-claim strength. The model analogously produces $\hat{\mathbf{z}}_i = (\hat{b}_i, \hat{o}_i, \hat{c}_i, \hat{q}_i)$.

Decision correctness is

$$D_i = \mathbb{1}[\hat{y}_i = y_i^*]. \tag{2}$$

Mechanism recovery is intentionally not reduced to a single binary variable: the four components can be evaluated individually, averaged at the component level, or required jointly. This distinction is important because exact joint recovery is a conjunction of several predictions and therefore becomes mechanically stricter as component errors accumulate. Section 5 defines the corresponding component-normalized, joint-exact, and conjunction-adjusted measures.

We use *decision–mechanism consistency* to describe the relationship between D_i and recovery of the externally specified mechanism targets.

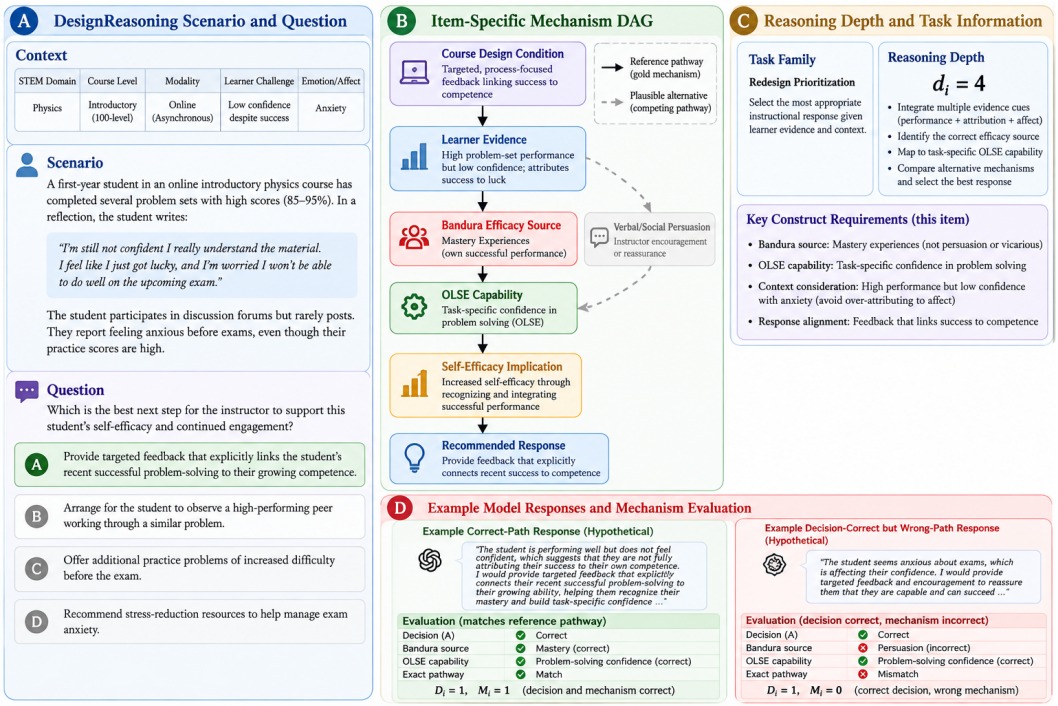


Figure 2: Example DesignReasoning item showing the item-specific mechanism DAG, reasoning depth, and the distinction between decision correctness and exact pathway recovery.

3.2 Theory-Grounded Mechanism Space

BanduraBench represents theory-grounded dependencies using a directed acyclic graph

$$G = (V, E), \tag{3}$$

whose typed nodes describe course-design conditions, learner evidence, Bandura sources of efficacy information, task-specific OLSE capabilities, self-efficacy implications, and aligned instructional or research responses. A canonical progression is

$$\textsc{Design} \rightarrow \textsc{Evidence} \rightarrow \textsc{Bandura} \rightarrow \textsc{OLSE} \rightarrow \textsc{Efficacy} \rightarrow \textsc{Response}. \tag{4}$$

Edges in E indicate dependencies licensed by the reviewed theoretical ontology; they should not be interpreted as empirically identified causal effects. An item instantiates only the task-relevant portion of this broader mechanism space.

To avoid assuming that every scenario has exactly one defensible theoretical realization, item i is associated with a human-adjudicated set of admissible pathways

$$\mathcal{P}_i = \left\{ P_i^{(1)}, \ldots, P_i^{(m_i)} \right\}, \tag{5}$$

where $m_i = 1$ for unambiguous items and $m_i > 1$ when multiple complete pathways are judged theoretically defensible. One pathway is retained as the primary reference for reproducibility, while accepted-path evaluation permits a prediction to match any *single complete* pathway in $\mathcal{P}_i$. Components from different accepted pathways are never mixed to construct an artificial match.

This representation therefore separates three objects:

$$\text{global theory space} \rightarrow \text{item-specific admissible pathways} \rightarrow \text{model-predicted structure}. \tag{6}$$

4 BENCHMARK CONSTRUCTION AND VALIDATION

BanduraBench contains 1,055 Core items and 3,475 DesignReasoning items. Core is split into 107 development, 159 validation, and 789 test items; DesignReasoning into 348 development, 521 validation, and 2,606 paper-test items. Task-family definitions and release fields are summarized in Appendix A.

4.1 GOLD-FIRST CONSTRUCTION

BanduraBench fixes the theoretical target before natural-language realization. Reviewed ontology pathways and human-verified seed banks determine the mechanism, learner evidence, decision, task family, and contextual constraints before an LLM realizes the scenario and options. Surface generators may vary wording but cannot redefine protected gold fields. DesignReasoning surface realization used five active generators—GPT-5.4, Claude Sonnet 5, Gemini 3.5 Flash, DeepSeek-V4 Flash, and Qwen3.7-Plus—under task-local failover; generator identity does not determine the gold answer or mechanism. Ontology inventories, ContextMatrix axes, complete generator provenance, and failover procedures are detailed in Appendix C.

Human adjudication. Generated candidates pass deterministic validation before expert review. Reviewers inspect answer support, mechanism coherence, distractor quality, causal humility, equity and anti-deficit framing, and other task-specific requirements. All Core and DesignReasoning items underwent final human adjudication, including review of complete alternative pathways when more than one interpretation is supported. Full procedures and quality-assurance analyses are in Appendix E.

5 EVALUATION PROTOCOL

5.1 METRICS AT A GLANCE

We report deterministic decision/mechanism metrics separately from human-calibrated semantic scores.

BanduraBench-Core. Let S_{jf} denote model j's primary score for Core family f. Bandura-source, course-design, and OLSE tasks use example-level F1; Quality Matters and redesign use exact match. We report

$$\text{Overall}_j = \frac{1}{6}\sum_{f\in\mathcal{F}} S_{jf}, \qquad \text{Structured}_j = \frac{1}{5}\sum_{f\in\mathcal{F}_{\text{str}}} S_{jf},$$

where $\mathcal{F}_{\text{str}}$ excludes explanation. For explanation items,

$$\text{Explanation}_j = \frac{1}{N_E}\sum_i \frac{1}{6}\sum_{d=1}^{6} c_{id},$$

and *Pass* requires at least four of six criteria, including Bandura alignment and effect-direction correctness. *Chain* is the proportion of eligible groups for which all tier-required components are correct. Full definitions are in Appendix F.

BanduraBench-DesignReasoning. Let $\mathcal{I}$ be the full paper-test set and $\mathcal{I}'$ the common four-field set ($|\mathcal{I}'| = 2{,}554$). For $c \in \{B, O, C, Q\}$, denoting Bandura, OLSE, course design, and causal claim,

$$A_{\text{ans}} = \frac{1}{|\mathcal{I}|}\sum_i \mathbb{1}[\hat{y}_i = y_i^*], \qquad A_c = \frac{1}{|\mathcal{I}'|}\sum_i \mathbb{1}[\hat{z}_{ic} = z_{ic}^*].$$

We summarize structured recovery with

$$S_{\text{comp}} = \frac{1}{|\mathcal{I}'|}\sum_i \frac{1}{4}\sum_c \mathbb{1}[\hat{z}_{ic} = z_{ic}^*], \qquad A_4 = \frac{1}{|\mathcal{I}'|}\sum_i \prod_c \mathbb{1}[\hat{z}_{ic} = z_{ic}^*].$$

Because A_4 is conjunctive, we also report

$$\Delta_{\text{comp}} = A_4 - \mathbb{E}_0[A_4],$$

where $\mathbb{E}_0[A_4]$ is the task-family×reasoning-depth permutation expectation. Accepted-path and leave-one-component-out variants are sensitivity analyses.

Semantic scores. For applicable rubric dimensions $\mathcal{A}_i^{(g)}$,

$$R_i^{(g)} = \frac{1}{|\mathcal{A}_i^{(g)}|} \sum_{d \in \mathcal{A}_i^{(g)}} \frac{s_{id}}{3}.$$

TRS uses theory/evidence dimensions, PAS uses pedagogical-action dimensions, and Strict Pass requires every task-required dimension to score at least 2. Full rubric and calibration details are in Appendix G.

6 Experimental Setup

We evaluate 15 closed- and open-weight LLMs zero-shot using common component-specific instructions and structured response schemas. Each model produces one response per item; protected answers, reference mechanisms, and admissible-path annotations are excluded from prompts. Retries are limited to transport or formatting failures, semantic errors are not regenerated, and unresolved malformed required fields remain in the denominator as incorrect. All 15 systems were freshly evaluated on the final DesignReasoning release through hosted APIs; full runtime identifiers, decoding settings, and prompt interfaces are in Appendices D and F.

Core uses its 789-item held-out test set. DesignReasoning answer accuracy uses all 2,606 paper-test items; the four-component analysis uses the same 2,554 eligible items for every model after the frozen OLSE rule excludes 52 items from OLSE-dependent scoring. Uncertainty uses 10,000 task-family-stratified bootstrap resamples and 10,000 task-family×reasoning-depth permutations; model-level permutation tests are Holm-adjusted (Holm, 1979). Core semantic calibration uses 125 human-scored development responses, while DesignReasoning calibration uses 400 annotations from 100 items spanning all ten task families. After calibration and targeted reannotation, 4,106 jointly applicable rubric cells reached 88.9% exact agreement and 100% agreement within one ordinal point; response-level pass agreement was about 95% ($\kappa \approx 0.90$). Additional inference, judge, and statistical details are in Appendices F–G.

Comparison policy. The primary claims concern broad differences between decision-level and structured recovery, not fine-grained model ranking. We therefore treat small cross-model differences descriptively, report bootstrap uncertainty for the principal structured metrics, preserve source-seed clusters in Core resampling where applicable, and keep judge-based semantic scores analytically separate from deterministic selected-answer and construct-recovery scores.

7 Results and Analysis

7.1 Main Results

BanduraBench-Core and DesignReasoning measure complementary capabilities and are reported separately rather than collapsed into one score.

Core shows stronger local construct and explanation performance than strict chain completion: the best strict-chain score is 23.2%. Across models, mean task-family scores are 87.3% for OLSE mapping, 54.6% for Bandura-source attribution, 49.6% for course-design attribution, 60.8% for Quality Matters identification, and 97.1% for redesign selection. The explanation rubric shows a similar asymmetry: design/evidence identification averages above 97%, whereas Bandura-source alignment averages 45.2%. Full task-family, explanation-dimension, and chain diagnostics are in Appendix H.1.

Table 1: BanduraBench-Core results (%). Structured and Explanation are aggregate scores; Pass is the frozen explanation pass rate; Chain is strict chain-group completion.

Model	Overall	Struct.	Expl.	Pass	Chain
GPT Luna	78.2	75.9	89.9	59.2	23.2
Gemini 3.0 Flash	77.8	76.5	84.4	41.5	18.3
GPT-4o	76.7	74.5	88.1	59.2	19.5
GPT-5 mini	76.3	73.5	90.3	68.5	18.3
Claude Haiku 4.5	75.3	72.9	87.2	41.5	12.2
Qwen3-235B-A22B-Thinking	74.0	72.3	82.8	36.2	8.5
DeepSeek-R1	72.7	70.5	83.8	40.8	11.0
Qwen3-32B	71.2	69.1	81.8	33.1	6.1
Phi-4 Reasoning	73.3	71.4	82.4	31.5	9.8
Qwen3.5 Flash	43.1	39.9	59.0	33.8	9.8
DeepSeek-V3	73.6	72.1	81.0	36.2	7.3
Mistral Large 3	72.7	70.8	82.4	40.0	13.4
Mistral Small 4	73.0	70.8	83.8	36.9	11.0
Llama 3.3 70B Instruct	73.5	71.8	82.3	27.7	3.7
Llama 4 Scout	68.5	66.2	79.9	23.8	3.7

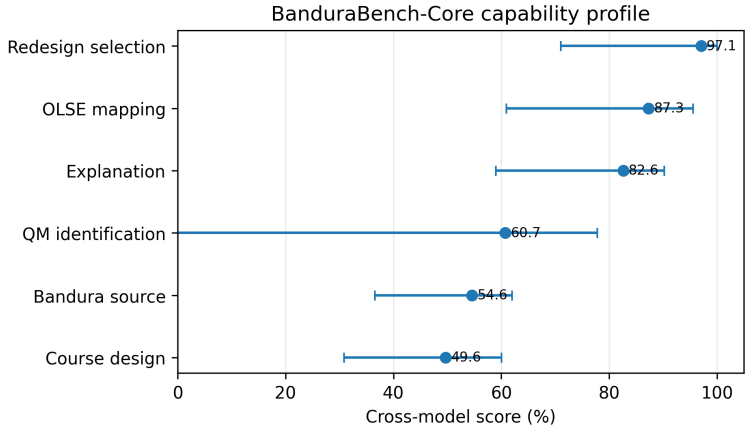


Figure 3: Cross-model BanduraBench-Core task-family performance.

7.2 Structured Recovery and Sensitivity

Role of selected-answer accuracy. Final-answer accuracy is nearly saturated (mean 96.6%, range 95.3–98.2%). We therefore do not treat selected-answer accuracy as the primary model-ranking metric; instead, it serves as a coarse decision-level anchor against which explicit mechanism recovery can be compared. The release also removes the severe option-internal shortcuts: the longest-option heuristic reaches 26.6%, while word- and character-TF–IDF option-only probes achieve 24.6% and 25.0% four-way accuracy. Full artifact and context-matching audits are in Appendix H.7.

CompNorm averages 65.5% and spans 45.1–84.7%. Across models, Bandura, OLSE, course-design, and causal-claim recovery average 67.0%, 66.9%, 54.2%, and 73.7%, respectively. Course design is the weakest field on average, although the limiting component differs by model. In leave-one-out analysis, omitting course design raises exact recovery by 16.1 points on average, versus 8.1 for OLSE, 7.1 for causal claim strength, and 4.6 for Bandura source; complete results are in Appendix H.3.

Reasoning-depth profiles. Stratifying the common structured-analysis set by annotated reasoning depth reveals a non-monotonic performance pattern rather than a simple linear decline. Mean CompNorm decreases from 71.1% at depth 1 to 57.0% at depth 5, while 4-Exact decreases from 27.0% to 13.4%; depths 2–4 are comparatively similar on both measures. The component-level profile further shows that this variation is not uniform across constructs: Bandura and OLSE recovery are lower at depth 5, whereas course-design and causal-claim recovery vary differently across depth strata. Because reasoning depth is associated with task-family and item composition, these differences are interpreted descriptively rather than as causal effects of greater reasoning depth.

Table 2: BanduraBench-DesignReasoning results (%). Ans. uses 2,606 items; structured metrics use 2,554. Null is the task-family×reasoning-depth permutation expectation

Model	Ans.	Band.	OLSE	Course	Causal	CompNorm	4-Exact	Null	Excess
GPT Luna	98.2	97.3	94.6	63.0	84.0	84.7	50.4	47.9	+2.5
Gemini 3.0 Flash	98.1	74.7	74.2	59.0	75.3	70.8	26.4	24.7	+1.7
GPT-4o	96.5	68.4	67.7	62.4	88.5	71.8	28.7	25.3	+3.4
GPT-5 mini	95.8	69.2	68.7	56.0	96.8	72.7	29.1	26.9	+2.2
Claude Haiku 4.5	97.0	69.3	67.7	51.4	90.3	69.7	23.2	21.4	+1.8
Qwen3-235B-A22B-Thinking	96.0	65.3	65.2	50.2	60.3	60.3	14.1	13.6	+0.5
DeepSeek-R1	95.3	67.7	67.6	54.2	59.3	62.2	15.7	15.0	+0.7
Qwen3-32B	96.5	64.5	59.7	48.0	48.7	55.2	9.6	9.0	+0.6
Phi-4 Reasoning	96.7	50.8	54.0	44.4	67.7	54.2	13.3	8.7	+4.6
Qwen3.5 Flash	96.8	70.9	69.7	56.4	83.3	70.1	24.3	22.5	+1.8
DeepSeek-V3	96.6	66.4	67.7	57.4	74.0	66.4	20.0	19.1	+1.0
Mistral Large 3	96.4	65.8	64.7	52.6	96.7	69.9	25.6	22.7	+2.9
Mistral Small 4	96.0	62.4	62.2	56.0	83.6	66.1	19.0	17.0	+2.1
Llama 3.3 70B Instruct	97.0	58.1	61.2	52.0	79.2	62.6	15.3	14.8	+0.5
Llama 4 Scout	95.9	54.3	58.4	50.2	17.5	45.1	5.8	4.5	+1.4

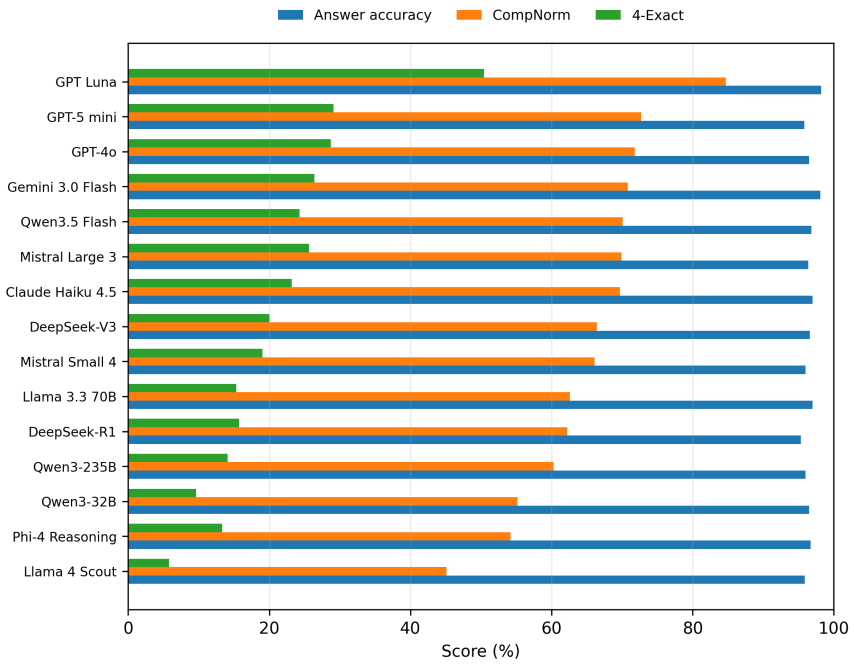
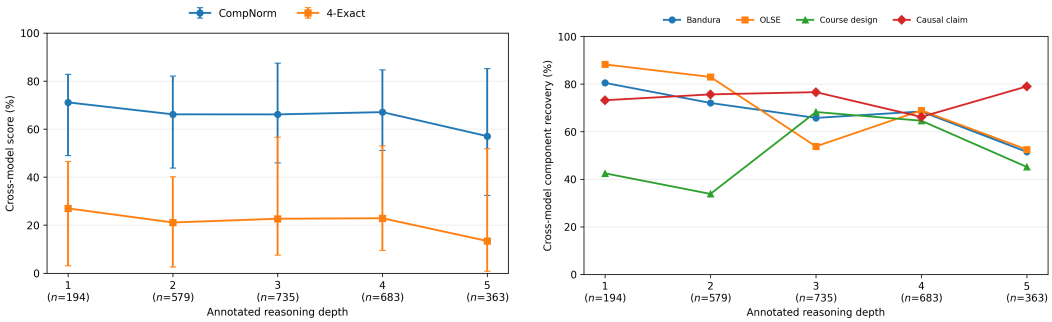


Figure 4: DesignReasoning final-answer accuracy, component-normalized mechanism recovery, and strict four-component recovery across 15 LLMs.

Strict 4-Exact averages 21.4%, but the product-of-marginals expectation is 19.4% and the stronger stratified permutation null is 19.5%, leaving only +1.84 points of mean observed-minus-null excess. Thus most of the CompNorm-to-4-Exact reduction follows from conjunction rather than an additional comparably large composition penalty. Ten model-level residuals remain significant after Holm correction (Holm, 1979), but these residuals are interpreted only as within-item dependence beyond the stratified marginals. Bootstrap intervals and complete null results are in Appendix H.2.

Alternative pathways and semantic quality. Human adjudication identifies 1,958 unique-path and 648 multi-admissible paper-test items. Allowing any complete accepted pathway increases the scorer's controlled Bandura–OLSE joint measure by 1.87 points on average (7.53 points within the multi-admissible subset), indicating that multiplicity matters locally but does not explain the broader dispersion in four-component recovery. Separately, human-calibrated rationale evaluation



(a) CompNorm and 4-Exact by reasoning depth (b) Component recovery by reasoning depth

Figure 5: DesignReasoning performance by reasoning depth.

yields mean TRS=81.3%, PAS=95.1%, and Strict Pass=52.4%. Full pathway and semantic tables are in Appendices H.4 and H.5.

8 Discussion and Interpretations

From local knowledge to structured recovery. BanduraBench-Core shows that models can perform well on individual constructs without consistently recovering the full theory-grounded structure. This pattern is stronger in DesignReasoning: final-answer accuracy averages 96.6%, while CompNorm averages 65.5% and varies much more across models. Likewise, PAS (95.1%) is higher than TRS (81.3%) and Strict Pass (52.4%), suggesting that appropriate pedagogical actions are easier to produce than fully theory- and evidence-grounded explanations.

Interpreting strict recovery. The low 4-Exact score should not be treated entirely as a separate reasoning failure. Its observed mean is 21.4%, while the stratified permutation null expects 19.5%. Most of the additional drop therefore reflects accumulated component errors. Final decisions, individual components, and joint recovery should consequently be interpreted separately.

Scope and implications. BanduraBench evaluates explicit model outputs against human-reviewed targets; it does not measure latent chain-of-thought or a human-performance ceiling. The results therefore characterize observable decision–mechanism consistency under the fixed zero-shot protocol. Overall, they show that models with very similar final decisions can differ substantially in the theory-grounded structure they explicitly recover.

9 Conclusion and Future Work

BanduraBench introduces a theory-grounded framework for evaluating educational LLM reasoning beyond final-answer accuracy. Across 15 LLMs, the results show that similar decisions can correspond to substantially different recovery of Bandura, OLSE, course-design, and causal-claim components. By separating decision accuracy, component recovery, conjunctive consistency, and semantic response quality, BanduraBench provides a more better view of theory-grounded reasoning in educational settings. Future work should examine extensions to additional educational theories and multi-turn educational interactions, including settings that integrate established social cognitive theories with mixed-methods evidence and reasoning.

REPRODUCIBILITY STATEMENT

We provide an anonymous reproducibility package containing the code, data, and artifacts required to inspect and reproduce the BanduraBench construction and evaluation pipeline. The package includes the reusable BanduraBench Python implementation (`src/`); executable scripts for dataset construction, validation, model evaluation, scoring, statistical analysis, and artifact verification (`scripts/`); frozen generation, evaluation, split, model, judge, and metric configurations (`configs/`); versioned generation, prescreening, target-model, and judge prompts (`prompts/`); and the machine-readable schemas used for dataset, generation, prescreening, and human review records (`schemas/`).

The package additionally contains the reviewed ontology, admissible pathways, hard-negative mappings, and controlled vocabularies (`ontology/`); canonical Core and DesignReasoning datasets and evaluation splits (`data/`); human-review, calibration, and adjudication records (`annotations/`); and the frozen experimental artifacts used for the reported evaluation (`experiments/`). These artifacts include model-facing inputs, outputs from the evaluated LLMs, judge inputs and outputs, rationale-dimension scores, and derived scoring records. The `results/` directory contains the corresponding aggregate metrics, item-level analyses, statistical tests, validation and shortcut audits, tables, and figures used in the paper. Cached evaluation outputs are included so that the reported scoring and analysis can be verified without requiring access to the original model APIs.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 Technical Report. *arXiv preprint arXiv:2412.08905*, 2024. URL https://arxiv.org/abs/2412.08905.

Alibaba Cloud. qwen3.5-flash Model Information, 2026. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-5-flash. Documents snapshot qwen3.5-flash-2026-02-23.

Anthropic. Introducing Claude Haiku 4.5, October 2025. URL https://www.anthropic.com/news/claude-haiku-4-5.

Albert Bandura. *Self-Efficacy: The Exercise of Control*. W. H. Freeman, New York, 1997.

Guizhen Chen, Weiwen Xu, Hao Zhang, Hou Pong Chan, Chaoqun Liu, Lidong Bing, Deli Zhao, Anh Tuan Luu, and Yu Rong. FineReason: Evaluating and improving LLMs' deliberate reasoning through reflective puzzle solving. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6685–6715, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.333. URL https://aclanthology.org/2025.acl-long.333/.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a. URL https://arxiv.org/abs/2501.12948.

DeepSeek-AI. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025b. URL https://arxiv.org/abs/2512.02556.

Michael D. Fetters, Leslie A. Curry, and John W. Creswell. Achieving integration in mixed methods designs—principles and practices. *Health Services Research*, 48(6pt2):2134–2156, 2013. doi: 10.1111/1475-6773.12117.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Google DeepMind. Gemini 3 Flash Model Card, December 2025. URL https://deepmind.google/models/model-cards/gemini-3-flash/. Published December 2025.

Timothy C. Guetterman, Michael D. Fetters, and John W. Creswell. Integrating quantitative and qualitative results in health science mixed methods research through joint displays. *The Annals of Family Medicine*, 13(6):554–561, 2015. doi: 10.1370/afm.1865.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2017.

Shaun R. Harper. An anti-deficit achievement framework for research on students of color in STEM. *New Directions for Institutional Research*, 2010(148):63–74, 2010. doi: 10.1002/ir.362.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020. URL https://miguelhernan.org/whatifbook.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Yilin Jiang, Mingzi Zhang, Xuanyu Yin, Sheng Jin, Suyu Lu, Zuocan Ying, Zengyi Yu, and Xiangjie Kong. EduGuardBench: A holistic benchmark for evaluating the pedagogical fidelity and adversarial safety of LLMs as simulated teachers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31356–31364, 2026. doi: 10.1609/aaai.v40i37.40399. URL https://ojs.aaai.org/index.php/AAAI/article/view/40399.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.248. URL https://aclanthology.org/2024.emnlp-main.248/.

Ron Legon. Measuring the impact of the quality matters rubric™: A discussion of possibilities. *American Journal of Distance Education*, 29(3):166–173, 2015. doi: 10.1080/08923647.2015.1058114.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Tong Lu, Zhichun Wang, Yuanhao Sun, Yaoyu Zhou, Mingrui Li, Yiming Guan, and Zhiyong Bai. CogBench: Benchmarking cognitive alignment of large language models in educational question answering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21242–21256, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1068. URL https://aclanthology.org/2026.findings-acl.1068/.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 204–221, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.11. URL https://aclanthology.org/2025.emnlp-main.11/.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1234–1251, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.57. URL https://aclanthology.org/2025.naacl-long.57/.

Meta AI. Llama 3.3 70B Instruct Model Card, 2024. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Model release date: December 6, 2024.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Mistral AI. Introducing Mistral 3, December 2025. URL https://mistral.ai/news/mistral-3/. Introduces Mistral Large 3.

Mistral AI. Introducing Mistral Small 4, March 2026. URL https://mistral.ai/news/mistral-small-4/.

OpenAI. GPT-4o System Card, August 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. GPT-5 System Card, August 2025. URL https://deploymentsafety.openai.com/gpt-5.

OpenAI. GPT-5.6 System Card, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6. Published July 9, 2026.

Quality Matters. QM Higher Education Rubric, Seventh Edition. Quality Matters, 2023. URL https://www.qualitymatters.org/qa-resources/resource-center/articles-resources/reference-use-qm-in-research. Used under license. All rights reserved. Retrieved from MyQM.

Demei Shen, Moon-Heum Cho, Chia-Lin Tsai, and Rose M. Marra. Unpacking online learning experiences: Online learning self-efficacy and learning satisfaction. *The Internet and Higher Education*, 19:10–17, 2013. doi: 10.1016/j.iheduc.2013.04.001.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32965–32973, 2026. doi: 10.1609/aaai.v40i39.40578. URL https://ojs.aaai.org/index.php/AAAI/article/view/40578.

Ellen L. Usher and Frank Pajares. Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research*, 78(4):751–796, 2008. doi: 10.3102/0034654308321456.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng

Zhou, and Zihan Qiu. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Yuqing Ye, Xuan Zhou, Zhifu Chen, Dandan Li, Hengnian Gu, Jin Peng Zhou, and Dongdai Zhou. K-12EduBench: A benchmark for evaluating large language models' knowledge, problem-solving, and educational goal cognition in K-12 education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34459–34466, 2026. doi: 10.1609/aaai.v40i40.40744. URL https://ojs.aaai.org/index.php/AAAI/article/view/40744.

Wanjia Zhao, Qinwei Ma, Jingzhe Shi, Shirley Wu, Jiaqi Han, Yijia Xiao, Si-Yuan Chen, Xiao Luo, Ludwig Schmidt, and James Y. Zou. PRISM-Physics: Causal DAG-based process evaluation for physics reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1a379b51d7ec80b32d8c14f40c281ce1-Abstract-Conference.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Yuhang Zhou, Xutian Chen, Yixin Cao, Yuchen Ni, Yu He, Siyu Tian, Xiang Liu, Yunwen Chen, Guangnan Ye, Xipeng Qiu, and Hongfeng Chai. Teach2Eval: An interaction-driven LLMs evaluation method via teaching effectiveness. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c98ef086dc70d528e1c1aa1e66893365-Abstract-Conference.html.

Whitney A. Zimmerman and Jonna M. Kulikowich. Online learning self-efficacy in students with and without online learning experience. *American Journal of Distance Education*, 30(3):180–191, 2016. doi: 10.1080/08923647.2016.1193801.

A BENCHMARK DEFINITION, TASK FAMILIES, AND RELEASE STRUCTURE

This appendix provides the benchmark inventory, task-family definitions, model-facing and protected fields, and evaluation denominators underlying the main-paper results. BanduraBench contains two complementary components: BanduraBench-Core for foundational construct knowledge and local relations, and BanduraBench-DesignReasoning for contextualized instructor- and researcher-facing mechanism reasoning.

A.1 FROZEN BENCHMARK INVENTORY AND EVALUATION DENOMINATORS

Table 3 summarizes the frozen counts used throughout the paper. These quantities are kept explicit because different metrics have different eligibility requirements.

Quantity	**Count**
BanduraBench-Core total	1,055
Core development / validation / test	107 / 159 / 789
Core held-out source-seed clusters	99
BanduraBench-DesignReasoning total	3,475
DesignReasoning development / validation / paper test	348 / 521 / 2,606
DesignReasoning common four-component set	2,554
Systematic OLSE-ineligible paper-test items	52
Unique-path paper-test items	1,958
Multi-admissible paper-test items	648
Evaluated target models	15

Table 3: Frozen BanduraBench inventory and the principal evaluation denominators.

Final-decision accuracy is defined on all 2,606 DesignReasoning paper-test items. Component-normalized recovery, four-component exact recovery, accepted-path sensitivity, Theory Reasoning Score (TRS), and Strict Pass use the common 2,554-item set unless stated otherwise. Pedagogical Action Score (PAS) is reported over the 1,048 items with at least one applicable pedagogical-action dimension. Core results use the 789-item held-out test set.

A.2 CAPABILITY THEMES, TASK FAMILIES, AND OUTPUTS

The benchmark covers three complementary capability themes: foundational theory knowledge, instructor-facing design reasoning, and researcher-facing evidence reasoning. The themes share the same reviewed ontology but place different demands on construct identification, evidence interpretation, mechanism recovery, and response selection.

Foundational theory knowledge. BanduraBench-Core contains six task families. Table 4 summarizes their model-facing inputs, required operation, and principal diagnostic purpose.

Instructor- and researcher-facing DesignReasoning. DesignReasoning contains ten contextualized task families. Seven are primarily instructor-facing and three are primarily researcher-facing. Table 5 records the frozen evidence/output defaults and the allowed reasoning-depth ranges. Depth ranges describe structural designs, not guaranteed empirical difficulty.

A.3 THREE CONDENSED SCENARIO ILLUSTRATIONS

The following examples illustrate how the three capability themes reuse the same theoretical vocabulary while changing the evidence and response demand. They are design illustrations rather than released test items.

1. **Foundational theory knowledge.** An online project prompt asks for a high-quality technical explanation but provides no rubric or annotated example. A student reports that they cannot judge whether their draft meets expectations. Different Core variants can ask for

Core family	Model-facing input	Required operation	Principal capability
Bandura-source attribution (Bandura, 1997; Usher & Pajares, 2008)	Short course/learner vignette	Select all supported sources	Distinguish mastery, vicarious, verbal/social persuasion, and physiological/affective evidence.
Course-design-cluster attribution (Legon, 2015; Quality Matters, 2023)	QM-aligned condition or artifact excerpt	Select affected design-impact cluster(s)	Separate neighboring course-design impacts such as clarity, alignment, usability, engagement, accessibility, knowledge acquisition, and learner support.
OLSE mapping (Shen et al., 2013; Zimmerman & Kulikowich, 2016)	Focal online-learning activity and learner challenge	Select task-specific capability	Distinguish task confidence from general motivation, emotion, performance, or satisfaction.
Quality Matters identification (Quality Matters, 2023)	Paraphrased course feature or review observation	Select one instructor-controlled condition	Map natural-language evidence to the intended QM-aligned design condition without relying on copied standard wording.
Redesign selection	Scenario plus candidate instructional actions	Select one mechanism-aligned redesign	Reject generally useful but mechanism-mismatched actions.
Mechanism explanation	Brief scenario requiring a concise explanation	Produce a short mechanism account	Connect design condition, learner evidence, Bandura source, OLSE capability, implication, and aligned response.

Table 4: Operational design of the six scored BanduraBench-Core task families. Test-set sizes are Bandura $n = 148$, course-design $n = 124$, OLSE $n = 148$, QM $n = 18$, redesign $n = 221$, and explanation $n = 130$.

Task family	Required evidence	Output	Depth	Principal capability
Instructor-facing design reasoning				
Mediation-chain completion	Student quote and course-artifact excerpt	Complete chain	1–5	Recover missing nodes or relations in an evidence–mechanism–response pathway.
Student voice to mechanism	Student quote	Classify	1–3	Infer the best-supported efficacy-information source and focal online-learning capability from learner voice.
Evidence selection	Student quote and course-artifact excerpt	Select evidence	2–4	Identify which details genuinely support the proposed mechanism and which are irrelevant or insufficient.
Pedagogical-mechanism repair	Student quote and course-artifact excerpt	Select repair	3–5	Choose an instructor-controlled action that targets the same mechanism and capability represented by the evidence.
Redesign prioritization	Course-artifact excerpt	Prioritize	3–5	Compare plausible course changes under timing, mechanism, and design constraints.
Competing-mechanism tradeoff	Student quote and course-artifact excerpt	Compare	4–5	Compare cases in which an intervention may strengthen one mechanism while weakening another.
Equity-mechanism invariance (Harper, 2010)	Student quote with controlled identity/access cue	Classify	3–4	Preserve an evidence-grounded diagnosis without replacing the mechanism with unsupported identity-based inference.
Researcher-facing evidence reasoning				
Redesign-evidence counterfactual reasoning	Comparative design-impact evidence	Compare	3–4	Interpret what a redesign or comparative pattern does and does not support.
Mixed-methods joint inference (Fetters et al., 2013; Guetterman et al., 2015)	Simplified quantitative result and student quote	Compare	4–5	Integrate quantitative and qualitative evidence, including convergence, complementarity, or tension.
Causal-humility reasoning (Hernán & Robins, 2020)	Simplified statistical result	Critique causal claim	3–5	Distinguish association, theory-consistent mediation, and an identified causal effect under limited evidence.

Table 5: DesignReasoning task families, model-facing evidence/output defaults, and allowed reasoning-depth ranges.

the relevant Bandura source, course-design cluster, OLSE capability, aligned redesign, or a concise mechanism explanation.

2. **Instructor-facing design reasoning.** A peer-review activity exposes students to successful peer strategies but also produces anxiety and unproductive comparison. A higher-depth item can ask the model to compare vicarious and affective interpretations and select a structured, low-stakes repair rather than simply remove peer interaction or add generic reassurance.
3. **Researcher-facing evidence reasoning.** A course redesign is followed by a positive association between perceived clarity and task confidence, while interviews indicate that students used a new rubric to judge successful performance. The model must integrate the evidence, identify a theory-consistent pathway, and explain why the design does not by itself identify causal mediation.

A.4 Model-Facing and Protected Fields

The model-facing record contains the scenario, question, response options when applicable, and the requested structured-output schema. Protected fields include the benchmark-preferred answer, controlled theoretical labels, primary and accepted pathway annotations, ontology and seed provenance, reviewer metadata, and scorer eligibility fields. Public and private variants preserve the same model-facing content while withholding protected gold metadata from the public evaluation interface.

B Theory, Ontology, DAGs, and Task Design

Section 3.2 defines the global mechanism representation. This appendix specifies why the DAG is used, how item-specific mechanisms are instantiated, how multiple admissible pathways are handled, and how reasoning depth differs from pathway identity and contextual variation.

B.1 Why a DAG Representation?

Educational reasoning mechanisms are difficult to evaluate using either a single class label or an unconstrained natural-language explanation. A model may identify a relevant construct while attaching it to unsupported evidence, map the correct online-learning confidence construct for the wrong reason, or select an appropriate instructional response without recovering the relations that support it. Independent component correctness therefore does not establish that the components form a coherent theory-grounded mechanism.

A single fixed linear chain is also too restrictive. Different scenarios may instantiate different subsets of the same theoretical representation, include multiple evidence sources, omit an intermediate node not required by the task, or support more than one complete interpretation. BanduraBench therefore uses a directed acyclic graph (DAG) to define the broader space of reviewed admissible dependencies while allowing each item to instantiate only the task-relevant portion of that space.

The canonical progression is

$$\textsc{Design} \rightarrow \textsc{Evidence} \rightarrow \textsc{Bandura} \rightarrow \text{OLSE} \rightarrow \textsc{Efficacy} \rightarrow \textsc{Response}. \tag{1}$$

DAG edges are benchmark-defined theory-grounded dependencies. They are not empirically identified causal effects and should not be interpreted as a structural causal model (Hernán & Robins, 2020).

B.2 Typed Nodes and Admissible Dependencies

Each node $v \in V$ is assigned a semantic type through $\tau : V \rightarrow \mathcal{T}$, with

$$\mathcal{T} = \{\textsc{Design}, \textsc{Evidence}, \textsc{Bandura}, \text{OLSE}, \textsc{Efficacy}, \textsc{Response}\}. \tag{2}$$

Type compatibility is necessary but not sufficient for an edge. For allowable type transitions $\mathcal{A} \subseteq \mathcal{T} \times \mathcal{T}$,

$$(u, v) \in E \Rightarrow (\tau(u), \tau(v)) \in \mathcal{A}, \tag{3}$$

and the reviewed ontology must additionally license the specific dependency. This prevents arbitrary connections between constructs that occupy compatible stages but are not theoretically supported.

B.3 Item-Specific Subgraphs and Admissible Pathways

The global graph G describes the complete set of admissible dependencies, but an individual item requires only the subset relevant to its scenario and task. For item i,

$$G_i = (V_i, E_i), \qquad G_i \subseteq G. \tag{4}$$

Human review associates the item with a set of complete admissible pathways

$$\mathcal{P}_i = \left\{ P_i^{(1)}, \ldots, P_i^{(m_i)} \right\}, \tag{5}$$

where $m_i = 1$ for unique-path items and $m_i > 1$ when multiple complete interpretations are judged comparably supported by the model-facing evidence. One pathway is retained as the primary reference for reproducibility, but accepted-path sensitivity permits a model prediction to match any one complete human-approved pathway. Components from different accepted pathways are never mixed to construct an artificial match.

The hierarchy is therefore

$$\text{global mechanism space} \longrightarrow G_i \longrightarrow \mathcal{P}_i \longrightarrow \hat{P}_i, \tag{6}$$

where $\hat{P}_i$ is the observable structured pathway returned by the model. The benchmark evaluates this submitted representation; it does not claim to recover a model's private chain of thought.

B.4 Controlled Signature versus Richer Pathways

The four-field signature used for the principal DesignReasoning component analysis is

$$z_i^* = (b_i^*, o_i^*, c_i^*, q_i^*), \tag{7}$$

covering Bandura source, OLSE capability, course-design cluster, and causal-claim strength. This signature is not a complete DAG path. It provides a common deterministic target across the 2,554-item structured-analysis set, whereas the richer pathway representation is retained for structural and alternative-path sensitivity analyses. This separation prevents 4-Exact from being mislabeled as full-path recovery.

B.5 Reasoning-Depth Definition

Each DesignReasoning item is assigned a structural reasoning-depth level

$$d_i \in \{1, 2, 3, 4, 5\}. \tag{8}$$

Depth is fixed during benchmark construction and represents the amount of structured composition required by the item, not prompt length or empirical item difficulty. It is broader than a literal path-edge count because some items require comparison of mechanisms, counterfactual reasoning, multiple evidence sources, or additional contextual constraints.

- **Depth 1:** one local construct or direct relation.
- **Depth 2:** two linked constructs or one constrained missing relation.
- **Depth 3:** a short evidence–mechanism–response chain.

- **Depth 4:** a competing mechanism, counterfactual, multiple-evidence relation, or contextual constraint.
- **Depth 5:** joint reasoning over several theoretical and contextual dependencies, potentially including multiple evidence sources, causal qualification, equity-sensitive constraints, or competing interpretations.

Conceptually,

Pathway: what mechanism	Depth: how much composition	Context: where

(9)

summarizes the three complementary design dimensions. Depth-conditioned performance is interpreted descriptively because task family, mechanism composition, evidence form, and other item properties vary across depth levels.

B.6 Theory and Interpretation Guardrails

Four guardrails apply throughout ontology construction and item realization: (1) learner reports and observed behavior are not treated as direct measurements of a latent trait; (2) identity cues cannot determine a mechanism without supporting evidence; (3) task wording distinguishes the primary, best-supported, or all-supported interpretation when mechanisms may co-occur; and (4) instructional responses must preserve learner agency and avoid deficit framing (Harper, 2010).

C Dataset Construction and Generation

BanduraBench uses a gold-first construction process. Reviewed ontology pathways, contextual anchors, task requirements, and scoring fields are fixed before natural-language realization. Generators can vary the model-facing surface form but cannot redefine protected theoretical targets.

C.1 Ontology Development and Pathway Inventory

We denote the frozen ontology by

$$\Omega = (\mathcal{C}, \mathcal{R}, \Pi^{+}, \Pi^{-}), \tag{10}$$

where $\mathcal{C}$ is the controlled construct inventory, $\mathcal{R}$ the reviewed admissible relations, Π^{+} the admissible mechanism pathways, and Π^{-} reviewed hard-negative mappings.

A broad triage covered 1,408 candidate records. The frozen ontology contains 136 canonical pathways. A focused review of 100 contextual candidates resulted in 74 direct accepts, 18 additional accepts after follow-up review, and eight deferred cases, yielding 92 context-dependent pathways. Thus,

$$|\Pi^{+}| = 136_{\text{canonical}} + 92_{\text{context}} = 228. \tag{11}$$

An additional 302 mappings are retained as hard negatives or failure probes. Canonical does not mean universally valid: every accepted path still requires scenario evidence for each task-required transition.

C.2 Human-Verified Seed Banks

The ontology defines abstract mechanisms, while reviewed seed banks provide realistic educational anchors for their realization. Conceptually,

$$\text{Ontology pathway} \rightarrow \text{Reviewed seed} \rightarrow \text{Generated item}. \tag{12}$$

BanduraBench-Core begins from 136 human-verified seed anchors. For DesignReasoning, 120 candidate seeds were reviewed; 118 were promoted after targeted repair and follow-up review, while two

were not promoted. Seed review considered theoretical alignment, contextual realism, sufficiency of learner evidence, Bandura and OLSE alignment, mechanism coherence, response alignment, distractor quality, causal humility, answer leakage, and anti-deficit framing.

C.3 CONTEXTMATRIX SPECIFICATION

BanduraBench-DesignReasoning uses a versioned **ContextMatrix** to vary the educational setting in which a reviewed mechanism is instantiated. For item i,

$$C_i = (c_i^{\text{inst}}, c_i^{\text{learner}}, c_i^{\text{decision}}), \tag{13}$$

where the three components describe instructional context, learner evidence, and decision context. The frozen specification contains 13 sampled axes and three derived axes.

Axis	Values	Function / representative values
course_level	4	Instructional level; lower/upper undergraduate, master's, doctoral.
stem_domain	9	Broad disciplinary family; computer science, engineering, life sciences, interdisciplinary STEM, and related domains.
stem_subdomain	158	Level- and domain-compatible course context; e.g., introductory programming, thermodynamics, genomics, research seminars.
modality	4	Online or hybrid delivery context.
course_artifact	67	Focal design object; syllabus, module, rubric, lab, discussion, feedback, technology tool, and related artifacts.
student_challenge_family	16	Higher-level challenge category; unclear expectations, feedback gap, platform difficulty, anxiety/overwhelm, and related families.
student_challenge	42	Level-aware concrete learner difficulty.
time_point	25	When the evidence occurs; orientation, after feedback, midterm recovery, proposal/manuscript review, and related points.
emotional_state	13	Reported or observed affective state.
audience	13	Intended decision-maker or reader; instructor, designer, QM reviewer, advisor, researcher, and related roles.
equity_cue	14	Controlled identity/access context, including first-generation, working, multilingual, rural/low-bandwidth, or accommodation contexts.
evidence_type	5	Artifact excerpt, student quote, Legon-CIS table (Legon, 2015), simplified SEM result, or mixed-methods joint display (Fetters et al., 2013; Guetterman et al., 2015).
output_type	7	Classify, complete chain, compare, prioritize, critique causality, select repair, or related structured operation.
Derived axes	3	audience_role_family, affective_direction, and equity_cue_relevance.

Table 6: Frozen ContextMatrix axes and representative functions. Complete controlled vocabularies and hierarchical mappings are maintained in the benchmark release files.

The ContextMatrix is sampled hierarchically rather than as an unrestricted Cartesian product. Construction first fixes the source pathway, task family, and reasoning depth. It then samples a compatible course level, STEM domain, subdomain, challenge family, concrete challenge, modality, time point, emotional state, audience, equity cue, evidence type, and output type. Derived fields are computed afterward.

Compatibility constraints prevent educationally or theoretically incoherent combinations. Course level constrains eligible subdomains, challenges, and time points; task-family defaults constrain evidence and output types; and fully in-person contexts are excluded because the benchmark is grounded in online and hybrid-online self-efficacy. Equity cues do not alter the gold mechanism merely because they are present; they affect the reference interpretation only when explicit scenario evidence makes the cue relevant to access, support, or interpretation.

The resulting construction separates

$$\underbrace{P_i^*}_{\text{mechanism}} \qquad \underbrace{d_i}_{\text{reasoning depth}} \qquad \underbrace{C_i}_{\text{context}}. \tag{14}$$

Thus, the pathway determines *what* mechanism is tested, reasoning depth controls the required amount of composition, and the ContextMatrix determines *where* the mechanism is instantiated.

C.4 GOLD-FIRST BLUEPRINT CONSTRUCTION

Four principles govern construction: (1) gold before prose; (2) immutable theory fields; (3) constrained contextualization; and (4) traceable transformations. For DesignReasoning item i, an admissible pathway P_i^*, human-verified seed S_i, ContextMatrix tuple C_i, task family T_i, and reasoning depth d_i define the structured blueprint

$$B_i = \mathcal{B}(P_i^*, S_i, C_i, T_i, d_i). \tag{15}$$

The blueprint fixes the required learner evidence, Bandura source, OLSE capability, course-design construct, self-efficacy implication, expected decision, required mechanism elements, allowed distractor relations, and output schema before surface generation.

C.5 CORE CANDIDATE CONSTRUCTION

Core uses reviewed seeds and task-family-specific variant plans to construct an immutable task blueprint before any optional surface enrichment. The programmatic layer determines the target task, gold labels, canonical answers, and permitted edits; an LLM may only realize or improve authorized surface fields.

Algorithm 1 Gold-preserving Core candidate construction

Require: Reviewed seeds $\mathcal{S}$; task-variant plan $\mathcal{V}$; validator G; authorized enrichment map A
Ensure: Candidate set $\mathcal{C}$

$\mathcal{C} \leftarrow \emptyset$
for all $(s, v) \in \text{AllowedVariants}(\mathcal{S}, \mathcal{V})$ **do**
  $b \leftarrow \text{FreezeCoreBlueprint}(s, v)$
  $x \leftarrow \text{RealizeProgrammaticVariant}(b)$
  if $\text{id}(b) \in \text{dom}(A)$ **then**
    $m \leftarrow A[\text{id}(b)]$
    $y \leftarrow \text{EnrichAuthorizedSurfaceFields}(m, x, b)$
    if $G(y, b) = \text{valid}$ **then**
      $x \leftarrow \text{MergeAuthorizedSurfaceFields}(x, y)$
    end if
  end if
  AssertImmutableGold(x, b)
  if $G(x, b) = \text{valid}$ **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{x\}$
  end if
end for
return $\mathcal{C}$

The canonical final Core release records verified later enrichment metadata for 470 of 1,055 records: 429 identify GPT-4.1-mini and 41 identify Qwen/Qwen2.5-14B-Instruct; 585 records contain no later LLM-enrichment record. These final-record metadata should not be interpreted as a reconstructed allocation for the complete 1,800-candidate generation pool.

C.6 DESIGNREASONING CANDIDATE CONSTRUCTION

DesignReasoning uses a pathway-first blueprint that separates the hidden gold representation from its natural-language realization.

C.7 TASK-PRESERVING MODIFIED ROUND-ROBIN SURFACE GENERATION

DesignReasoning used five active surface generators: GPT-5.4, Claude Sonnet 5, Gemini 3.5 Flash, DeepSeek-V4 Flash, and Qwen3.7-Plus. Each task was assigned a preferred generator before inference. The assigned model was attempted first; a failed generation could then trigger task-local cyclic failover across the remaining eligible active models and, when configured, an eligible reserve pool. The next task restarted from its own preassigned model rather than inheriting a global pointer.

Algorithm 2 Gold-first DesignReasoning candidate construction

Require: Frozen admissible pathways Π^+; reviewed seeds $\mathcal{S}$; task families $\mathcal{T}$; ContextMatrix $\mathcal{C}tx$; generator pool $\mathcal{M}$; validator G
Ensure: Candidate records $\mathcal{O}$

$\mathcal{O} \leftarrow \emptyset$
for all planned item specifications r **do**
  $P^* \leftarrow \mathrm{SelectAdmissiblePath}(\Pi^+, r)$
  $S \leftarrow \mathrm{SelectCompatibleSeed}(\mathcal{S}, P^*, r)$
  $(T, d) \leftarrow \mathrm{AssignTaskAndDepth}(r, P^*)$
  $C \leftarrow \mathrm{SampleCompatibleContext}(\mathcal{C}tx, P^*, T, d)$
  $B \leftarrow \mathcal{B}(P^*, S, C, T, d)$
  FREEZEGOLD(B)
  $m_0 \leftarrow \mathrm{PreassignedGenerator}(r, \mathcal{M})$
  $x \leftarrow \mathrm{TaskPreservingGenerate}(B, m_0, \mathcal{M})$
  if $x \neq \perp$ and $G(x, B) = \mathrm{valid}$ **then**
    $\mathcal{O} \leftarrow \mathcal{O} \cup \{\mathrm{AttachProvenance}(x, B)\}$
  end if
end for
return $\mathcal{O}$

Algorithm 3 Task-preserving modified round-robin surface generation

Require: Task blueprint t; active models $\mathcal{M}$; reserve models $\mathcal{R}$; eligibility predicate E; validator G
Ensure: Surface realization x or unresolved state $\perp$

$m_0 \leftarrow t.\mathrm{assigned_prompt_model}$
$Q_a \leftarrow [m_0] \,\|\, \mathrm{CyclicEligible}(\mathcal{M} \setminus \{m_0\}, m_0, t, E)$
$Q_r \leftarrow \mathrm{EligibleReserve}(\mathcal{R}, t, E, Q_a)$
$Q \leftarrow Q_a \,\|\, Q_r$
for all $m \in Q$ **do**
  $y \leftarrow \mathrm{InvokeGenerator}(m, t)$
  if y is returned and $G(y, t) = \mathrm{valid}$ **then**
    return y
  end if
end for
return $\perp$

A valid surface return remains only a benchmark candidate; it does not bypass later deterministic checks or human adjudication.

C.8 Automated Post-Generation Checks

Generated candidates are subjected to deterministic checks before final human adjudication. The validation layer checks parseability and required fields, controlled-vocabulary membership, preservation of frozen gold fields, required mechanism structure, task-family evidence and output requirements, answer-option consistency, prohibited leakage, ContextMatrix compatibility, hard-negative and distractor use, duplicate or near-duplicate candidates, and whether identity or affective cues improperly replace the evidence-supported mechanism. Causal-humility and anti-deficit checks are also applied. Model- assisted prescreening may flag ambiguous or suspicious records, but its outputs are non-authoritative and do not create or modify gold labels.

The construction sequence is therefore

$$\boxed{\text{Ontology} \rightarrow \text{Reviewed seed} \rightarrow \text{Blueprint} \rightarrow \text{Surface realization} \rightarrow \text{Automated validation} \rightarrow \text{Human adjudication}}. \tag{16}$$

D Prompt Templates and Model-Facing Interfaces

This section documents the prompt interfaces used for candidate generation, target-model evaluation, and semantic response evaluation. Generation prompts may receive protected construction metadata because they realize an already specified benchmark target. Target-model prompts do not receive item-specific gold answers, reference pathways, reviewer notes, generator identities, task-family labels, or reasoning-depth labels. Dynamic runtime fields are written in monospace braces below. The released prompt files preserve the complete static text; the templates here reproduce the operative instructions and schemas used for the reported experiments.

Stage	Component	Recipient	Principal dynamic inputs
Candidate generation	Core	Surface generator	Reviewed ontology context, audience/perspective, difficulty, and reviewer feedback when regenerating
Candidate generation	DesignReasoning	Surface generator	ContextMatrix summary, private generation specification, fixed answer position, assigned generator, and runtime output skeleton
Benchmark evaluation	Core	Evaluated LLM	Scenario, question, and task-specific response options when applicable
Benchmark evaluation	DesignReasoning	Evaluated LLM	Scenario, question, A–D options, and fixed global construct catalogs supplied in the system prompt
Semantic evaluation	Core	Calibrated judge	Scenario, question, approved reference, model explanation, and six assessment principles
Semantic evaluation	DesignReasoning	Calibrated judge	Human-adjudicated item/reference, target-model rationale, and task metadata

Table 7: Prompt interfaces used in BanduraBench. Protected construction fields are available only to generation/adjudication stages; target-model evaluation uses model-facing content and fixed global vocabularies.

D.1 Core Candidate-Generation Prompt

The ontology-conditioned Core prompt is used for LLM-assisted scenario realization or regeneration from a verified pathway. It supports both first- pass generation and reviewer-directed repair. Its principal runtime field is {`ontology_context_json`}, which contains the reviewed pathway, hidden construct targets, valid redesign actions, audience/perspective, difficulty information, and, when present, reviewer feedback. The generator is explicitly prohibited from inventing, revising, or exposing ontology-derived gold labels.

Frozen instruction structure.

> **Role.** "You are generating one BanduraBench-Core benchmark item from a verified ontology pathway." Return only valid JSON and no commentary.
>
> **Goal.** Create a realistic online or hybrid STEM scenario containing enough observable evidence for an expert reviewer to infer the hidden course- experience mechanisms, task-specific online-learning confidence issue, and aligned redesign need. Hidden ontology labels shape the scenario but must not appear verbatim in the public scenario, question, rationale, or student quote.
>
> **Runtime context.** {ontology_context_json}
>
> **Evidence coverage.** Provide a separate observable cue for every hidden Bandura source, course-design cluster, OLSE item, and valid redesign action. When reviewer_notes or human_review_feedback_for_regeneration is present, explicitly repair the stated defect while preserving the protected interpretation.

The prompt contains source-specific evidence rules (Bandura, 1997; Usher & Pajares, 2008). Mastery requires evidence about the learner's own task completion, practice, progress, feedback use, or success/failure experience; vicarious experience requires a peer/model/example or demonstration; social persuasion requires concrete feedback, coaching, encouragement, or help-seeking communication; and physiological/affective states require observable affective evidence. Analogous evidence requirements are supplied for the eight course-design clusters (Legon, 2015; Quality Matters, 2023): clarity of purpose, ease of use, course alignment, learner engagement, accessibility, knowledge acquisition, compliance, and learner support. The prompt separately requires evidence for the exact OLSE task (Shen et al., 2013; Zimmerman & Kulikowich, 2016) and for the problem addressed by the approved redesign action.

The requested perspective is also frozen. Supported frames include student_quote, qm_review_excerpt, course_artifact, instructional_designer_note, instructor_self_review, redesign_plan, and researcher_evaluation_excerpt. The scenario must identify an online or hybrid setting, a concrete STEM domain, and a course artifact; include exactly one plausible but irrelevant detail; avoid demographic stereotyping and deficit framing; and remain approximately 120–190 words. The generation question is fixed as: "Select all course-experience mechanisms most likely affected in this scenario."

Required generation schema. The generator returns the following record; the downstream construction script attaches benchmark answer options and protected labels from the verified ontology rather than allowing the generator to redefine them.

```
{
  "scenario": "...",
  "question": "...",
  "scenario_context": {
    "course_domain": "...",
    "modality": "...",
    "course_artifact": "...",
    "student_level": "...",
    "time_point": "...",
    "student_challenge": "...",
    "emotional_state": "...",
    "irrelevant_detail": "..."
  },
  "surface_features": {
    "student_quote_used": true|false,
    "course_artifact_evidence": "...",
    "design_issue_evidence": "...",
    "distractor_evidence": "...",
    "input_perspective_used": "...",
    "bandura_evidence_plan": {...},
    "legon_evidence_plan": {...},
    "olse_task_evidence": "...",
    "redesign_problem_evidence": "...",
    "review_feedback_addressed": "..."
  },
  "difficulty": {
    "level": <integer>,
    "label": "easy|medium|hard",
    "reasoning_hops": <integer>,
    "distractor_count": <integer>
  },
  "generation_notes": "..."
```

```
}
```

D.2 DESIGNREASONING CANDIDATE-GENERATION PROMPT

DesignReasoning uses one common blueprint-to-item prompt. Task-family behavior is supplied through the private generation specification, ContextMatrix realization, output type, reasoning depth, and fixed answer position rather than through ten separate long-form prompts. The dynamic fields are {context_summary}, {private_generation_spec_json}, {target_gold_answer_label}, {generator_model}, {required_output_json_skeleton}, and {non_gold_labels_json}.

Frozen instruction structure.

> **Role.** Generate one raw BanduraBench-DesignReasoning candidate item and return exactly one JSON object.
>
> **Context interpretation.** The supplied context tuple controls STEM domain/subdomain, modality, course artifact, student challenge, time point, emotional state, audience, equity cue, evidence type, and output type. Convert these fields into natural educational prose rather than exposing raw field names.
>
> **Private specification.** Use {private_generation_spec_json} to preserve the hidden mechanism and gold answer, but do not expose hidden theory or implementation identifiers unless the task explicitly requires a visible theory label.
>
> **Surface-realization requirements.** Preserve the hidden self-efficacy mediation mechanism and the fixed option {target_gold_answer_label}; create exactly four options A–D; make only the fixed option correct; make the remaining three plausible but incorrect distractors; respect the audience and output type; retain the assigned generator routing in {generator_model}; and avoid answer leakage, student-deficit framing, and unsupported causal-proof language.

The complete runtime JSON skeleton is injected programmatically. The static prompt independently enforces the following field contract, which therefore holds across candidate generations:

```
{
  "scenario": "...",
  "question": "...",
  "options": {"A": "...", "B": "...", "C": "...", "D": "..."},
  "gold_answer": "<target_gold_answer_label>",
  "rationale_gold": "...",
  "distractor_rationales": {
    "<non-gold label 1>": "...",
    "<non-gold label 2>": "...",
    "<non-gold label 3>": "..."
  },
  "leakage_self_check": {
    "no_gold_label_leakage": true,
    "no_deficit_framing": true,
    "no_unsupported_causal_claim": true
  }
}
```

The generator may not expose the hidden mediation chain, pathway/seed IDs, target answer variable, internal ontology IDs, raw QM standard IDs, or raw course-design cluster IDs in student-facing text. Raw implementation labels must instead be rendered as ordinary instructional-design language.

D.3 CORE TARGET-MODEL EVALUATION PROMPTS

The Core prompt suite uses the same common system instruction for all scored families:

> "You are answering BanduraBench-Core items about online STEM course design and student self-efficacy. Use only the scenario, question, and answer choices provided. Do not assume hidden metadata. Do not reveal chain-of-thought. Return only valid JSON in the requested schema."

The six scored families use three response forms. The three set-valued families share the following body:

```
Task: Select all answer choices directly supported by the scenario.

Scenario:
{scenario}

Question:
{question}

Options:
{options_text}

Return JSON in exactly this format:
{
  "answer": ["A", "C"]
}
```

Family-specific rules are preserved rather than collapsed into a single generic instruction. Bandura-source attribution requires every directly supported source and excludes merely possible sources; course-design attribution excludes broad or plausible choices not supported by evidence; OLSE mapping prefers the most specific supported capability and permits multiple selections only when multiple choices are clearly supported. Each form returns option letters only and prohibits explanations or extra keys.

Quality Matters identification and redesign selection use a single-choice schema:

```
Scenario:
{scenario}

Question:
{question}

Options:
{options_text}

Return JSON in exactly this format:
{
  "answer": "A"
}
```

For Quality Matters identification, the selected option must be the one most directly supported by the scenario or course artifact. For redesign selection, the model is instructed to prefer an action that addresses the represented course-design mechanism rather than generic encouragement or student blame.

The open-ended mechanism-explanation family uses:

```
Task: Write a concise mechanism explanation.

Scenario:
{scenario}

Question:
{question}

Return JSON in exactly this format:
{
  "explanation": "..."
}
```

The explanation is limited to two or three sentences and must ground the mechanism in scenario evidence, identify the focal course-design feature, connect it to the learner experience or task-confidence belief, identify a self-efficacy source when inferable, state the supported direction of the self-efficacy implication, and include a mechanism-aligned redesign or implication when requested. Hidden metadata, ontology IDs, markdown, and extra keys are prohibited. The archived standalone effect-direction prompt is not a seventh scored Core family in the present six-family analysis.

D.4 DesignReasoning Target-Model Evaluation Prompt

All ten DesignReasoning families use one common system prompt. The model sees only the scenario, question, and A–D options for the current item. The system prompt supplies the same global closed-set catalogs to every model: four Bandura sources (Bandura, 1997; Usher & Pajares, 2008);

25 OLSE task-confidence indicators (Shen et al., 2013; Zimmerman & Kulikowich, 2016); eight course-design clusters (Legon, 2015; Quality Matters, 2023); and four causal-claim-strength categories (Hernán & Robins, 2020). These catalogs expose the allowed vocabulary but not the correct values for any item.

The per-item model-facing payload is

```
Scenario:
{scenario}

Question:
{question}

Options:
{options_text}
```

and the frozen system instruction requires exactly one answer together with the observable structured mechanism fields used by the scorer:

```
{
  "predicted_answer": "A|B|C|D",
  "rationale": "2-5 sentences explaining the mediation reasoning.",
  "cited_evidence": "short supporting phrase",

  "predicted_bandura_source_id": "<Bandura catalog ID>",
  "predicted_bandura_source": "<plain-language label>",

  "predicted_online_learning_self_efficacy_indicator_id":
    "<OLSE catalog ID>",
  "predicted_online_learning_self_efficacy_indicator":
    "<plain-language label>",

  "predicted_course_design_cluster_id": "<cluster ID or UNKNOWN>",
  "predicted_course_design_issue": "<specific design issue or unknown>",

  "predicted_causal_claim_strength_id": "<causal catalog ID>",
  "predicted_causal_claim_strength": "<plain-language label>",

  "predicted_mediation_chain": [
    {"node_type": "student_experience_signal", "value": "..."},
    {"node_type": "bandura_source",
     "construct_id": "<same Bandura ID>", "value": "..."},
    {"node_type": "online_learning_self_efficacy_indicator",
     "construct_id": "<same OLSE ID>", "value": "..."},
    {"node_type": "pedagogical_or_design_repair", "value": "..."},
    {"node_type": "self_efficacy_trajectory",
     "value": "optional if inferable"},
    {"node_type": "performance_or_persistence_implication",
     "value": "optional if inferable"}
  ],

  "confidence": 0.0
}
```

The confidence field is the model's estimated probability that its selected answer is correct. Models are instructed to choose exactly one answer, keep the rationale and ordered chain concise, use `UNKNOWN` only when a construct genuinely cannot be inferred, and avoid causal overclaiming, identity overreach, deficit framing, and commentary outside JSON. The submitted mediation chain is therefore an explicitly requested observable response, not access to a model's private chain of thought.

D.5 Semantic-Evaluation Prompt Interfaces

Core explanation judge. The Core judge uses a point-wise evaluator prompt. Its system instruction is:

> "You are an expert evaluator for BanduraBench-Core mechanism explanations. Use point-wise evaluation. Score the model explanation using the provided gold reference and assessment principles. Be conservative: do not reward fluent text that is unsupported by the scenario, misidentifies the Bandura source, misstates the OLSE task confidence, gets the effect direction wrong, or shifts responsibility to the student when the mechanism is course-design related. Return only valid JSON in the requested schema."

The user-message template is

```
Scenario:
{scenario}

Question:
{question}

Gold reference:
{gold_reference_json}

Model explanation:
{model_explanation}

Assessment criteria and scoring principles:
{principles_json}
```

The judge returns six independent binary principle scores – course-design feature identification, Bandura-source connection, learner-experience or capability evidence, OLSE task-confidence connection, effect direction, and aligned redesign/implication – with a reason for each, their total and normalized scores, a Boolean pass recommendation, and a brief overall rationale. The pass recommendation is true only when the total is at least four and both Bandura-source connection and effect direction are correct; the benchmark recomputes this rule from criterion outputs.

DesignReasoning rationale judge. The DesignReasoning judge is instructed to evaluate the target model's rationale against the frozen human-adjudicated item and approved gold metadata, not to rewrite or re-adjudicate the benchmark. Its input contains the response-level calibration ID, benchmark item ID, task-family/validation metadata, scenario, question, options, approved gold answer/plan/mediation chain, and the target model's predicted answer, rationale, and cited evidence.

The output has the following structure:

```
{
  "calibration_id": "<copy exactly>",
  "scores": {
    "<each of the 16 frozen rationale dimensions>": 0|1|2|3|null
  },
  "overall_pass_recommendation": true|false,
  "failure_modes": ["<controlled failure-mode value>", "..."],
  "judge_comment": "brief evidence-based explanation"
}
```

The common ordinal anchors are 0=incorrect/contradicted/harmful or absent when required, 1=weak, 2=partially correct, and 3=strong, specific, evidence-grounded, coherent, and appropriately qualified; `null` is reserved for genuinely inapplicable dimensions. The 16 dimension names and frozen task-family applicability matrix are given in Tables 14 and 15. The controlled failure-mode vocabulary covers Bandura/OLSE/course-design confusion, chain breaks, wrong causal direction, student-deficit attribution, generic-support failure, evidence-retrieval error, context-prior override, multihop collapse, competing-mechanism failure, identity overreach, causal overclaim, audience mismatch, mixed-methods integration failure, CIS interpretation error, pedagogical-repair mismatch, vague rationale, and hallucinated evidence. Faithful paraphrases are accepted; answer correctness alone neither guarantees nor precludes a strong rationale score. Authoritative TRS, PAS, and Strict Pass values are recomputed from the criterion-level outputs as described in Appendix G.

E Human Review, Annotation, and Adjudication

BanduraBench uses human review at ontology, seed, realized-item, and human-reference stages. Automated validators and model-assisted prescreeners identify potential defects and organize review, but final acceptance, revision, rejection, alternative-path, and disputed-reference decisions remain human decisions.

E.1 Expert Review Roles

A six-member multidisciplinary team contributed according to domain expertise. Collectively, the team covered STEM and science education, online-course design and Quality Matters, Social Cognitive Theory and psychology, mixed-methods research, structural equation modeling, active-learning

interventions, and educational equity. Not every expert reviewed every record; assignments were matched to the construct or methodological decision under review.

Reviewer	Relevant expertise and primary contribution	Cadence
Expert A	Teaching and learning, physics education, STEM equity; study design, benchmark scope, implementation quality, and educational relevance	Weekly
Expert B	STEM education, mixed methods, qualitative research, equity; scenario realism, evidence integration, intervention relevance, and equity-sensitive interpretation	Monthly
Expert C	Psychology and theoretical evaluation; Social Cognitive Theory, self-efficacy alignment, theoretical consistency, and interpretation	Monthly; increased during analysis
Expert D	Socio-cognitive learning, structural equation modeling, and mixed methods; methodological design and quantitative–qualitative integration	Monthly
Expert E	Science education, active learning, educational equity; disciplinary realism, inclusive interventions, and anti-deficit framing	As needed
Expert F	Quality Matters and online-course design; QM-aligned conditions, online-course artifacts, design-quality criteria, and feasible mechanism-aligned redesigns	Monthly

Table 8: Anonymized expertise and primary review responsibilities of the BanduraBench expert team.

The team used a frozen, source-integrated annotation handbook containing a theory codebook and operational review criteria. The handbook specified ContextMatrix realism, course-design construct boundaries, Bandura-source evidence thresholds, OLSE task-confidence boundaries, mechanism/DAG coherence, mechanism-aligned repair, causal humility, answer leakage, equity safeguards, and anti-deficit framing.

E.2 Ontology and Seed Review

Human review begins before realized benchmark items are generated. The ontology and seed inventories are reported once in Appendix C.1 and Appendix C.2. At this stage reviewers verify theoretical grounding, contextual realism, evidence–mechanism consistency, construct boundaries, response alignment, causal humility, equity/safety, and whether a candidate is suitable to anchor later realization. LLM-assisted seed generation does not create new gold labels or make the final promotion decision.

E.3 Model-Assisted Prescreening

Before final human review, a frozen prescreen procedure can classify a candidate as `pass`, `minor_warning`, `needs_revision`, or `reject` and flag possible construct misalignment, ContextMatrix mismatch, pathway inconsistency, answer leakage, weak distractors, causal overclaim, equity concerns, or ambiguous gold. The prescreener also assigns a review priority. Every field is advisory: human reviewers may override the prescreen, and prescreen outputs never establish the final gold interpretation.

E.4 Item-Level Human Adjudication

Reviewers inspect the model-facing scenario, question, and options together with the protected answer, rationale, mechanism pathway/DAG summary, distractor rationales, relevant ContextMatrix fields, and validation or prescreen flags. The protocol proceeds through three sequential decisions.

Step 1: Gold-answer support. The designated answer is classified as **correct**, **ambiguous**, or **incorrect**. Correct means the designated option is uniquely the best-supported response under the visible evidence and task request. Ambiguous means another option or pathway is comparably defensible or the evidence is under-specified. Incorrect means the designated option is contradicted by the evidence or a different option is clearly better supported.

Step 2: Gold-pathway validity. The complete evidence–mechanism–response representation is classified as **valid**, **partially valid/repairable**, or **invalid**. A valid pathway requires every task-required node and transition to be explicitly supported or strongly inferable, correctly ordered, and consistent with the item-specific mechanism representation. A repairable pathway preserves the

central interpretation but requires clarification of a node, edge, rationale, or response. An invalid pathway cannot be supported without replacing the underlying interpretation.

The reference structure is

$$\text{course condition} \rightarrow \text{learner evidence} \rightarrow \text{Bandura source} \rightarrow \text{OLSE capability} \rightarrow \text{self-efficacy implication} \rightarrow \text{aligned response}. \tag{17}$$

Step 3: Item quality and final disposition. Reviewers then check scenario realism, ContextMatrix compatibility, task-family fit, reasoning-depth conformity, distractor plausibility, accidental validity, answer leakage, causal humility, equity/safety, and learner-agency preservation. Each item receives one of three dispositions:

- **Accept:** answer and pathway are supported and the item is realistic, answerable, safe, and sufficiently clear without substantive revision.
- **Revise:** the intended answer and pathway remain defensible, but a repairable problem affects wording, evidence foregrounding, distractors, leakage, causal strength, contextual framing, or mechanism-to-response alignment.
- **Reject:** the answer or pathway is unsupported, the item is fundamentally ambiguous or unsafe, or correction would require replacing the underlying gold interpretation.

For the current ICLR paper-test evaluation, all 2,606 DesignReasoning paper-test items underwent final human adjudication before model evaluation. Human adjudication means a qualified reviewer verified the item's answer, mechanism, and release disposition; it does not imply that every benchmark item was independently double-coded by the same pair of reviewers.

Decision point	Primary reviewer question
Gold answer	Is the designated response uniquely best supported by the model-facing evidence?
Gold pathway	Does every required node and dependency form a coherent evidence–mechanism–response pathway?
Construct boundaries	Are course-design, learner-evidence, Bandura, OLSE, self-efficacy, and response roles kept distinct?
Distractors	Are alternatives plausible but inferior, non-leaking, safe, and not accidentally valid?
Responsible interpretation	Are causal claims appropriately limited, and are identity or demographic cues prevented from replacing evidence?
Final disposition	Can the record be accepted as written, repaired without changing its gold interpretation, or must it be rejected?

Table 9: Principal decision points used during DesignReasoning human adjudication.

Reviewers record evidence-based notes for substantive decisions, including the relevant course condition, learner evidence, Bandura source, OLSE capability, answer/pathway judgment, distractor concerns, and any causal, equity, or leakage issue requiring adjudication.

E.5 Revision and Repair Verification

A revise disposition does not automatically promote an item. Authorized repairs may modify scenario wording, question wording, response options, evidence foregrounding, distractors, or other surface fields when the intended theoretical interpretation remains valid. Repaired records remain outside the accepted set until a human reviewer verifies that the corrected model-facing text remains consistent with the protected gold structure.

A targeted revision-effectiveness audit examined 453 DesignReasoning records whose initial revise disposition had not yet produced a substantive surface change. In the verified second pass, all 453 received revised questions and revised options A–D, while 258 also received revised scenarios. Overlapping repair categories included distractor plausibility or safety (434 records), answer leakage or cueing (339), scenario or context clarification (133), equity or anti-deficit wording (9), and causal-humility wording (6). The audit verifies that *revise* corresponded to an observable correction rather than only an administrative status change.

E.6 PATHWAY-MULTIPLICITY ADJUDICATION

Final paper-test review includes a dedicated pathway-multiplicity decision. Reviewers first verify the primary pathway inherited from the construction blueprint and then ask whether the visible evidence supports another complete theory-grounded interpretation at comparable validity. An alternative is admitted only when its entire evidence–construct–response sequence is coherent and supported; isolated plausible labels are insufficient. The final 2,606-item paper-test set contains 1,958 unique-path items and 648 multi-admissible items. These annotations are frozen before target-model scoring.

E.7 ANNOTATION QUALITY ASSURANCE

Quality assurance distinguishes benchmark-item adjudication from human-response annotation. The 2,606 paper-test items receive final human adjudication under the item-review protocol above, while the quantitative inter-rater statistics below apply specifically to the separate DesignReasoning human reference used to calibrate semantic response evaluation.

The human-reference set contains 400 evaluated responses formed from 100 benchmark items and four target-model responses per item. The sample spans all ten DesignReasoning task families and all 16 rationale dimensions. Two trained raters independently scored every response using the frozen rubric while blinded to target-model identity, deterministic answer-correctness fields, candidate-judge outputs, and one another's ratings.

The reference was partitioned by benchmark item into 300 development responses from 75 items and an untouched 100-response audit partition from 25 items, with no item crossing partitions. After rubric calibration and targeted reannotation, 4,106 jointly applicable numeric cells showed 88.9% exact agreement, 100% agreement within one ordinal point, and 1.9% threshold-crossing disagreement; no remaining numeric disagreement exceeded one point. Derived response-level pass agreement was approximately 95%, with pass/fail Cohen's $\kappa \approx 0.90$. In total, 456 disputed score cells were resolved by human adjudication, with no unresolved final disagreements. These figures describe convergence after calibration and targeted reannotation rather than untouched first-pass inter-rater reliability.

Agreement / adjudication quantity	**Result**
Calibration reference	400 responses / 100 items
Jointly applicable numeric cells	4,106
Exact dimension-level agreement	88.9%
Within-one-point agreement	100.0%
Threshold-crossing disagreement	1.9%
Derived response-level pass agreement	≈95%
Pass/fail Cohen's κ	0.899 (≈0.90)
Human-adjudicated disputed score cells	456
Unresolved final disagreements	0
Benchmark-reference escalations	30 items
Reference metadata corrected / retained	24 / 6

Table 10: Human-reference agreement and adjudication statistics for the DesignReasoning rationale-calibration study. These statistics describe post-calibration convergence after targeted reannotation and adjudication; they are not an untouched first-pass reliability estimate for the full benchmark release. Cohen's κ is computed on the derived response-level pass/fail decision. No Kendall τ_b statistic is used in the frozen primary agreement analysis.

Response annotation also served as an additional benchmark-reference audit. In the 100-item calibration sample, 30 items were escalated for project-level review; reference metadata were corrected for 24 and retained for six after adjudication. Potential benchmark-reference issues were therefore resolved at the item level rather than by changing a response score.

Scope of the human-reference statistics. These agreement statistics establish reproducibility of the semantic-response annotations and provide an additional audit of benchmark references. They

are not a blind human-solver baseline: benchmark-item adjudicators may inspect protected reference information, whereas a human recoverability study would require independent experts to answer model-facing items without access to gold labels or pathways. We therefore do not use these statistics as a human ceiling or as evidence that the absolute model–reference gap is human-sized.

F Models, Scoring, and Statistical Analysis

F.1 Evaluated Model Inventory and Inference Policy

The evaluation contains the same 15 systems reported in the main paper. All frozen model-evaluation runs were executed through provider APIs; no evaluated model was downloaded, fine-tuned, or locally served for the benchmark comparison. The only local-vLLM compatibility entry in the project configuration was disabled and was not part of the frozen evaluation. Models were evaluated zero-shot and single-turn with benchmark-component-specific instructions and a fixed structured response schema. The exact generation, target-model, and semantic-evaluation prompt interfaces are documented in Appendix D. Protected answers, reference mechanisms, accepted-path annotations, reviewer metadata, task-family labels, and reasoning-depth labels were excluded from target-model prompts. No demonstrations, self-consistency voting, external tool use, answer refinement, or correctness-triggered regeneration was used.

The settings below are *inference/decoding settings*, not training hyperparameters. BanduraBench did not train or fine-tune any of the 15 evaluated models. The table reports the runtime identifiers and request settings recorded for the DesignReasoning paper-test evaluation. Temperature is reported as 0.0 when the frozen request did not explicitly transmit a temperature value and therefore used the provider/model-side behavior. No explicit `max_tokens` or `max_completion_tokens` cap was sent in these runs.

Across the frozen DesignReasoning reports, the request timeout was 300 seconds, no explicit token-limit parameter was supplied, and model-specific structured output was requested only where supported. The evaluation protocol therefore standardized the task, schema, and scoring while respecting provider-specific request constraints rather than forcing unsupported decoding parameters.

F.2 Eligibility and Deterministic Scoring Details

For evaluation set $\mathcal{I}$, decision accuracy and component accuracy are

$$A_{\text{dec}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{1}[\hat{y}_i = y_i^*], \qquad A_c = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \mathbb{1}[\hat{z}_{ic} = z_{ic}^*]. \tag{18}$$

The common four-component set $\mathcal{I}'$ contains 2,554 items. Component- normalized and strict four-component recovery are

$$S_{\text{comp}} = \frac{1}{|\mathcal{I}'|} \sum_{i \in \mathcal{I}'} \frac{1}{4} \sum_{c \in \{B,O,C,Q\}} \mathbb{1}[\hat{z}_{ic} = z_{ic}^*], \tag{19}$$

$$A_4 = \frac{1}{|\mathcal{I}'|} \sum_{i \in \mathcal{I}'} \prod_{c \in \{B,O,C,Q\}} \mathbb{1}[\hat{z}_{ic} = z_{ic}^*]. \tag{20}$$

A simple independence diagnostic is

$$A_{\text{ind}} = \prod_c A_c. \tag{21}$$

Leave-one-component-out diagnostics recompute the conjunction after omitting one controlled field while retaining the same item set. For accepted-path sensitivity,

Paper label	Runtime model ID	API route	API?	Temp.	Structured-output mode
GPT Luna (OpenAI, 2026)	`gpt-5.6-luna`	OpenAI	Yes	1.0	JSON object
Gemini 3.0 Flash (Google DeepMind, 2025)	`google/gemini-3-flash-preview`	OpenRouter	Yes	0.0	JSON object
GPT-4o (OpenAI, 2024)	`gpt-4o`	OpenAI	Yes	0.0	JSON object
GPT-5 mini (OpenAI, 2025)	`gpt-5-mini-2025-08-07`	OpenAI	Yes	1.0	JSON object
Claude Haiku 4.5 (Anthropic, 2025)	`claude-haiku-4-5`	Anthropic	Yes	0.0	Native/no OpenAI JSON mode
Qwen3-235B-A22B-Thinking (Yang et al., 2025)	`qwen/qwen3-235b-a22b-thinking-2507`	OpenRouter	Yes	not sent	Native/no JSON mode
DeepSeek-R1 (DeepSeek-AI, 2025a)	`deepseek/deepseek-r1`	OpenRouter	Yes	not sent	Native/no JSON mode
Qwen3-32B (Yang et al., 2025)	`qwen/qwen3-32b`	OpenRouter	Yes	not sent	Native/no JSON mode
Phi-4 Reasoning (Abdin et al., 2024)	`microsoft/phi-4`	OpenRouter	Yes	not sent	Native/no JSON mode
Qwen3.5 Flash (Alibaba Cloud, 2026)	`qwen/qwen3.5-flash-02-23`	OpenRouter	Yes	0.0	JSON object
DeepSeek-V3 (DeepSeek-AI, 2025b)	`deepseek/deepseek-v3.2`	OpenRouter	Yes	0.0	JSON object
Mistral Large 3 (Mistral AI, 2025)	`mistralai/mistral-large-2512`	OpenRouter	Yes	0.0	JSON object
Mistral Small 4 (Mistral AI, 2026)	`mistralai/mistral-small-2603`	OpenRouter	Yes	0.0	JSON object
Llama 3.3 70B Instruct (Meta AI, 2024)	`meta-llama/llama-3.3-70b-instruct`	OpenRouter	Yes	0.0	JSON object
Llama 4 Scout (Meta AI, 2025)	`meta-llama/llama-4-scout`	OpenRouter	Yes	0.0	JSON object

Table 11: Frozen API routing and decoding configuration for the 15-model DesignReasoning evaluation. OpenAI calls used `https://api.openai.com/v1`; Claude used the Anthropic API compatibility endpoint; the remaining hosted models used `https://openrouter.ai/api/v1`. All 15 evaluated systems were invoked through APIs. "Not sent" indicates that temperature was not explicitly included in the frozen request.

Evaluation setting	Frozen policy
Evaluation mode	Zero-shot, single-turn; one retained terminal response per item
Model adaptation	None; no benchmark-specific training or fine-tuning
API execution	All 15 target LLMs evaluated through hosted APIs
Temperature	Default 0.0 when explicitly supported; GPT-5 mini and GPT Luna use 1.0; selected reasoning-model requests omit the parameter
Other sampling controls	Provider defaults unless explicitly overridden by the frozen model configuration
Token-limit policy	No explicit `max_tokens` or `max_completion_tokens` cap in the frozen evaluation suites
Structured output	JSON-object mode where supported; prompt-level structured JSON otherwise
Request timeout	300 seconds
Tools / retrieval	None
Self-consistency / voting	None

Table 12: Common inference and decoding policy for the evaluated model panel. These are evaluation-time settings; BanduraBench does not train or fine-tune the target models.

$$M_i^{\mathrm{adm}} = \max_{P \in \mathcal{P}_i} \mathbb{1}[\hat{P}_i = P], \tag{22}$$

so credit is assigned only when the prediction matches one complete human-approved pathway; components from different accepted pathways are never combined.

Core set-valued families use example-level F1 and single-answer families use normalized exact match. Core strict-chain diagnostics are defined over frozen source-seed–difficulty groups and require all tier-specified components to be correct.

F.3 Bootstrap Uncertainty and the Conjunction Null

For DesignReasoning structured metrics, uncertainty is estimated with 10,000 task-family-stratified bootstrap resamples. Each bootstrap replicate resamples items with replacement within task family and then recomputes the complete metric on the resulting stratified sample. Percentile 95% intervals use the 2.5th and 97.5th percentiles of the bootstrap distribution.

Because A_4 is conjunctive, it is compared with both the simple product of component marginals and a stronger task-family×reasoning-depth permutation null. Within each stratum, component correctness indicators are permuted independently, preserving the observed success total for every component while disrupting within-item co-occurrence.

Algorithm 4 Task-family×reasoning-depth conjunction null

Require: Item-level correctness matrix X for components $\{B, O, C, Q\}$; strata S; repetitions $R = 10{,}000$

Ensure: Null distribution $\{A_4^{(r)}\}_{r=1}^{R}$

```
for r = 1 to R do
    X^(r) ← X
    for all strata s ∈ S do
        for all components c ∈ {B, O, C, Q} do
            independently permute X_{s,c} across items in stratum s
        end for
    end for
    A_4^(r) ← FourExact(X^(r))
end for
return {A_4^(r)}_{r=1}^R
```

The diagnostic residual is

$$\Delta_{\mathrm{comp}} = A_4 - \mathbb{E}_0[A_4]. \tag{23}$$

The one-sided upper-tail permutation p-value is computed with the finite- replicate correction

$$p = \frac{1 + \sum_{r=1}^{R} \mathbb{1}[A_4^{(r)} \geq A_4^{\mathrm{obs}}]}{R + 1}. \tag{24}$$

Model-level permutation tests are Holm-adjusted across the 15 systems (Holm, 1979). Because the permutation breaks all within-item association, including association due to shared item difficulty, Δ_{comp} is interpreted as dependence beyond the stratified component marginals rather than as a causal or cognitive synergy estimate.

Where paired Core uncertainty is reported, resampling is performed over the 99 held-out source-seed clusters so that closely related task variants remain together within a bootstrap replicate.

Model-comparison interpretation. The task-family-stratified DesignReasoning bootstrap quantifies uncertainty for the principal structured metrics but is not used to declare a total ordering among

near-tied systems. Likewise, Core resampling preserves source-seed dependence when related variants are compared. We emphasize cross-model dispersion and mechanism-specific profiles; differences of a few tenths of a percentage point are treated as descriptive unless supported by an appropriate paired uncertainty analysis. Judge-based TRS/PAS/Strict-Pass values remain complementary semantic evidence and are not used to resolve close deterministic model rankings.

G Human Calibration and LLM-as-a-Judge Evaluation

LLM judges are used only for outputs that require semantic evaluation: Core mechanism explanations and DesignReasoning rationales/pedagogical responses (Zheng et al., 2023; Liu et al., 2023; Kim et al., 2024). Selected answers, controlled construct fields, and structured mechanism targets remain deterministically scored. Judge-provided overall ratings are not used as primary benchmark outcomes; authoritative aggregate scores are recomputed from criterion-level outputs.

G.1 Core Explanation Judge

Core explanation calibration uses 125 human-scored responses under the six binary criteria defined by the frozen judge prompt in Appendix D.5. GPT-5.6 Sol (OpenAI, 2026) is the production Core explanation judge. The official explanation score is the normalized sum of the six criterion outputs; the pass rule is recomputed deterministically and requires a total of at least four together with correct Bandura-source alignment and effect direction.

G.2 DesignReasoning Human Reference

Candidate-judge calibration uses the same 400-response human reference described in Appendix E.7. The item-disjoint split contains 300 development responses from 75 items and an untouched 100-response audit set from 25 items. Agreement and adjudication statistics are reported once in Table 10; no Kendall τ_b statistic is part of the frozen primary agreement analysis.

G.3 Candidate-Judge Calibration and Selection

Six candidate LLM judges were compared with the adjudicated 300-response development reference using the same prompt, ordinal rubric, applicability rules, structured-output schema, and deterministic aggregation. The predeclared selection criteria emphasized (i) quadratic-weighted agreement over the 0–3 scale, (ii) agreement on theory-critical dimensions, and (iii) agreement on the derived response-level pass decision. Coverage and mean absolute error were secondary diagnostics.

Candidate judge	Rank	Disposition	Selection record
GPT-5.6 Sol	1	Primary production judge	Highest predeclared composite rank on ordinal-scale, theory-critical, and derived-pass agreement.
GPT-5.6 Terra	2	Complementary production judge	Close second under the same primary criteria; retained to reduce dependence on one judge model.
GPT-5.5	3	Not selected	Favorable secondary diagnostics but lower on the primary construct-sensitive criteria.
DeepSeek-V4-Pro	4	Not selected	Lower composite ordinal and pass-level agreement.
DeepSeek-V4-Flash	5	Not selected	Lower composite ordinal and pass-level agreement.
Qwen3.7-Plus	6	Not selected	Lower composite ordinal and pass-level agreement.

Table 13: Frozen candidate-judge selection record on the 300-response development reference. Candidate-level point estimates are not reproduced when the complete per-response candidate outputs are not available in the compact artifact.

G.4 Held-Out Judge Audit and Production Allocation

After the judge prompt, rubric, applicability rules, parser, and selection criterion were frozen, GPT-5.6 Sol was evaluated on the untouched 100-response audit partition. Applicable-cell coverage was 98.4%, exact ordinal agreement 69.5%, within-one-point agreement 97.8%, quadratic-weighted $\kappa = 0.685$, and derived-pass Cohen's $\kappa = 0.677$.

For production DesignReasoning rationale scoring, each target model's 2,554 eligible responses are divided into balanced, disjoint subsets of 1,277 for Sol and 1,277 for Terra. The judges do not vote on the same response and their scores are not averaged. Both use the same frozen rubric, task-family applicability matrix, and aggregation rules. Consequently, deterministic selected-answer and controlled-field findings remain analytically separate from judge-based semantic results.

G.5 DesignReasoning Rubric

Each applicable rationale dimension is scored on

$$s_{id} \in \{0, 1, 2, 3\}, \tag{25}$$

where 0 denotes incorrect or absent evidence, 1 weak alignment, 2 partial or substantively acceptable alignment, and 3 strong alignment. N/A is allowed only when a dimension is genuinely inapplicable under the frozen task-family policy.

Code	Dimension	Primary score
B	Bandura-source alignment	TRS
O	OLSE alignment	TRS
M	Mediation-chain completeness	TRS
E	Evidence-grounding quality	TRS
C	Causal humility	TRS
Q	Equity/non-deficit framing	TRS
T	Tradeoff recognition	TRS
X	Mixed-methods integration	TRS
H	Rationale coherence	TRS
R	Mechanism-to-repair alignment	PAS
A	Actionability	PAS
G	Guidance quality	PAS
Y	Agency preservation	PAS
F	Affective appropriateness	PAS
L	Cognitive-load fit	PAS
V	Active-learning support	PAS

Table 14: Frozen DesignReasoning rationale dimensions and their assignment to Theory Reasoning Score (TRS) or Pedagogical Action Score (PAS).

TRS measures whether the rationale correctly identifies, grounds, and composes the theory-relevant mechanism, including task-specific epistemic requirements. PAS measures whether the proposed response is aligned with the diagnosed mechanism and provides an actionable educational response.

G.6 Task-Family Applicability

Applicability is determined from the frozen task family rather than from whether the judge elects to return a numeric score.

G.7 Theory Reasoning Score, Pedagogical Action Score, and Strict Pass

Let $\mathcal{A}_i^{(g)}$ denote the dimensions from score group $g \in \{\mathrm{TRS}, \mathrm{PAS}\}$ applicable to item i. For model j,

$$R_{ij}^{(g)} = \frac{1}{|\mathcal{A}_i^{(g)}|} \sum_{d \in \mathcal{A}_i^{(g)}} \frac{s_{ijd}}{3}. \tag{26}$$

Model-level scores average item-level scores,

$$R_j^{(g)} = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} R_{ij}^{(g)}, \tag{27}$$

Task family	Required dimensions	Codes
Evidence selection	Bandura, OLSE, evidence grounding, rationale coherence	B, O, E, H
Student voice to mechanism	Bandura, OLSE, evidence grounding, rationale coherence	B, O, E, H
Mediation-chain completion	Bandura, OLSE, mediation chain, evidence grounding, rationale coherence	B, O, M, E, H
Redesign-evidence counterfactual reasoning	Bandura, OLSE, mediation chain, evidence grounding, causal humility, rationale coherence	B, O, M, E, C, H
Mixed-methods joint inference	Bandura, OLSE, mediation chain, evidence grounding, causal humility, mixed-methods integration, rationale coherence	B, O, M, E, C, X, H
Redesign prioritization	Mediation chain, evidence grounding, mechanism-to-repair alignment, rationale coherence, actionability, guidance quality	M, E, R, H, A, G
Competing-mechanism tradeoff	Bandura, OLSE, mediation chain, evidence grounding, mechanism-to-repair alignment, causal humility, tradeoff recognition, rationale coherence, actionability, guidance quality, agency preservation	B, O, M, E, R, C, T, H, A, G, Y
Equity-mechanism invariance	Bandura, OLSE, mediation chain, evidence grounding, causal humility, equity/non-deficit framing, rationale coherence, agency preservation	B, O, M, E, C, Q, H, Y
Causal-humility reasoning	Evidence grounding, causal humility, rationale coherence	E, C, H
Pedagogical-mechanism repair	Bandura, OLSE, mediation chain, evidence grounding, mechanism-to-repair alignment, causal humility, rationale coherence, actionability, guidance quality, agency preservation	B, O, M, E, R, C, H, A, G, Y

Table 15: Frozen task-family applicability used for primary DesignReasoning rationale scoring.

so each eligible item has equal weight even when task families require different numbers of dimensions. Inapplicable dimensions are excluded. Missing required dimensions are assigned zero under the frozen scoring policy.

For Strict Pass, let $\mathcal{A}_i$ denote all dimensions required for item i. The raw applicable-dimension mean is

$$\bar{s}_{ij} = \frac{1}{|\mathcal{A}_i|} \sum_{d \in \mathcal{A}_i} s_{ijd}. \tag{28}$$

The frozen thresholds are $\tau_{\text{mean}} = 2.0$ and $\tau_{\text{dim}} = 2$. An item passes when

$$Q_{ij} = \mathbb{I}\left[\bar{s}_{ij} \geq 2.0 \wedge s_{ijd} \geq 2 \;\; \forall d \in \mathcal{A}_i\right]. \tag{29}$$

The model-level Strict Pass rate is

$$\text{SP}_j = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} Q_{ij}. \tag{30}$$

Answer correctness is not part of this rule. A correct answer does not force a rationale pass, and an incorrect answer does not automatically force a rationale failure. Judge-provided holistic pass recommendations are ignored in favor of the deterministic rule above.

H Extended Results, Sensitivity, and Robustness Analyses

This appendix expands the main-paper result tables without changing their primary interpretation. Evaluation denominators follow Table 3 unless a subsection states otherwise.

H.1 BanduraBench-Core Extended Results

The main paper reports Overall, Structured, Explanation, Strict Pass, and strict Chain scores. Here we decompose the Core results by structured task family, explanation criterion, and chain difficulty.

Model	Bandura F1	Course-design F1	OLSE F1	QM Acc.	Redesign Acc.
GPT Luna	58.3	53.4	95.6	72.2	100.0
Gemini 3.0 Flash	53.6	60.1	91.1	77.8	100.0
GPT-4o	61.4	53.8	90.5	66.7	100.0
GPT-5 mini	59.4	53.0	88.4	66.7	100.0
Claude Haiku 4.5	53.3	51.2	88.9	72.2	99.1
Qwen3-235B-A22B-Thinking	59.5	50.9	90.2	61.1	99.5
DeepSeek-V3	56.2	47.6	89.9	66.7	100.0
Llama 3.3 70B Instruct	62.0	49.1	87.5	61.1	99.1
Phi-4 Reasoning	51.8	49.6	87.6	72.2	95.9
Mistral Small 4	48.2	50.5	89.1	66.7	99.5
Mistral Large 3	56.1	54.3	88.4	55.6	99.5
DeepSeek-R1	54.6	47.5	90.4	61.1	98.6
Qwen3-32B	55.1	48.0	87.0	55.6	100.0
Llama 4 Scout	52.5	44.8	84.5	55.6	93.7
Qwen3.5 Flash	36.6	30.8	60.9	0.0	71.0

Table 16: BanduraBench-Core structured task-family scores on the held-out test set (%). Multi-label families use example-level F1; QM and redesign use exact-match accuracy. Family sizes are Bandura $n = 148$, course-design $n = 124$, OLSE $n = 148$, QM $n = 18$, and redesign $n = 221$.

The package-authoritative Core recomputation is used throughout; in particular, Llama 3.3 70B Instruct has Explanation=82.3 and Overall=73.5.

Across the 15 models, the multi-label families also show different error profiles. Mean precision/recall are 51.6/64.4 for Bandura sources, 42.9/63.6 for course-design clusters, and 83.6/95.2 for OLSE mapping. Exact-set accuracy is 25.3%, 12.6%, and 72.7%, respectively, indicating substantially more over-selection and boundary confusion in the first two families.

Model	Design	Evidence	Bandura	OLSE	Direction	Redesign
GPT Luna	100.0	100.0	68.5	89.2	84.6	96.9
Gemini 3.0 Flash	100.0	100.0	43.8	82.3	85.4	94.6
GPT-4o	100.0	99.2	63.1	80.8	93.1	92.3
GPT-5 mini	100.0	100.0	89.2	80.8	77.7	93.8
Claude Haiku 4.5	99.2	100.0	46.2	90.0	91.5	96.2
Qwen3-235B-A22B-Thinking	96.2	99.2	37.7	85.4	93.8	84.6
DeepSeek-V3	99.2	100.0	40.0	71.5	90.0	85.4
Llama 3.3 70B Instruct	97.7	96.9	32.3	88.5	88.5	90.0
Phi-4 Reasoning	100.0	100.0	33.1	80.8	90.8	90.0
Mistral Small 4	99.2	100.0	39.2	86.9	90.0	87.7
Mistral Large 3	99.2	100.0	43.1	80.8	88.5	83.1
DeepSeek-R1	100.0	100.0	43.8	79.2	91.5	88.5
Qwen3-32B	98.5	99.2	35.4	73.8	91.5	92.3
Llama 4 Scout	98.5	100.0	26.9	82.3	89.2	82.3
Qwen3.5 Flash	68.5	68.5	36.2	55.4	63.1	62.3

Table 17: Core explanation-rubric criterion success rates (%; $n = 130$ explanations per model). The aggregate explanation score and Strict Pass rate are reported in Table 1.

The gap between mean component recovery and strict chain completion shows why Core chain accuracy should be interpreted as a conjunctive endpoint rather than as a normalized measure of local theoretical knowledge.

Model	Mean Component	Easy Strict	Medium Strict	Hard Strict
GPT Luna	85.6	14.3	39.0	0.0
Gemini 3.0 Flash	82.2	14.3	26.8	5.0
GPT-4o	84.5	9.5	29.3	10.0
GPT-5 mini	83.6	0.0	36.6	0.0
Claude Haiku 4.5	81.5	9.5	19.5	0.0
Qwen3-235B-A22B-Thinking	81.2	0.0	17.1	0.0
DeepSeek-V3	79.6	9.5	9.8	0.0
Llama 3.3 70B Instruct	80.2	4.8	4.9	0.0
Phi-4 Reasoning	78.5	4.8	17.1	0.0
Mistral Small 4	81.1	9.5	17.1	0.0
Mistral Large 3	81.9	9.5	22.0	0.0
DeepSeek-R1	79.7	0.0	22.0	0.0
Qwen3-32B	80.2	4.8	9.8	0.0
Llama 4 Scout	77.2	4.8	4.9	0.0
Qwen3.5 Flash	54.2	0.0	19.5	0.0

Table 18: Core mechanism-chain diagnostics (%). Mean Component averages required component scores within eligible linked chains. Strict columns require every component in the tier template to be correct. Each model has 21 easy, 41 medium, and 20 hard eligible chains.

H.2 DesignReasoning Bootstrap and Conjunction-Null Results

Table 19 reports task-family-stratified bootstrap intervals for the two principal structured scores and the 10,000-replicate task-family×reasoning-depth null. The permutation p-value is the one-sided upper-tail probability of observing at least the empirical 4-Exact score under the stratified marginal null. Holm adjustment is applied across the 15 model-level tests.

Model	CompNorm [95% CI]	4-Exact [95% CI]	Null [95% CI]
GPT Luna	84.7 [84.1, 85.4]	50.4 [48.7, 52.2]	47.9 [47.3, 48.5]
Gemini 3.0 Flash	70.8 [70.0, 71.6]	26.4 [24.9, 27.9]	24.7 [23.9, 25.6]
GPT-4o	71.8 [70.9, 72.6]	28.7 [27.1, 30.3]	25.3 [24.5, 26.2]
GPT-5 mini	72.7 [72.0, 73.4]	29.1 [27.5, 30.7]	26.9 [26.1, 27.7]
Claude Haiku 4.5	69.7 [68.9, 70.4]	23.2 [21.7, 24.7]	21.4 [20.6, 22.2]
Qwen3-235B-A22B-Thinking	60.3 [59.4, 61.1]	14.1 [12.8, 15.3]	13.6 [12.8, 14.4]
DeepSeek-R1	62.2 [61.3, 63.1]	15.7 [14.5, 17.0]	15.0 [14.2, 15.9]
Qwen3-32B	55.2 [54.3, 56.1]	9.6 [8.6, 10.7]	9.0 [8.3, 9.8]
Phi-4 Reasoning	54.2 [53.1, 55.3]	13.3 [12.1, 14.6]	8.7 [7.9, 9.4]
Qwen3.5 Flash	70.1 [69.3, 70.9]	24.3 [22.8, 25.8]	22.5 [21.7, 23.3]
DeepSeek-V3	66.4 [65.6, 67.2]	20.0 [18.7, 21.4]	19.1 [18.3, 19.9]
Mistral Large 3	69.9 [69.2, 70.7]	25.6 [24.0, 27.2]	22.7 [21.9, 23.5]
Mistral Small 4	66.1 [65.2, 66.9]	19.0 [17.6, 20.5]	17.0 [16.1, 17.8]
Llama 3.3 70B Instruct	62.6 [61.8, 63.4]	15.3 [14.0, 16.6]	14.8 [14.0, 15.6]
Llama 4 Scout	45.1 [44.1, 46.1]	5.8 [5.0, 6.7]	4.5 [4.0, 5.0]

Table 19: DesignReasoning bootstrap intervals and stratified permutation-null intervals (%).

Ten of the 15 model-level tests remain below .05 after Holm correction. The positive residual is therefore reproducible for many systems, but its magnitude is small relative to the raw CompNorm-to-4-Exact reduction. The null disrupts within-item association, including association induced by shared item difficulty, so the residual is interpreted only as dependence beyond the stratified marginals.

Model	Excess (pp)	Raw p	Holm p
GPT Luna	+2.5	0.0001	0.0015
Gemini 3.0 Flash	+1.7	0.0001	0.0015
GPT-4o	+3.4	0.0001	0.0015
GPT-5 mini	+2.2	0.0001	0.0015
Claude Haiku 4.5	+1.8	0.0002	0.0015
Qwen3-235B-A22B-Thinking	+0.5	0.1494	0.2466
DeepSeek-R1	+0.7	0.0489	0.1956
Qwen3-32B	+0.6	0.0662	0.1986
Phi-4 Reasoning	+4.6	0.0001	0.0015
Qwen3.5 Flash	+1.8	0.0001	0.0015
DeepSeek-V3	+1.0	0.0125	0.0625
Mistral Large 3	+2.9	0.0001	0.0015
Mistral Small 4	+2.1	0.0001	0.0015
Llama 3.3 70B Instruct	+0.5	0.1233	0.2466
Llama 4 Scout	+1.4	0.0001	0.0015

Table 20: Model-level stratified permutation tests. Excess is observed 4-Exact minus the null mean.

H.3 Leave-One-Component-Out Sensitivity

Model	4-Exact	Omit Bandura	Omit OLSE	Omit Course	Omit Causal
GPT Luna	50.4	50.7	52.6	78.8	58.9
Gemini 3.0 Flash	26.4	30.7	35.4	43.2	35.4
GPT-4o	28.7	34.5	39.8	44.9	33.3
GPT-5 mini	29.1	35.5	39.2	49.3	30.3
Claude Haiku 4.5	23.2	28.2	32.5	45.7	27.4
Qwen3-235B-A22B-Thinking	14.1	18.1	20.9	27.1	24.1
DeepSeek-R1	15.7	19.5	23.0	28.3	28.0
Qwen3-32B	9.6	12.8	15.5	20.0	20.9
Phi-4 Reasoning	13.3	19.7	20.0	25.8	17.5
Qwen3.5 Flash	24.3	29.0	35.0	43.5	30.9
DeepSeek-V3	20.0	24.3	29.2	36.2	28.7
Mistral Large 3	25.6	30.9	35.8	45.9	27.3
Mistral Small 4	19.0	24.6	31.2	34.9	23.5
Llama 3.3 70B Instruct	15.3	22.4	24.1	29.5	21.7
Llama 4 Scout	5.8	8.8	8.5	8.6	20.0

Table 21: Leave-one-component-out exact recovery (%). Each omission reports three-component conjunction accuracy on the same 2,554 items; it is a sensitivity diagnostic rather than a causal contribution estimate.

Table 21 provides the complete per-model sensitivity pattern.

H.4 Human-Adjudicated Alternative-Path Sensitivity

The paper-test set contains 1,958 unique-path and 648 multi-admissible items; 634 multi-admissible cases remain after applying the common OLSE eligibility rule. Accepted-path credit is computed against one complete human-approved path at a time. No metric combines nodes or fields from different accepted paths.

Table 22 provides the per-model primary-versus-accepted comparison; aggregate effects are summarized in Section 7.2.

Model	Primary B–OLSE Joint	Accepted B–OLSE Joint	Δ Accepted
GPT Luna	94.0	94.1	0.12
Gemini 3.0 Flash	58.6	61.6	3.01
GPT-4o	52.0	53.6	1.57
GPT-5 mini	51.1	52.6	1.53
Claude Haiku 4.5	51.3	52.9	1.53
Qwen3-235B-A22B-Thinking	45.3	47.7	2.47
DeepSeek-R1	49.1	52.1	3.05
Qwen3-32B	41.0	43.3	2.35
Phi-4 Reasoning	33.6	34.4	0.82
Qwen3.5 Flash	52.5	55.9	3.45
DeepSeek-V3	49.3	50.8	1.53
Mistral Large 3	47.9	49.1	1.17
Mistral Small 4	41.0	43.0	2.04
Llama 3.3 70B Instruct	38.4	39.1	0.67
Llama 4 Scout	37.7	40.4	2.74

Table 22: Sensitivity of the scorer's controlled Bandura–OLSE joint exact measure to human-approved alternative pathways (%). Δ Accepted is the accepted-minus-primary change in percentage points. This is not the four-component 4-Exact metric.

H.5 Open-Ended Rationale Results

Model	TRS	PAS	Strict Pass	Missing required
GPT Luna	89.9	98.7	64.8	0
Gemini 3.0 Flash	84.0	96.7	63.3	0
GPT-4o	82.1	93.3	56.3	0
GPT-5 mini	85.5	98.4	57.3	0
Claude Haiku 4.5	83.8	98.8	57.5	0
Qwen3-235B-A22B-Thinking	79.9	93.1	47.1	0
DeepSeek-R1	80.8	96.6	54.4	0
Qwen3-32B	78.5	87.3	40.7	0
Phi-4 Reasoning	75.7	94.7	44.7	0
Qwen3.5 Flash	79.9	91.0	57.2	0
DeepSeek-V3	82.8	97.6	54.5	0
Mistral Large 3	82.0	96.8	52.3	1
Mistral Small 4	78.4	96.4	46.2	0
Llama 3.3 70B Instruct	77.8	94.3	45.7	0
Llama 4 Scout	77.7	93.2	44.6	0

Table 23: DesignReasoning rationale results (%). TRS and Strict Pass use 2,554 items; PAS uses the 1,048 items with an applicable pedagogical-action dimension. Missing required counts responses with at least one required judge dimension absent before the frozen zero rule.

One Mistral Large 3 response contains a missing required dimension; all other model outputs have complete task-required judge fields. Judge-provided holistic pass recommendations are not used in these aggregates.

H.6 Reasoning-Depth Profiles

Reasoning-depth analyses use the common 2,554-item structured-analysis set. Table 24 reports cross-model means after first computing each model's score within each depth stratum. The profile is non-monotonic over depths 1–4 and falls more sharply at depth 5. Because task family, evidence format, and mechanism composition vary across depth levels, these results are descriptive diagnostics rather than causal estimates of the effect of greater reasoning depth.

The depth-conditioned component profiles further indicate that the aggregate pattern is not produced by a uniform shift in all four fields. Bandura, OLSE, course-design, and causal-claim recovery vary

Depth	n	CompNorm (%)	4-Exact (%)
1	194	71.1	27.0
2	579	66.1	21.1
3	735	66.1	22.7
4	683	67.0	22.9
5	363	57.0	13.4

Table 24: Cross-model DesignReasoning structured recovery by annotated reasoning depth. Means are computed across the 15 target models on the common 2,554-item set. Depth is a structural annotation and is not interpreted as an isolated causal difficulty variable.

differently across the five strata, reinforcing the use of depth as a diagnostic slice rather than a single scalar difficulty ranking.

H.7 Surface-Artifact, Overlap, and Context-Matching Audits

Option length and option-only probes. (Gururangan et al., 2018; Geirhos et al., 2020) Across all 3,475 items, mean gold-option length is 76.65 words / 571.17 characters versus 76.72 words / 570.83 characters for the mean distractor. The corresponding gold-minus-distractor differences are -0.07 words and +0.34 characters.

Length heuristic	Accuracy	95% CI
Longest by words	26.56	25.15–28.01
Shortest by words	28.19	26.77–29.60
Longest by characters	27.68	26.20–29.18
Shortest by characters	26.32	24.85–27.76

Table 25: Tie-adjusted length heuristics on the 3,475-item release (%).

Learned option-only probes train on dev+public validation ($n = 869$) and are evaluated once on paper test ($n = 2,606$). Answer letters and task-family labels are excluded; four-way selection chooses the independently scored option with highest predicted gold probability.

Probe	Balanced Acc.	AUROC	Four-way
Word TF–IDF	49.99	.495	24.60
Character TF–IDF	49.65	.495	25.02
Handcrafted option-only	50.90	.512	27.40

Table 26: Position-invariant option-only probes on paper test (%). Binary chance is 50%; four-way chance is 25%. Bootstrap 95% CIs are [48.92,51.05]/[23.06,26.21], [48.65,50.79]/[23.44,26.59], and [49.73,52.01]/[25.63,29.16] for balanced/four-way accuracy, respectively.

Position, split overlap, and context matching. The diagnostics below separate option-internal balance from signals that use task family, scenario, or question context.

Diagnostic	Result	Additional detail
Global answer-position majority	25.33%	near uniform over A–D
Task-family position majority	40.14%	residual family-level structure
Exact cross-split duplicates	0	scenario, S+Q, and full item
Max cosine: scenario+question	.922	one pair $\geq .90$; none $\geq .95$
Max cosine: full model-facing item	.867	none $\geq .90$
Question→option overlap	45.98%	95% CI: 44.10–47.86

Table 27: Position, cross-split, and context-matching diagnostics. The context-overlap heuristics are distinct from option-only probes because they use scenario or question text.