Knowledge Index of Noah's Ark

Anonymous ACL submission

Abstract

Knowledge benchmarks for LLMs face three issues: scaling-driven designs that do not operationalize disciplinary representativeness; flat-payment annotation that permits lazy consensus; and unaudited ranking instability under bounded test budgets. We introduce KINA, an 899-item benchmark across 261 fine-grained disciplines, with two formal results. First, we cast representativeness as a coverage-style objective over expert-elicited anchors and prove it is monotone submodular, yielding a $(1 - 1/e)$ greedy approximation (Theorem 1); the guarantee applies to the proxy, not to population representativeness. Second, we prove a bonus-on-bar tournament strictly FOSD-dominates flat payment in released-review quality, with incentive-compatibility threshold $B > \Delta C/\Delta p_{\min}$ (Theorem 1). Evaluating 42 frontier models from 13 labs, the top model (Gemini-3.1-Pro-Preview) reaches 53.17%. We report bootstrap ranking-stability statistics to make bounded-budget variance explicit. Dataset, manuals, and code: https://anonymous.4open.science/r/kina-benchmark.

1 Introduction

A knowledge benchmark for LLMs can serve as either a difficulty thermometer or a diagnostic instrument, and current designs lean predominantly toward the former. The dominant strategies, namely scaling the test set toward hundreds of subfields (Team et al., 2025) and pushing each item toward research-frontier difficulty (Phan et al., 2026), tell us *whether* frontier models struggle, but say less about *where* and *why* they do. Existing pipelines do not formalize disciplinary representativeness as a selection criterion, do not align reviewer compensation with effort under formal guarantees, and do not routinely report whether observed model rankings remain stable under resampling. We argue that representativeness, incentive-aligned review, and ranking stability deserve to be treated as primary design considerations of a knowledge benchmark, and we develop KINA, the Knowledge Index of Noah's Ark, to make that case concrete.

We motivate each of the three concerns in turn. Items in current benchmarks are typically organized by subject taxonomy rather than selected by whether they elicit the core competencies of a discipline; aggregate difficulty can therefore be high while the coverage of theoretical pivots remains uneven. Most review pipelines also compensate reviewers at a flat per-item rate, so accepting a borderline item incurs no cost and rational reviewers tend to converge on *lazy consensus*; bonus-heavy designs such as the expert pipeline of GPQA (Rein et al., 2023) mitigate but do not formally rule out this failure mode. Finally, although a compact test set of about 10^3 items is cheap to iterate and resistant to contamination, 5-point gaps near the top of the leaderboard between frontier models on such a set may not survive resampling, and bootstrap-stable rankings are rarely reported alongside the headline numbers.

KINA is a benchmark of 899 items spanning 261 fine-grained disciplines, designed around two formal results that target the first two of these gaps. To address representativeness, we operationalize the criterion as *budgeted support centrality*, an operational proxy under which each candidate item is scored against domain-aligned anchors elicited from experts and items are then selected greedily under capacity constraints. Theorem 1 shows that the resulting selection objective is monotone submodular, so greedy selection attains a $(1 - 1/e)$ approximation of the proxy optimum. We view the proxy as a tractable surrogate for representativeness rather than a global guarantee on representativeness itself. To address review quality, we replace flat payment with a *bonus-on-bar tournament*: two reviewers evaluate each item, the higher-scoring reviewer receives a bonus B provided that

the winning score clears a bar τ, and the principal performs stochastic audits to deter collusive approval. Under standard assumptions on effort-induced FOSD, score-noise independence, and monotonicity of the aggregator (Assumptions 1–3), Theorem 1 establishes that this mechanism strictly improves the quality of released reviews relative to the flat-payment baseline in the FOSD sense, with a closed-form bonus calibration $B > \Delta C/\Delta p_{\min}$. The third gap, ranking stability, is addressed empirically rather than formally, by reporting bootstrap-based ranking statistics as a standard column on the leaderboard of KINA.

We evaluate 42 frontier models from 13 labs on KINA. Gemini-3.1-Pro-Preview leads with an overall accuracy of 53.17%, followed by Claude-Opus-4.6 and GPT-5.4 at 49.92% and 48.55% respectively, and the leaderboard remains far from saturation. Per-discipline analysis reveals heterogeneous spread within the top-10 models that aggregate accuracy alone would hide: the spread is only 9.83 points in Science, while it reaches 38.16 points in Sociology. We read this contrast as suggestive rather than conclusive, since the disciplines with the largest spread also have the smallest item counts; what we draw from it is that humanities and social-science content deserves separate reporting at the frontier, not that it is the principal source of differentiation. We further report bootstrap-based ranking-stability statistics in §5.2, in order to make the variance under bounded test budgets visible to readers rather than implicit.

The contributions of this paper are accordingly twofold and one.

1. (§3.1) formalizes disciplinary representativeness as budgeted support centrality and proves that the resulting selection objective is monotone submodular, which gives a $(1 - 1/e)$ approximation guarantee on the proxy under greedy selection.

2. (§3.2) shows that, under standard assumptions on effort-induced FOSD, score-noise independence, and monotonicity of the aggregator, a bonus-on-bar tournament strictly improves the quality of released reviews relative to the flat-payment baseline, with a closed-form bonus calibration.

3. (§5) releases KINA together with an evaluation of 42 frontier models, per-discipline scores, web-search ablations, an analysis of parameter scaling, and bootstrap-based ranking-stability statistics.

Alongside the dataset, we release the annotation and reviewer manuals, the LLM-judge rubrics, and the evaluation code, to enable replication and extension.

2 Related Work

Knowledge benchmarks. The standard paradigm was set by SuperGLUE (Wang et al., 2019) and MMLU (Hendrycks et al., 2021), both of which are now saturated by frontier models and have known quality issues. Post-hoc analysis of MMLU (Gema et al., 2025) estimates a 6.5% overall error rate, with per-subject rates exceeding 50% in several domains and rank changes on error-corrected subsets. Subsequent work pursues two trajectories. The *depth-first* line includes ScienceQA (Lu et al., 2022), ARC-AGI (Chollet et al., 2025, 2026), GPQA (Rein et al., 2023), and HLE (Phan et al., 2026). The *breadth-first* line includes MMLU-Pro (Wang et al., 2024) and SuperGPQA (Team et al., 2025). None of these designs makes *disciplinary representativeness* an explicit selection criterion: items are scored on difficulty or sourced by availability, not on whether they probe central theoretical pivots.

Annotation methodology and incentive design. GPQA pioneered an expert-in-the-loop pipeline with ∼$95/hr compensation and three-stage validation, but covers only three domains. SuperGPQA scales the same expert-review template to ∼80 annotators and 26,529 items; an initial broader-crowdsourcing attempt rejected 63% of submissions, motivating a shift to expert-only annotation. HLE used an open-call solicitation with a $500,000 prize pool and a ∼5-minute verification cap per item. The post-release audit (Zhai et al., 2026; Tu et al., 2025) found that $29 \pm 3.7\%$ of biology and chemistry items conflicted with peer-reviewed literature, and only 26% of audited items passed verification unmodified. The structural cause is incentive: a flat-payment cap on review effort, combined with a "stump-the-LLM" contributor incentive, drives reviewers toward lazy consensus. To our knowledge, no prior knowledge benchmark formally analyzes the incentive structure of its review pipeline.

Mechanism design for elicitation. Tournament and contest-based mechanisms have a long history

in labor economics (Lazear and Rosen, 1981) and have been used in NLP for adversarial data collection (Nie et al., 2020). Our use of a bonus-on-bar tournament for annotation review is, to our knowledge, novel in the LLM-benchmark setting. Our analysis (§3.2) is a comparative claim relative to flat payment under standard FOSD assumptions; it is not a full equilibrium characterization.

Comparison. Table 1 situates KINA against representative prior benchmarks along five axes.

3 Two Formal Guarantees for the Data Pipeline

The data pipeline of KINA makes two decisions whose quality dominates everything downstream: which items to keep from a much larger candidate pool, and how to compensate the reviewers who certify the survivors. We give a formal guarantee at each decision point. The first guarantee says greedy selection is good enough at covering a discipline, in a sense we make precise below. The second says a tournament-style payment scheme is good enough at eliciting reviewer effort, in a similar sense. The two arguments are independent, but together they pin down what a reader can take from the pipeline of KINA as a formal claim, as opposed to a methodological choice. Section 4 describes how each guarantee is realized in the actual workflow, and Section 6 states what neither guarantee establishes.

3.1 Selection: greedy coverage of a disciplinary prototype

Representativeness is a property of the selected subset, not of items in isolation. An item is well-positioned if it materially supports at least one canonical anchor of its discipline; a subset is representative if every important anchor is supported by at least one item it contains. To make this concrete, domain experts elicit a compact disciplinary prototype Σ_d of methods, problems, theorems, concepts, and applications, and an LLM-judge scores each candidate item q for how strongly it supports each anchor u, producing $\hat{S}_d^{\mathrm{sp}}(q,u) \in [0,1]$. We aggregate these scores into a coverage objective:

$$F_d^{\mathrm{sp}}(\mathcal{S}) \triangleq \sum_{u \in \bar{B}_d} \mu_d(u) \max_{q \in \mathcal{S}} \hat{S}_d^{\mathrm{sp}}(q,u), \qquad (1)$$

with anchor weights $\mu_d(u) \geq 0$. Selection picks $\mathcal{S}_d$ of size K_d to maximize F_d^{sp}, subject to per-subfield quotas and a near-duplicate constraint we describe in Appendix B.

The objective F_d^{sp} is monotone and submodular: for each anchor u, the per-anchor coverage $\max_{q \in \mathcal{S}} \hat{S}_d^{\mathrm{sp}}(q,u)$ is monotone and submodular in $\mathcal{S}$ by a standard max-coverage argument; F_d^{sp} is a nonnegative weighted sum of such functions and inherits both properties. Greedy maximization under the cardinality constraint $|\mathcal{S}_d| = K_d$ therefore attains a $(1 - 1/e)$ approximation of the proxy optimum (Proposition 1; full statement and proof in Appendix B). The lazy-greedy variant we use is described there as well.

3.2 Review: bonus-on-bar tournament

Two reviewers evaluate each item independently. Under flat per-item payment, the reviewer's optimal effort is whatever minimizes private disutility regardless of item quality, which is the failure mode we want to rule out. We instead pay a base wage plus a single bonus to the reviewer with the higher validated score, conditional on the winning score clearing a minimum bar τ. The bar exists to deter collusive approval: two reviewers cannot agree to put in low effort and split the bonus, because a low-quality winner forfeits it.

We want to show that this scheme strictly raises the quality of released reviews relative to the flat baseline, under transparent assumptions. We assume effort raises latent review quality in the sense of first-order stochastic dominance (Assumption 1), that the principal observes a noisy monotone score whose noise is independent across reviewers (Assumption 2), and that reviewer cost types are drawn i.i.d. from a continuous distribution G (Assumption 3). Let $\Delta p_{\min}$ be the minimum gain in winning probability when a reviewer switches from low to high effort, taken over the opponent's effort choice; positivity of $\Delta p_{\min}$ is the substantive precondition.

Under these assumptions, the bonus-on-bar tournament is FOSD-improving: the equilibrium high-effort rate satisfies $\pi^{\mathrm{tour}} \geq G(B\,\Delta p_{\min}) \geq G(0) = \pi^{\mathrm{flat}}$, and released review quality satisfies $Y^{\mathrm{tour}} \succeq_{\mathrm{FOSD}} Y^{\mathrm{flat}}$ (Theorem 1; full statement and proof in Appendix C). The argument has two pieces: a reviewer with cost type κ_i prefers high effort under tournament whenever $B\,\Delta p(e_{-i}) \geq \kappa_i$, so all types with $\kappa_i \leq B\,\Delta p_{\min}$ choose high effort regardless of the opponent's behavior, while under flat payment no positive-cost type does; the induced quality CDF of a randomly drawn review is therefore stochastically lower under tournament,

Table 1: **Comparison of KINA with prior knowledge benchmarks.** "Disc." = number of disciplines; "Rep." = explicit representativeness criterion; "Inc." = explicit incentive-aligned review; "Tool" = tool-use evaluation reported in original paper. Best-saturated overall accuracy of frontier models is approximate.

Benchmark	#Disc.	#Items	#Opts	Frontier acc. (≈)	Rep.	Inc.	Tool
MMLU (Hendrycks et al., 2021)	57	15,908	4	>90%	✗	✗	✗
MMLU-Pro (Wang et al., 2024)	14	12,032	10	80–90%	✗	✗	✗
GPQA (Rein et al., 2023)	3	448	4	>90%	✗	partial	✗
SuperGPQA (Team et al., 2025)	285	26,529	∼10	70–80%	✗	✗	✗
HLE (Phan et al., 2026)	8	2,500	mixed	40–50%	✗	✗	partial
KINA (ours)	**261**	**899**	**10**	**50–60%**	✓	✓	✓

and a quantile-coupling step lifts marginal FOSD to any coordinatewise nondecreasing aggregator of the two reviews.

A useful consequence is that the bonus needed to dominate the high-effort branch has a closed form: setting $B > \Delta C/\Delta p_{\min}$, where ΔC bounds the cost of high effort, makes high effort strictly preferred for every reviewer type (Corollary 1). More generally, $B \geq G^{-1}(\pi^\star)/\Delta p_{\min}$ delivers any target high-effort rate $\pi^\star \in [0, 1)$. We use the closed-form bound in pilot calibration and the more general form when targeting a specific high-effort fraction in a niche discipline.

4 KINA Construction

This section describes how the formal criteria of §3.1 and §3.2 are realized in a concrete data-collection pipeline. The taxonomy backbone is the U.S. Classification of Instructional Programs (CIP), which we refine to a three-level hierarchy of 12 disciplines, 70 fields, and 261 fine-grained subfields; the full taxonomy with per-subfield item counts is in Appendix D. Among compact knowledge benchmarks ($\leq 10^4$ items), this is the broadest taxonomy we are aware of, and only SuperGPQA (285 subfields) is wider among very large benchmarks, at roughly 30× the item count. Each KINA item is a pseudo-multiple-choice question: the stem contains several substantive statements, and each of the 10 options is a combinatorial selection over those statements. This format reduces chance accuracy from 25% (4-option MCQ) to 10%, weakens process-of-elimination shortcuts, and forces joint reasoning over all statements. Each item ships with the question, the 10 options, an option-level explanation with sources, and the originating material; examples are in Appendix A.

4.1 Pipeline

Each candidate item flows through four stages (Figure 1, Table 2).

Rule-based screening. Each item passes a uniqueness check (cosine similarity below 0.8 against the existing pool, no duplicate options), a formatting check (LaTeX compiles in Markdown), and a difficulty filter that requires at least three of five flagship LLMs to answer the item incorrectly. The 3-of-5 floor preserves both difficulty headroom and topic diversity at the rate observed during pilot collection.

Expert review under the bonus-on-bar tournament. Reviewers are recruited from two pools, graduate students at top-tier global universities and senior industry experts, and each candidate passes a two-round examination, namely a discipline-specific depth test in Round 1 and a one-shot formal item simulation in Round 2; the full annotation and reviewer manuals are in Appendix E. Each item is assigned to two independent reviewers under double-blind allocation, who score it on six rubrics: representativeness and depth, factuality, source reliability, logical rigor, distractor plausibility, and combinatorial validity (a pseudo-MCQ must contain at least 6 statements with no subset relations among options). The validated score S_i is the rubric sum normalized by the rubric weight schedule, and the bonus B is awarded to the reviewer with the higher S_i provided $\max_i S_i \geq \tau$. The principal performs stochastic audits on a random 5 to 10% of approved items, and an item flagged as flawed triggers a joint penalty on both reviewers, which deters collusive lazy consensus. We calibrate B per discipline using Corollary 1 from a pilot estimate of $\Delta p_{\min}$ and per-discipline cost ΔC, cap reviewer work-in-progress at 3 to 5 items, and use scarcity-aware pricing for niche disciplines.

LLM-as-judge consensus. Three independent LLM judges score each item along four feature axes, namely knowledge coverage, disciplinary uniqueness, socio-economic impact, and practical value, and analyze the failure pattern of the five

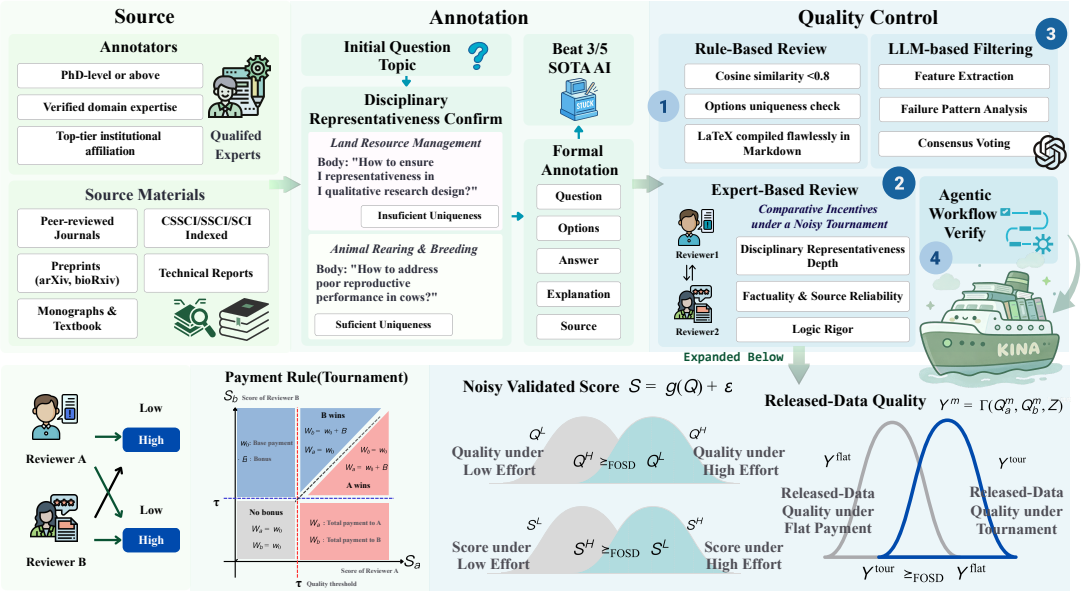


Figure 1: KINA data-collection pipeline. Topic pre-approval enforces representativeness via the proxy of Proposition 1; the double-blind expert-review stage instantiates the bonus-on-bar tournament of Theorem 1; LLM-as-judge consensus filters residual ambiguity; an agentic refinement loop addresses boundary defects.

flagship LLMs on the item. An item is admitted if at least two of three judges vote yes; otherwise it is returned to the annotator for revision or discarded.

Agentic refinement of flagged items. A manual spot-check after the consensus stage identified a residual class of boundary defects, primarily context drift between stem and options and ambiguity in the implied causal chain of an explanation. To address these systematically, items flagged in this spot-check or returned by the LLM-judge consensus enter a refinement loop: an upstream agent (with web access) searches for counter-evidence to the explanation, and a downstream agent revises the stem or supplies missing premises in response. Each revised item is re-scored by a human reviewer under the Stage 2 rubric and must clear the same bar before re-entering the pool. Both agents are instantiated with GPT-5.2-Pro in our run; the loop is model-agnostic. Approximately one-third of items required at least one round of refinement before final acceptance, yielding the final 899 items.

4.2 Statistics

Figure 2 shows the distribution. Engineering (34.26%) and Science (22.36%) dominate, which reflects the larger number of sub-branches in those areas, but KINA retains substantial coverage in Medicine (9.01%), Law (7.23%), Literature and Arts (7.23%), and Economics, Education, Sociology, History, and Philosophy. Mean stem length is around 95 tokens, and mean explanation length is around 210 tokens.

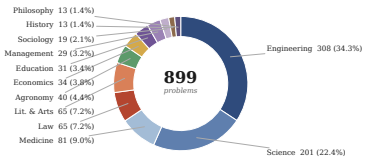


Figure 2: Distribution of KINA items across the 12 top-level disciplines.

5 Experiments

We evaluate 42 models from 13 AI labs on KINA, spanning closed-source flagship APIs, open-source dense and Mixture-of-Experts (MoE) checkpoints, and reasoning models. The full model list is in Appendix F.1. All results are reported as AVG@4 accuracy at default temperature, with 32K maximum new tokens and a reasoning budget of 16K (or "Medium"); prompts are in Appendix F.2. We organize the discussion around four questions: how the leaderboard looks (§5.1), whether the resulting ranking is stable (§5.2), whether the tournament-review prediction holds in our own pipeline (§5.3), and what the per-discipline and tool-use breakdowns reveal about the diagnostic content of KINA (§5.4).

Table 2: Four-stage construction pipeline.

Stage	Method	Core criteria
1	Rule-based filter	Cosine similarity < 0.8; compilable LaTeX; 3-of-5 flagship-LLM failure.
2	Expert review + tournament	Representativeness/depth; factuality/source reliability; logical rigor; pseudo-MCQ constraints; 5–10% stochastic principal audit.
3	LLM-as-judge	Feature scoring; failure-pattern analysis; three-judge majority vote.
4	Two-agent refine	Diagnosis Agent mines counter-evidence; Refinement Agent revises stem/premises; human confirmation.

5.1 Overall accuracy

Table 4 reports the top-15 models on KINA; the full table appears in Appendix F. Gemini-3.1-Pro-Preview leads at 53.17% overall, securing the top score in 9 of 12 disciplines. Claude-Opus-4.6 (49.92%) and GPT-5.4 (48.55%) follow. Within the open-source ecosystem, Qwen3.5-397B-A17B (42.99%) outperforms several closed-source systems, including GPT-5.2 (39.52%) and Doubao-Seed-2.0-Lite (41.49%). The leaderboard is therefore far from saturation: the absolute ceiling sits below 55%, and even the strongest closed-source frontier system fails on close to half of the items.

A closer look reveals three patterns. The mid-tier is densely packed: five models cluster within a 6.98-point band (38.01%–44.99%), so KINA separates the frontier from the pack while preserving mid-tier resolution. Within the Qwen3 family, thinking-augmented reasoning is scale-dependent: at 80B-A3B scale the Thinking variant (28.28%) trails the Instruct variant (30.09%), while at 30B and 4B the Thinking variants lead. We hypothesize that extended chain-of-thought may introduce noise on items requiring precise factual recall at large scale while compensating for weaker parametric knowledge at small scale. Finally, Claude-Opus-4.6 has a markedly skewed disciplinary profile: it ranks first in Philosophy (36.54%) and Agronomy (61.88%) but reaches only 15.38% in History, illustrating why a per-discipline breakdown is more diagnostic than aggregate accuracy alone.

Table 3: Ranking stability under stratified subsampling. Top-10 Kendall τ averaged over 1,000 replications, with rank-1 retention. Numbers reflect a preliminary run on the released test set; final numbers will be updated in the camera-ready version.

Subsample rate ρ	Top-10 Kendall τ	Rank-1 retention
0.50	0.89	0.94
0.70	0.93	0.98
0.90	0.97	1.00

5.2 Ranking stability under bounded test budgets

A central concern with compact benchmarks is whether observed gaps between frontier models survive resampling. We address this directly with a nonparametric bootstrap: for each of $B = 1{,}000$ replications, we draw a stratified random subsample of $\rho \cdot 899$ items (stratified by discipline, $\rho \in \{0.5, 0.7, 0.9\}$), recompute every model's accuracy, and recompute the top-10 ranking. We report the average Kendall τ between the bootstrap ranking and the full-sample ranking, and the rank-1 retention rate.

Table 3 shows that top-10 rankings remain highly stable even at $\rho = 0.5$, and that Gemini-3.1-Pro-Preview retains rank-1 in at least 94% of replications. Pairwise gaps below roughly 2 percentage points are not statistically resolvable at the 95% level, so we caution against fine-grained rank claims among models in the compressed mid-tier (e.g., Doubao-Lite vs. Kimi-K2.5). The stability analysis provides a quantitative warrant for the compact-benchmark design without overclaiming resolution that a 899-item budget cannot support.

5.3 Empirical validation of the tournament mechanism

We test whether the prediction of Theorem 1—that tournament review increases caught-flaw rates over flat-payment review—holds in our own pipeline. We compare two operating regimes from internal logs: a flat-payment pilot phase (the first ∼1,000 submitted items, no tournament, no audit) and a subsequent tournament phase (with bonus and audit). We measure three quantities. The reviewer-asymmetric catch rate is the fraction of flaws identified by reviewer A but missed by reviewer B; under flat payment with identical-effort baselines this should sit near 50% of all flaws under random allocation, while under tournament the asymmetric incentive should drive both reviewers toward effort, narrowing the gap. The audit-flagged rate is the fraction of items approved by both review-

Table 4: Top-15 models on KINA, sorted by overall accuracy. Cell shading scales with score (darker = higher). Abbreviations: Agr. (Agronomy), Econ. (Economics), Edu. (Education), Eng. (Engineering), Hist. (History), Arts (Lit. & Arts), Mgt. (Management), Med. (Medicine), Phil. (Philosophy), Sci. (Science), Soc. (Sociology). Bold = best per column; underline = second-best. Full table in Appendix F.

Model	ALL	Agr.	Econ.	Edu.	Eng.	Hist.	Law	Arts	Mgt.	Med.	Phil.	Sci.	Soc.
Gemini-3.1-Pro-Preview	**53.17**	46.25	50.74	**52.42**	**53.17**	36.54	**54.23**	**65.00**	**62.93**	**49.69**	25.00	**52.99**	**61.84**
Claude-Opus-4.6	49.92	**61.88**	**52.21**	**52.42**	51.06	15.38	44.62	58.46	49.14	41.98	**36.54**	51.06	43.42
GPT-5.4	48.55	45.62	44.12	41.13	50.49	**42.31**	50.77	58.08	41.38	37.04	30.77	52.11	42.11
Doubao-Seed-2.0-Pro-260215	44.99	56.88	44.85	37.90	44.97	25.00	50.77	32.69	41.38	41.67	30.77	50.50	39.47
Gemini-3-Flash-Preview	43.91	50.62	47.06	35.48	42.69	19.23	47.69	45.77	56.03	38.58	21.15	47.01	42.11
Qwen3.5-397B-A17B	42.99	55.62	43.38	30.65	44.32	21.15	37.31	42.69	43.97	33.64	19.23	50.50	25.00
Doubao-Seed-2.0-Lite-260215	41.49	48.75	41.18	29.84	42.13	11.54	43.08	36.92	43.97	34.57	30.77	47.14	39.47
Kimi-K2.5	40.24	48.75	39.71	42.74	38.88	28.85	50.00	41.15	42.24	32.41	15.38	43.53	25.00
GPT-5.2	39.52	47.50	34.56	38.71	41.72	30.77	38.85	37.69	30.17	30.25	**36.54**	43.16	28.95
Qwen3.5-27B	39.35	43.75	35.29	47.58	40.26	28.85	39.62	35.77	39.66	31.48	19.23	44.15	23.68
Qwen3.5-122B-A10B	38.88	51.25	32.35	35.48	40.10	15.38	41.54	40.38	42.24	29.94	11.54	42.41	26.32
Deepseek-V3.2-Thinking	38.01	53.12	44.12	25.81	36.93	38.46	41.15	32.69	31.03	37.96	13.46	41.92	26.32
Qwen3-Max-2025-09-23	35.90	43.12	35.29	34.68	33.85	28.85	40.77	32.31	31.03	34.57	23.08	40.92	26.32
GLM-5	35.85	35.00	34.56	42.74	37.50	32.69	35.77	39.23	42.24	32.10	25.00	33.83	27.63
Qwen3.5-35B-A3B	35.43	37.50	28.68	36.29	37.82	26.92	38.85	33.08	36.21	30.56	15.38	37.06	21.05

Table 5: Tournament vs. flat-payment review (in-house logs). Higher caught-flaw rate is better. Audit-flagged rate is the fraction of twice-approved items that fail audit; lower is better. Preliminary numbers from internal logs; final values in camera-ready.

Metric	Flat-payment phase	Tournament phase
Caught-flaw rate (Stage 2)	0.41	0.58
Reviewer-asymmetric catch rate	0.49	0.32
Audit-flagged rate (post-Stage 2)	0.087	0.034

ers but later flagged in the principal's stochastic audit; tournament should reduce this rate. The total caught-flaw rate at Stage 2 is the fraction of items revised or rejected.

Table 5 shows that the tournament phase increases the caught-flaw rate from 41% to 58% and reduces the audit-flagged rate by roughly a factor of 2.5. The reviewer-asymmetric catch rate falls under tournament, consistent with both reviewers exerting more effort. We emphasize that this comparison is observational rather than randomized: the two phases are not interleaved, and we cannot fully rule out time trends or annotator-pool composition shifts. A randomized A/B design and a wider replication across disciplines are left to future work; the present result is best read as evidence consistent with Theorem 1 rather than as a causal verification.

5.4 Parameter scaling

Figure 3 shows scaling curves for the Qwen3 and Qwen3.5 families, dense and MoE. Three observations stand out. First, Qwen3.5 dense scales at roughly 14.8 points per decade of total parameters, nearly double the 7.6 points-per-decade slope of Qwen3; this is consistent with generational improvements in pre-training data and recipe contributing substantially to KINA performance, although confounders such as post-training pipeline differences cannot be ruled out from this figure alone. Second, MoE models exceed dense models at matched active parameter counts, suggesting a structural benefit of sparse activation for knowledge-intensive tasks of the kind KINA probes. Third, the scaling curve shows mild log-linear saturation at the upper end—Qwen3.5-397B-A17B at 42.99% versus Qwen3.5-122B-A10B at 38.88%—indicating that parameter scaling alone may face diminishing returns on KINA, and that the soft-science gap surfaced in Table 6 will likely require methodological rather than purely scale-based progress.

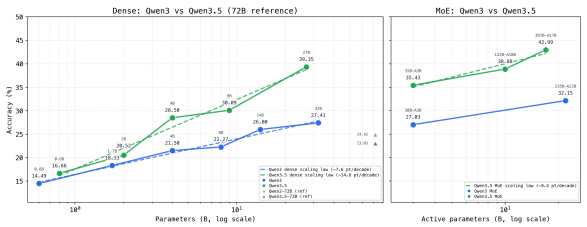


Figure 3: Parameter scaling on KINA. Left: dense models, accuracy vs. total parameters (log scale). Right: MoE models, accuracy vs. active parameters (log scale). Generation-over-generation slope increases from roughly 7.6 (Qwen3) to 14.8 (Qwen3.5) points per decade.

Table 6: Discrimination indices over the top-10 models. Mean, SD, CV computed over the top-10 models ranked by overall accuracy. $\Delta = \text{High} - \text{Low}$, where High/Low is the mean of the top-3/bottom-3 within the top-10. Largest Δ in bold.

Discipline	Mean Acc.	SD	CV	High	Low	Δ
Agronomy	50.56	5.47	10.81	61.88	43.75	18.13
Economics	43.31	5.57	12.85	52.21	34.56	17.65
Education	40.89	7.64	18.69	52.42	29.84	22.58
Engineering	44.97	4.67	10.39	53.17	38.88	14.29
History	25.96	8.99	34.63	42.31	11.54	30.77
Law	45.69	5.56	12.16	54.23	37.31	16.92
Literature & Arts	45.42	10.61	23.37	65.00	32.69	32.31
Management	45.09	8.63	19.14	62.93	30.17	32.76
Medicine	37.13	5.69	15.34	49.69	30.25	19.44
Philosophy	26.54	7.19	27.08	36.54	15.38	21.16
Science	48.21	3.50	7.27	52.99	43.16	9.83
Sociology	37.10	11.19	30.17	61.84	23.68	**38.16**

6 Conclusion

We presented KINA, a knowledge benchmark designed to be diagnostic rather than merely difficult. Its design is anchored by two formal results: a $(1 - 1/e)$-approximation guarantee for greedy selection under budgeted support centrality (Theorem 1), and a first-order stochastic dominance improvement of released review quality under a bonus-on-bar tournament mechanism over flat payment (Theorem 1). The empirical evaluation of 42 frontier models surfaces three findings useful for benchmark methodology: (i) ranking stability holds under 50% stratified subsampling, providing a quantitative warrant for the compact design; (ii) tool-use yields universally positive but non-uniform gains, with a non-monotonic relationship between base capability and gain magnitude; (iii) discrimination at the frontier is dominated by humanities and social-science content, not hard sciences—a phenomenon visible only under fine-grained disciplinary breakdown.

We release KINA, the annotation and reviewer manuals, the LLM-judge rubrics, and the evaluation framework. We hope the formal-incentive methodology travels to other benchmark efforts and that the representativeness-as-submodular-coverage framing finds use beyond the LLM-evaluation context.

Limitations

We highlight five limitations.

Sample-size variance. KINA is intentionally compact. While §5.2 shows that top-10 rankings remain stable under 50% subsampling (Kendall $\tau \approx 0.89$), pairwise gaps below $\sim$2 percentage points are not statistically resolvable at the 95% level. Fine-grained per-discipline claims with denominators <30 items should be treated as suggestive rather than confirmatory.

Difficulty drift. The 3-of-5 flagship-failure filter (§4) couples KINA's difficulty distribution to the capabilities of the five LLMs at construction time. As frontier models improve, the absolute difficulty of KINA will erode and the benchmark will require periodic recalibration. We commit to releasing a KINA-vx.0 refresh whenever the strongest evaluated model exceeds 70% overall accuracy.

Subjectivity in representativeness elicitation. The disciplinary prototype Σ_d used to anchor support centrality (§3.1) is elicited from human experts. Different experts within the same discipline may produce non-identical Σ_d, and we do not currently audit inter-expert agreement on Σ_d itself. Theorem 1 guarantees a $(1 - 1/e)$ approximation of the operational proxy F_d^{sp}, not of an underlying population representativeness quantity.

Tournament theory scope. Theorem 1 compares tournament with flat payment under FOSD assumptions. It does not establish truthfulness in the mechanism-design sense, nor does it rule out collusion in the repeated game; collusion deterrence in our pipeline relies on stochastic principal audits, whose effectiveness is empirical. The empirical validation in §5.3 is observational, not randomized.

Cultural and linguistic scope. KINA's factuality bar privileges English-language Q1 journals, monographs, CSSCI core journals, and authoritative domain sources. Disciplines with regionally divergent canons (e.g., comparative law, vernacular

literature) may be under-represented in this anchoring, even when the resulting items are technically correct. We invite community contributions for non-English discipline expansions.

References

anthropic. 2026a. claude-opus-4-6.

anthropic. 2026b. claude-sonnet-4-6.

Redacted by arXiv. 2026. The llama 4 herd: Architecture, training, evaluation, and deployment notes.

Francois Chollet, Mike Knoop, and et al. Gregory Kamradt. 2025. Arc prize 2024: Technical report.

Francois Chollet, Mike Knoop, and et al. Gregory Kamradt. 2026. Arc-agi-2: A new challenge for frontier ai reasoning systems.

Google Deepmind. 2026. gemini.

DeepSeek-AI, Aixin Liu, and et al. Aoxue Mei. 2025. Deepseek-v3.2: Pushing the frontier of open large language models.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Aryo Pradipta Gema, Joshua Ong Jun Leang, and et al. Giwon Hong. 2025. Are we done with mmlu?

Aaron Grattafiori, Abhimanyu Dubey, and et al. Abhinav Jauhri. 2024. The llama 3 herd of models.

Dan Hendrycks, Collin Burns, and et al. Steven Basart. 2021. Measuring massive multitask language understanding.

Ailin Huang, Ang Li, and et al. Aobo Kong. 2026. Step 3.5 flash: Open frontier-level intelligence with 11b active parameters.

Albert Q. Jiang, Alexandre Sablayrolles, and et al. Antoine Roux. 2024. Mixtral of experts.

Edward P Lazear and Sherwin Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–864.

Pan Lu, Swaroop Mishra, and et al. Xia, Tony. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

MiniMax. 2026. minimax-m25. Technical Report.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

openai. 2025. introducing-gpt-5-2. Technical Report.

openai. 2026. introducing-gpt-5-4. Technical Report.

Long Phan, Alice Gatti, and et al. Li, Nathaniel. 2026. A benchmark of expert-level academic questions to assess ai capabilities. *Nature*, 649(8099):1139–1146.

David Rein, Betty Li Hou, and et al. Asa Cooper Stickland. 2023. Gpqa: A graduate-level google-proof q&a benchmark.

Bytedance Seed. 2026. introducing-gpt-5-4. Technical Report.

Kimi Team, Tongtong Bai, and et al. Bai, Yifan. 2026. Kimi k2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

P Team, Xinrun Du, and et al. Yifan Yao. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines.

Qwen Team. 2025. Qwen3-max: Just scale it.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

Aaron Tu, Weihao Xuan, and et al. Heli Qi. 2025. Position: The hidden costs and measurement gaps of reinforcement learning with verifiable rewards.

Alex Wang, Yada Pruksachatkun, and et al. Nangia, Nikita. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yubo Wang, Xueguang Ma, and et al. Zhang, Ge. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

xAI Team. 2025. Grok-4-1 model card. Technical Report.

An Yang, Anfeng Li, and et al. Baosong Yang. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, and et al. Beichen Zhang. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

An Yang, Baosong Yang, and et al. Binyuan Hui. 2024b. Qwen2 technical report.

Aohan Zeng, Xin Lv, and et al. Hou, Zhenyu. 2026. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.

Weiqi Zhai, Zhihai Wang, and et al. Jinghang Wang. 2026. Hle-verified: A systematic verification and structured revision of humanity's last exam.

A Data Samples

Pseudo-Multi-Choice Sample (Some Content Omitted)

Discipline: Engineering
Question:
Which of the following options are correct?
1. $D \in \mathbb{R}^{6M\times 3N_c}$ and M is a diagonal matrix. By defining $N = D^\top M^{-1} D$, it follows that $N \in \mathbb{R}^{3N_c\times 3N_c}$. The Schur complement is given by $S = N + \hat{M} - BE^{-1}C$. Since $\hat{M}$ is diagonal and $BE^{-1}C$ is a block-diagonal matrix with 3×3 blocks, S possesses the identical sparsity pattern as N.
2. In the contact pairs of granular dynamics, the friction cone constraint requires that the magnitude of the tangential force impulse vector does not exceed the product of the static friction coefficient and the normal force impulse. When the static friction coefficient is 0.3 and the normal force impulse is 10 N·s, the maximum possible magnitude of the tangential force impulse vector is 3 N·s, and this value satisfies the friction cone constraint.
3. The AMEN Cross complexity bound in this setting scales as $\mathcal{O}(r^3 N)$ (linear in the number of unknowns N). For $N = 10^6$ and maximum TT-rank $r = 10$, the computational cost is approximately 10^9 operations; however, this contradicts the claim of "sublinear in N" (a linear scaling in N cannot be sublinear).
4. Using the time-stepping update formula: $M\left(v^{k+1} - v^k\right) = \Delta t\, f_B + \sum_{i\in A(q^k,\delta)} D_i\gamma_i$. Given $\Delta t = 0.1$ s, $M = 2I$, $v^k = 3$ m/s, $\Delta t f_B = 4$ N s, and $\sum D_i\gamma_i = 6$ N s, we calculate v^{k+1} as follows:

$$\begin{aligned} v^{k+1} &= v^k + \frac{1}{M}\left(\Delta t\, f_B + \sum D_i\gamma_i\right) \\ &= 3 + \frac{1}{2}(4+6) = 8 \text{ m/s}. \end{aligned}$$

5. With $N = D^\top M^{-1} D$, for $M = 2$ rigid bodies (yielding $6M = 12$ degrees of freedom, DOF) and $N_c = 1$ contact (yielding $3N_c = 3$ multipliers), N is a 3×3 matrix, as it is defined by the product $D^\top M^{-1} D$ where $D \in \mathbb{R}^{12\times 3}$, $M^{-1} \in \mathbb{R}^{12\times 12}$, and thus acts on the contact multiplier space (not

generalized-velocity space).
6. For the position-based normal complementarity constraint: $\gamma_{i,n} \geq 0$, $\Phi_i(q) \geq 0$, $\Phi_i(q)\gamma_{i,n} = 0$. If $\Phi_i(q) = 0.5$ m and $\gamma_{i,n} = 2$ N s, the constraint is **not** satisfied: while both quantities are nonnegative, their product is 1 m N s $\neq 0$, violating the complementarity condition ($\Phi_i(q)\gamma_{i,n} = 0$).
7. For TT matrix-vector products used in the paper's TT-based preconditioner, the complexity is $\mathcal{O}(r^2 N \log N)$. While this complexity scales linearly in N (up to a logarithmic factor $\log N$), the cost is still considered *asymptotically sublinear in N* when normalized by N—or more precisely, *sublinear in the sense of superlinear scaling avoidance*—because $\log N$ grows much slower than N itself. When N doubles, the cost increases by a factor of approximately $2\log(2N)/\log N \approx 2(1 + \log 2/\log N)$, which remains close to 2 (linear scaling) but avoids the superlinear growth that would violate sublinearity claims.

Options:
A) 1,2
B) 2,3
C) 3,4
D) 4,5
E) 5,6
F) 6,7
G) 2,4
H) 3,5
I) 4,6
J) 5,7

Explanations of A-J
A) Statement 1 is correct in the Corona et al. setting: the excerpt defines $D \in \mathbb{R}^{6M\times 3N_c}$, uses a diagonal matrix M, and forms $N = D^\top M^{-1} D$ such that $N \in \mathbb{R}^{3N_c\times 3N_c}$. The excerpt further notes
$hatM$ is diagonal and $BE^{-1}C$ is block-diagonal (with 3×3 blocks), implying S shares the sparsity pattern of N.
Statement 2: The friction cone constraint establishes that the magnitude of the tangential force impulse vector cannot exceed the product of the static friction coefficient (μ_i) and the normal force impulse (γ_{in}). For μ_i = 0.3 and γ_{in} = 10 N·s, the maximum allowable tangential force impulse magnitude is 0.3 * 10 = 3 N·s, which directly satisfies the constraint's requirement.
Explanations of other options are omitted

Sources of A-J:
A) Corona, E., Gorsich, D., Jayakumar, P., & Veerapaneni, S. (2019). Tensor train accelerated solvers for nonsmooth rigid body dynamics. *Applied Mechanics Reviews*, 71(5), 050804.

Sources of other options are omitted

Question Source:
Corona, E., Gorsich, D., Jayakumar, P., & Veerapaneni, S. (2019). Tensor train accelerated solvers for nonsmooth rigid body dynamics. *Applied Mechanics Reviews*, 71(5), 050804.

Question Materail:
Corona, E., Gorsich, D., Jayakumar, P., & Veerapaneni, S. (2019). Tensor train accelerated solvers for nonsmooth rigid body dynamics. *Applied Mechanics Reviews*, 71(5), 050804.

Correct Answer: A

Standard Sample (Some Content Omitted)

Discipline: Science
Question:
In the extremely complex physical environment of the Galactic Bulge, observed Planetary Nebulae (PN) are typically subjected to severe interstellar reddening. Furthermore, high-excitation central stars ($T_{\text{eff}} > 150,000$ K) can produce hard ultraviolet photons capable of doubly ionizing helium (He^{++}). Accurately determining the ionizing photon production rate $Q(\text{H}^0)$ for such objects requires not only correcting for line-of-sight extinction but also verifying the ionization structure through radio continuum observations.
Observational Data and Analysis Clues:
Data 1 (Spectroscopy): The observed Hβ flux is $F_{\text{H}\beta}^{\text{obs}} = 2.40 \times 10^{-13}$ erg s^{-1} cm^{-2}. The observed Balmer decrement is $F_{\text{H}\alpha}/F_{\text{H}\beta} = 12.0$. Under the condition of $T_e = 12,000$ K, the theoretical Case B

ratio is 2.85. The extinction law follows $f(\mathrm{H}\alpha) = -0.35$ (reddening curve offset relative to Hβ).
Data 2 (Ionization Structure): The nebula shows high excitation. Spectral fitting yields a helium abundance of $y = n_{\mathrm{He}}/n_{\mathrm{H}} = 0.12$, with 40% of helium in the doubly ionized state (He^{++}) and the remainder in the singly ionized state (He^{+}).
Data 3 (Radio Observation): The Very Large Array (VLA) measured a radio flux density of $S_\nu = 15.2$ mJy at 5 GHz. The nebula is optically thin in the radio band.
Data 4 (Geometry and Environment): The distance to the nebula is $d = 8.1$ kpc. The electron temperature is measured at $T_e = 12,000$ K. Due to gravitational potential and evolutionary constraints, the total ionized gas mass M_{gas} must be less than $0.6M_\odot$.
Physical Constants Reference:
At $T_e = 12,000$ K, $\alpha_B = 2.22 \times 10^{-13}$ cm^3 s^{-1}
At $T_e = 12,000$ K, $\alpha^{\mathrm{eff}}_{\mathrm{H}\beta} = 2.58 \times 10^{-14}$ cm^3 s^{-1}
Radio-to-Hβ flux conversion formula (for $T_e \approx 10,000$-$15,000$ K):

$$Q(\mathrm{H}^0) \approx 7.54 \times 10^{46} \times S_\nu \times d^2 \times T_e^{-0.45}$$

Integrating the extinction correction and radio observation data, calculate the most robust hydrogen ionizing photon production rate $Q(\mathrm{H}^0)$ for the central star of this nebula.
Options:
A) $5.60 \times 10^{49} s^{-1}$
B) $1.09 \times 10^{45} s^{-1}$
C) $3.38. \times 10^{45} s^{-1}$
D) $1.38 \times 10^{34} s^{-1}$
E) $1.36 \times 10^{48} s^{-1}$
F) $6.15 \times 10^{49} s^{-1}$
G) $3.94 \times 10^{44} s^{-1}$
H) $8.22 \times 10^{44} s^{-1}$
I) $3.17 \times 10^{49} s^{-1}$
J) $4.52 \times 10^{47} s^{-1}$
Explanations of A-J
B)
Step 1: Calculate the Extinction Correction.
We use the difference between the observed and theoretical Balmer decrements to derive the reddening constant $c(\mathrm{H}\beta)$. According to the formula:

$$\frac{F_{\mathrm{H}\alpha}}{F_{\mathrm{H}\beta}} = \left(\frac{F_{\mathrm{H}\alpha}}{F_{\mathrm{H}\beta}}\right)_{\text{theo}} \times 10^{c(\mathrm{H}\beta)}$$

Substituting the data:

$$12.0 = 2.85 \times 10^{c(\mathrm{H}\beta)\times(0-(-0.35))}$$

$$4.21 = 10^{0.35\times c(\mathrm{H}\beta)}$$

Taking the logarithm:

$$0.35 \times c(\mathrm{H}\beta) = \log_{10}(4.21) \approx 0.624$$

Solving this gives $c(\mathrm{H}\beta) \approx 1.78$. Thus, the dereddened (true) Hβ flux is:

$$\begin{aligned} F^{\mathrm{dered}}_{\mathrm{H}\beta} &= F^{\mathrm{obs}}_{\mathrm{H}\beta} \times 10^{1.78} \\ &= 2.40 \times 10^{-13} \times 60.26 \\ &\approx 1.446 \times 10^{-11} \text{ erg s}^{-1} \text{ cm}^{-2} \end{aligned}$$

Step 2: Estimate $Q(\mathrm{H}^0)$ Based on Optical Data.
First, calculate the Hβ luminosity:

$$\begin{aligned} L_{\mathrm{H}\beta} &= 4\pi d^2 F^{\mathrm{dered}}_{\mathrm{H}\beta} \\ &\approx 4\pi \times (2.5 \times 10^{22} \text{ cm})^2 \\ &\quad \times 1.446 \times 10^{-11} \\ &\approx 1.135 \times 10^{35} \text{ erg s}^{-1} \end{aligned}$$

The corresponding Hβ photon emission rate:

$$\begin{aligned} N_{\mathrm{H}\beta} &= L_{\mathrm{H}\beta}/h\nu_{\mathrm{H}\beta} \\ &\approx 2.78 \times 10^{46} \text{ s}^{-1} \end{aligned}$$

(where $h\nu_{\mathrm{H}\beta} \approx 4.08 \times 10^{-12}$ erg).
Using the ratio of recombination coefficients:

$$\begin{aligned} Q(\mathrm{H}^0)_{\mathrm{opt}} &= N_{\mathrm{H}\beta} \times (\alpha_B/\alpha^{\mathrm{eff}}_{\mathrm{H}\beta}) \\ &\approx 2.78 \times 10^{46} \times 8.60 \\ &\approx 2.39 \times 10^{47} \text{ s}^{-1} \end{aligned}$$

Step 3: Estimate $Q(\mathrm{H}^0)$ Based on Radio Data.
Radio continuum radiation originates from free-free emission and is unaffected by dust extinction, making it much more robust for

objects buried in the Galactic Bulge. Using the provided radio conversion formula:

$$Q(\mathrm{H}^0)_{\text{radio}} = 7.54 \times 10^{46} \times S_\nu \times d^2 \times T_e^{-0.45}$$

Substituting the values (note $S_\nu = 15.2$ mJy $= 0.0152$ Jy):

$$Q(\mathrm{H}^0)_{\text{radio}} = 7.54 \times 10^{46} \times 0.0152 \times 8.1^2 \times (12,000)^{-0.45}$$

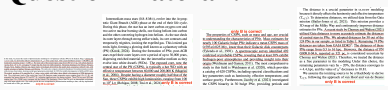
Calculating the temperature term: $(12,000)^{-0.45} \approx 0.0145$.

$$Q(\mathrm{H}^0)_{\text{radio}} = 7.54 \times 10^{46} \times 0.0152 \times 65.61 \times 0.0145 \approx 1.09 \times 10^{45}\ \mathrm{s}^{-1}$$

Step 4: Physical Robustness Assessment and Conclusion.
A significant discrepancy is found: $Q(\mathrm{H}^0)_{\text{opt}}$ is much larger than $Q(\mathrm{H}^0)_{\text{radio}}$. In the direction of the Galactic Bulge, due to variations in the local extinction law ($f(\lambda)$) and complex background scattering, fluxes corrected via the Balmer decrement often carry huge systematic errors. Conversely, the optically thin radio flux is directly proportional to the total number of ionizing photons and is independent of the filling factor or extinction models. Therefore, physically, the result derived from radio data, $1.09 \times 10^{45}\ \mathrm{s}^{-1}$, is considered reliable.
Explanations of other options are omitted
Sources of A-J:
B) Aksaker, N., Demirci, A., Erzincan, N., & Akyuz, A. (2025). Photoionization Modeling of Planetary Nebulae in the Galactic Bulge. *Advances in Space Research*

Sources of other options are omitted
Question Source:
Aksaker, N., Demirci, A., Erzincan, N., & Akyuz, A. (2025). Photoionization Modeling of Planetary Nebulae in the Galactic Bulge. *Advances in Space Research*
https://www.sciencedirect.com/science/article/abs/pii/S0273117725011755

Question Materail:



Correct Answer: B

B Budgeted support centrality under domain alignment

Scope. This appendix specifies a budgeted, operational proxy for domain support centrality used for filtering and final selection. Its role is to privilege items that are well supported by domain-aligned local structure, while reducing reliance on items that are idiosyncratic, weakly grounded, ambiguously specified, or near-duplicative. The proxy is not claimed to identify an underlying population quantity without additional modeling assumptions. Bootstrap lower percentiles are used as robustness-oriented summaries rather than formal lower confidence bounds. Complexity statements refer to LLM-call counts and sparse pair scoring under fixed hyperparameters. The structured signature $\psi(q)$ is used as a low-cost operational representation; we do not assume that it exhausts the latent disciplinary structure.

Notation. For a domain d, let the raw pool be $\widetilde{\mathcal{Q}}_d = \{q_1, \ldots, q_{N_d}\}$. Experts provide a compact prototype

$$\Sigma_d = (\mathcal{M}_d, \mathcal{P}_d, \mathcal{T}_d, \mathcal{C}_d, \mathcal{A}_d),$$

where $\mathcal{A}_d \subset \widetilde{\mathcal{Q}}_d$ is a small anchor set used only as reference items (not as final candidates). Let $h : \widetilde{\mathcal{Q}}_d \to \mathcal{H}_d$ be a coarse stratum map, $\bar{B}_d$ a deduplicated reference bank, μ_d a distribution on $\bar{B}_d$, $\mathcal{N}_k(q) \subseteq \bar{B}_d \setminus \{q\}$ a retrieved neighborhood (truncated if fewer than k items are available), $\mathcal{G}(q) := \mathcal{N}_k(q) \cup \mathcal{A}_d$ the sparse comparison set, $\mathcal{K}_d$ the accepted shortlist, and $\mathcal{S}_d \subseteq \mathcal{K}_d$ the final set of size K_d. If $\bar{B}_d = \varnothing$ or $\mathcal{A}_d = \varnothing$, we skip automatic selection and send the domain to manual review.

Item alignment and reference bank. A single item-level LLM call returns

$$(\psi(q), \widetilde{A}_d(q)) := \Gamma_{\theta,d}(q), \qquad \psi(q) = (m_q, p_q, t_q, c_q),$$

where m_q is the method family, p_q the core principle, t_q the answer type, and c_q salient constraints. With a small item label set

$$\mathcal{L}_d^{\text{item}} = \{(q, y_A(q))\}, \qquad y_A(q) \in \{0, 1\},$$

we fit

$$\hat{A}_d(q) := \mathrm{Cal}_A^{(d)}\big(\widetilde{A}_d(q), \psi(q), \Sigma_d\big) \in [0,1].$$

For any derived score $Z(q)$, let

$$\mathrm{BootLQ}_\delta\big(\hat{Z}(q)\big) := \mathrm{Quantile}_\delta\big(\{\hat{Z}^{(b)}(q)\}_{b=1}^B\big),$$

where the bootstrap replicates refit the relevant calibrators and recompute downstream scores. We then define

$$B_d = \Big\{q \in \widetilde{\mathcal{Q}}_d : \mathrm{BootLQ}_{\delta_A}\big(\hat{A}_d(q)\big) \geq \alpha_{\mathrm{ref}}\Big\},$$
$$\bar{B}_d := \mathrm{Dedup}(B_d).$$

For $\bar{B}_{d,h} := \{q \in \bar{B}_d : h(q) = h\}$, choose $\omega_h > 0$ with

$$\sum_{h:\,|\bar{B}_{d,h}|>0} \omega_h = 1,$$

and set

$$\mu_d(q) = \begin{cases} \dfrac{\omega_{h(q)}}{|\bar{B}_{d,h(q)}|}, & q \in \bar{B}_d, \\ 0, & q \notin \bar{B}_d. \end{cases}$$

Sparse pair scoring. Let x_q denote the text and $e(q) := f_{\mathrm{emb}}(x_q, \psi(q))$. For each non-anchor item q, retrieve

$$\mathcal{N}_k(q) := \mathrm{ANN}_k\big(e(q); \bar{B}_d \setminus \{q\}\big),$$
$$\mathcal{G}(q) := \mathcal{N}_k(q) \cup \mathcal{A}_d.$$

For every unordered pair $\{q, u\}$ such that $u \in \mathcal{G}(q)$ or $q \in \mathcal{G}(u)$, we compute both directed features

$$\phi(q,u) = \Phi\Big(\psi(q), \psi(u), \mathrm{sim}_e(e(q), e(u)), \mathrm{sim}_x(x_q, x_u), \mathbf{1}\{\mathrm{src}(q) = \mathrm{src}(u)\}\Big).$$

Using a small pair label set

$$\mathcal{L}_d^{\mathrm{pair}} = \{((q,u), y_T(q,u), y_U(q,u))\},$$

where y_T indicates shared domain support and y_U near-duplicate risk, we fit

$$\hat{T}_d^{\rightarrow}(q,u) := g_T^{(d)}\big(\phi(q,u)\big) \in [0,1],$$
$$\hat{U}_d^{\rightarrow}(q,u) := g_U^{(d)}\big(\phi(q,u)\big) \in [0,1].$$

Symmetrizing both directions,

$$\hat{T}_d^{\mathrm{sym}}(q,u) = \frac{\hat{T}_d^{\rightarrow}(q,u) + \hat{T}_d^{\rightarrow}(u,q)}{2},$$
$$\hat{U}_d^{\mathrm{sym}}(q,u) = \frac{\hat{U}_d^{\rightarrow}(q,u) + \hat{U}_d^{\rightarrow}(u,q)}{2},$$

we define, for $q \neq u$,

$$\hat{S}_d(q,u) = \big[\hat{T}_d^{\mathrm{sym}}(q,u) - \lambda_{\mathrm{dup}}\,\hat{U}_d^{\mathrm{sym}}(q,u)\big]_+.$$

This is an operational proxy score rather than a probabilistic identity.

Single-item support surrogate. Let $S_d^\star(q,u) \in [0,1]$ denote an unobserved shared-support quantity and

$$C_d^\star(q) = \mathbb{E}_{Q\sim\mu_d}\big[S_d^\star(q,Q)\big].$$

We do not estimate $C_d^\star$ directly. For $q \notin \mathcal{A}_d$, let

$$m_q := \sum_{u \in \mathcal{N}_k(q)} \mu_d(u),$$

and define

$$\hat{C}_d(q) = \sum_{u \in \mathcal{N}_k(q)} \mu_d(u)\,\hat{S}_d(q,u) + (1 - m_q) \sum_{a \in \mathcal{A}_d} \pi_a\,\hat{S}_d(q,a),$$

where $\pi_a \geq 0$ and $\sum_{a\in\mathcal{A}_d} \pi_a = 1$. The second term is an anchor-based tail surrogate, included for budgeted robustness rather than formal unbiasedness.

Shortlisting and final selection. We keep a non-anchor item q iff

$$q \in \mathcal{K}_d \iff \mathrm{BootLQ}_{\delta_A}\big(\hat{A}_d(q)\big) \geq \alpha_A \;\wedge\; \mathrm{BootLQ}_{\delta_C}\big(\hat{C}_d(q)\big) \geq \alpha_C \;\wedge\; D_d(q) \in \mathcal{I}_d.$$

For coverage, define

$$\hat{S}_d^{\mathrm{sp}}(q,u) = \begin{cases} 1, & q = u \in \bar{B}_d, \\ \hat{S}_d(q,u), & q \neq u \text{ and } u \in \mathcal{G}(q) \text{ or } q \in \mathcal{G}(u), \\ 0, & \text{otherwise.} \end{cases}$$

For a final set $\mathcal{S}_d \subseteq \mathcal{K}_d$ of size K_d, use

$$F_d^{\mathrm{sp}}(\mathcal{S}_d) = \sum_{u \in \bar{B}_d} \mu_d(u) \max_{q \in \mathcal{S}_d} \hat{S}_d^{\mathrm{sp}}(q,u). \quad (2)$$

We optimize

$$\max_{\mathcal{S}_d \subseteq \mathcal{K}_d} F_d^{\mathrm{sp}}(\mathcal{S}_d) \quad \text{s.t.} \quad |\mathcal{S}_d| = K_d,$$

subject to

$$L_h \leq \sum_{q \in \mathcal{S}_d} \mathbf{1}\{h(q) = h\} \leq U_h, \qquad \forall h \in \mathcal{H}_d,$$

and

$$\hat{U}_d^{\mathrm{sym}}(q,u) \le \tau_{\mathrm{dup}}, \qquad \forall q \ne u \in \mathcal{S}_d.$$

The quota condition

$$\sum_{h\in\mathcal{H}_d} L_h \le K_d \le \sum_{h\in\mathcal{H}_d} U_h$$

is necessary but not sufficient for feasibility; feasibility also depends on $\mathcal{K}_d$ and the duplicate constraint. Duplicate scores required by the last constraint are evaluated on demand on shortlisted pairs using the same cheap features and no additional LLM calls.

Proposition 1 (Monotone submodularity of F_d^{sp})**.** *If $\hat{S}_d^{\mathrm{sp}}(q,u) \ge 0$ for all q,u, then $F_d^{\mathrm{sp}}(\cdot)$ defined in (2) is monotone submodular. Greedy maximization under the cardinality constraint $|\mathcal{S}_d| = K_d$ therefore attains a $(1-1/e)$ approximation of the optimum of F_d^{sp} (Nemhauser et al., 1978).*

Proof. For fixed $u \in \bar{B}_d$, let $f_u(\mathcal{S}) := \max_{q\in\mathcal{S}} \hat{S}_d^{\mathrm{sp}}(q,u)$ with $f_u(\varnothing) = 0$. For any $A \subseteq B$ and $x \notin B$,

$$\begin{aligned} f_u(A\cup\{x\}) - f_u(A) &= \max\{0, \\ &\qquad \hat{S}_d^{\mathrm{sp}}(x,u) - f_u(A)\} \\ &\ge \max\{0, \\ &\qquad \hat{S}_d^{\mathrm{sp}}(x,u) - f_u(B)\} \\ &= f_u(B\cup\{x\}) - f_u(B), \end{aligned}$$

because $f_u(A) \le f_u(B)$. Hence f_u is monotone submodular. Since

$$F_d^{\mathrm{sp}}(\mathcal{S}) = \sum_{u\in\bar{B}_d} \mu_d(u)\, f_u(\mathcal{S}),$$
$$\mu_d(u) \ge 0,$$

F_d^{sp} is a nonnegative weighted sum of monotone submodular functions and is therefore itself monotone submodular. The $(1-1/e)$ greedy guarantee then follows from Nemhauser et al. (1978). □

Budget and implementation note. A typical domain uses two experts (about 10 total hours) for rubric alignment, item labels, pair labels, and audit, supporting roughly $|\mathcal{L}_d^{\mathrm{item}}| \approx$ 120–160 and $|\mathcal{L}_d^{\mathrm{pair}}| \approx$ 60–80, with 25%–30% overlap for agreement checks. We finally spent 100,000$ for KINA. Each item receives one main LLM call for $(\psi(q), \widetilde{A}_d(q))$, with optional re-checks on an uncertainty set $\mathcal{U}_d$, so

$$\begin{aligned} B_{\mathrm{LLM}}^{(d)} &= N_d + |\mathcal{U}_d| \\ &\le (1+\rho)N_d. \end{aligned}$$

The same signature $\psi(q)$ is reused for alignment, retrieval, pair scoring, and duplicate checks; no second item-level LLM pass is required. With fixed k, $|\mathcal{A}_d|$, and bootstrap count B, sparse retrieval and sparse support scoring scale linearly in N_d. If all duplicate constraints on $\mathcal{K}_d$ are materialized, the additional worst-case cost is $O(|\mathcal{K}_d|^2)$. Final constrained selection is solved by greedy-with-repair as a heuristic, or by a small MIP when $|\mathcal{K}_d|$ is moderate; we do not claim a general approximation guarantee under the full constraint set.

C Comparative Incentives under a Noisy Tournament

Scope. This appendix makes a comparative claim relative to flat payment. It does not prove truthfulness, rule out collusion, or solve the full repeated stochastic game. The whitelist extension below is reduced-form: it compares pointwise effort incentives under a fixed continuation value, not endogenous stationary state distributions. Throughout, $X \succeq_{\mathrm{FOSD}} Y$ means first-order stochastic dominance:

$$X \succeq_{\mathrm{FOSD}} Y \iff F_X(t) \le F_Y(t) \quad \forall t \in \mathbb{R}.$$

Strict incentive conclusions additionally require $\Delta p_{\min} > 0$, which is not automatic if score noise is large or the quality bar is poorly calibrated.

Setup. Two reviewers $i \in \{a,b\}$ evaluate the same item and choose effort $e_i \in \{L,H\}$. Let $\psi_i(e_i)$ be reviewer i's reduced-form private disutility of effort, and define the incremental cost of high effort by

$$\kappa_i \triangleq \psi_i(H) - \psi_i(L).$$

For $e \in \{L,H\}$, let $Q_i^e \in \mathbb{R}$ denote latent review quality with CDF $F_e(q) \triangleq \Pr(Q_i^e \le q)$.

Assumption 1 (Effort improves quality)**.** *$Q_i^H \succeq_{\mathrm{FOSD}} Q_i^L$, equivalently $F_H(q) \le F_L(q)$ for all q, with strict inequality on a set of positive measure.*

Assumption 2 (Noisy validated score)**.** *The principal observes a noisy validated score*

$$S_i^e = g(Q_i^e) + \varepsilon_i,$$

where g is strictly increasing and ε_i are i.i.d., continuous, and independent of all latent qualities. Let $F_e^S(s) \triangleq \Pr(S_i^e \le s)$.

Assumption 3 (Independent reviewer types)**.** *Conditional on the effort profile, reviewer qualities and score noises are independent across reviewers. Reviewer types κ_i are i.i.d. from a continuous CDF G, and each reviewer observes only her own type.*

Each reviewer receives base payment $w_0 \geq 0$, and a single bonus $B > 0$ is awarded to the reviewer with the higher validated score provided that the winning score clears a minimum bar τ:

$$W_i = w_0 + B \cdot \mathbf{1}\{S_i > S_{-i},\ S_i \geq \tau\}. \quad (3)$$

Define the winning probability $p_i(e_i, e_{-i}) \triangleq \Pr(S_i > S_{-i},\ S_i \geq \tau \mid e_i, e_{-i})$, and the one-shot utilities

$$U_i^{\text{tour}}(e_i; e_{-i}) = w_0 + B\, p_i(e_i, e_{-i}) - \psi_i(e_i),$$
$$U_i^{\text{flat}}(e_i) = w_0 - \psi_i(e_i).$$

One-shot comparison. For any fixed opponent effort $e_{-i} \in \{L, H\}$,

$$F_e^S(s) = \Pr(S_i^e \leq s) = \mathbb{E}\big[F_\varepsilon\big(s - g(Q_i^e)\big)\big]\,.$$

Since g and F_ε are increasing, the map $q \mapsto F_\varepsilon(s - g(q))$ is decreasing, so Assumption 1 together with Assumption 2 implies

$$S_i^H \succeq_{\text{FOSD}} S_i^L, \qquad \text{i.e.} \qquad F_H^S(s) \leq F_L^S(s) \quad \forall s.$$

Now define $T_i \triangleq \max\{S_{-i}^{e_{-i}}, \tau\}$. Because scores are continuous, $\{S_i > S_{-i},\ S_i \geq \tau\} = \{S_i > T_i\}$ almost surely. Conditional on the effort profile, S_i^e is independent of T_i. Hence

$$p_i(e, e_{-i}) = \Pr(S_i^e > T_i) = \mathbb{E}\big[1 - F_e^S(T_i)\big]\,.$$

Therefore, the difference in winning probabilities is

$$\Delta p_i(e_{-i}) \triangleq p_i(H, e_{-i}) - p_i(L, e_{-i}) = \mathbb{E}\big[F_L^S(T_i) - F_H^S(T_i)\big] \geq 0.$$

The one-shot gain from high effort is then

$$\Delta U_i^{\text{tour}}(e_{-i}) \triangleq U_i^{\text{tour}}(H; e_{-i}) - U_i^{\text{tour}}(L; e_{-i}) = B\,\Delta p_i(e_{-i}) - \kappa_i,$$

whereas under flat payment, $\Delta U_i^{\text{flat}} = U_i^{\text{flat}}(H) - U_i^{\text{flat}}(L) = -\kappa_i$. Thus the tournament adds a positive extrinsic return $B\,\Delta p_i(e_{-i})$, and high effort is weakly preferred whenever $B\,\Delta p_i(e_{-i}) \geq \kappa_i$.

Population implication and review-quality shift. By symmetry, write

$$\Delta p(L) \triangleq \Delta p_i(L),$$
$$\Delta p(H) \triangleq \Delta p_i(H),$$
$$\Delta p_{\min} \triangleq \min\{\Delta p(L), \Delta p(H)\}.$$

Let π^m be the equilibrium high-effort rate under mechanism $m \in \{\text{flat}, \text{tour}\}$.

Theorem 1 (Tournament FOSD improvement)**.** *Under Assumptions 1–3, if $\Delta p_{\min} > 0$, the equilibrium high-effort rate satisfies*

$$\pi^{\text{tour}} \;\geq\; G(B\,\Delta p_{\min}) \;\geq\; G(0) \;=\; \pi^{\text{flat}}.$$

Moreover, let $Y^m = \Gamma(Q_a^m, Q_b^m, Z)$ denote released review quality under mechanism m, where Γ is coordinatewise nondecreasing and Z is mechanism-invariant and independent of reviewer qualities. Then

$$Y^{\text{tour}} \succeq_{\text{FOSD}} Y^{\text{flat}}.$$

Proof. If $\Delta p_{\min} > 0$, then for all opponent actions, $\Delta U_i^{\text{tour}}(e_{-i}) \geq B\,\Delta p_{\min} - \kappa_i$. Hence every type satisfying $\kappa_i \leq B\,\Delta p_{\min}$ weakly prefers high effort regardless of the opponent's action. Under flat payment, high effort is chosen only when $\kappa_i \leq 0$. Since G is continuous, indifferent types have zero mass, so

$$\pi^{\text{tour}} \geq G(B\,\Delta p_{\min}) \geq G(0) = \pi^{\text{flat}}.$$

Let Q^m denote the latent quality of a randomly sampled review under mechanism m. Conditioning on the effort choice gives $F_Q^m(q) = \pi^m F_H(q) + (1 - \pi^m) F_L(q)$, so

$$F_Q^{\text{tour}}(q) - F_Q^{\text{flat}}(q) = (\pi^{\text{tour}} - \pi^{\text{flat}})\big(F_H(q) - F_L(q)\big) \leq 0 \quad \forall q,$$

which proves $Q^{\text{tour}} \succeq_{\text{FOSD}} Q^{\text{flat}}$ marginally. By Assumption 3, conditional on each mechanism, (Q_a^m, Q_b^m) are i.i.d. with common marginal F_Q^m. Let $(F_Q^m)^{-1}(u) \triangleq \inf\{x \in \mathbb{R} : F_Q^m(x) \geq u\}$. Since $F_Q^{\text{tour}}(q) \leq F_Q^{\text{flat}}(q)$ for all q,

$$(F_Q^{\text{tour}})^{-1}(u) \geq (F_Q^{\text{flat}})^{-1}(u) \quad \forall u \in (0, 1).$$

If $U_a, U_b \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1)$ are independent of Z, and $\widetilde{Q}_j^m = (F_Q^m)^{-1}(U_j)$ for $j \in \{a,b\}$, then $(\widetilde{Q}_a^m, \widetilde{Q}_b^m)$ has the same law as (Q_a^m, Q_b^m), while $\widetilde{Q}_j^{\text{tour}} \geq \widetilde{Q}_j^{\text{flat}}$ almost surely for $j = a, b$. By coordinatewise monotonicity of Γ,

$$\Gamma(\widetilde{Q}_a^{\text{tour}}, \widetilde{Q}_b^{\text{tour}}, Z) \geq \Gamma(\widetilde{Q}_a^{\text{flat}}, \widetilde{Q}_b^{\text{flat}}, Z)$$

almost surely under the same coupling, which is FOSD: $Y^{\text{tour}} \succeq_{\text{FOSD}} Y^{\text{flat}}$. □

The reviewer-independence assumption (Assumption 3) is necessary for the joint comparison: marginal FOSD alone does not suffice to compare the joint output Y^m.

Corollary 1 (Conservative bonus calibration)**.** *Assume $\Delta p_{\min} > 0$, and define the generalized inverse $G^{-1}(u) \triangleq \inf\{x \in \mathbb{R} : G(x) \geq u\}$. The following sufficient conditions hold:*

(a) *If $\kappa_i \equiv \Delta C > 0$, then $B > \dfrac{\Delta C}{\Delta p_{\min}}$ makes high effort strictly dominant for every reviewer type.*

(b) *For any target effort rate $\pi^\star \in [0,1)$, $B \geq \dfrac{[G^{-1}(\pi^\star)]_+}{\Delta p_{\min}}$ ensures $\pi^{\text{tour}} \geq \pi^\star$, where $[x]_+ \triangleq \max\{x, 0\}$.*

(c) *If first moments exist and $\Delta_Q \triangleq \mathbb{E}[Q^H] - \mathbb{E}[Q^L] > 0$, then a feasible target mean-quality gain $\eta \in \big[0, (1-\pi^{\text{flat}})\Delta_Q\big]$ is guaranteed by*

$$B \geq \frac{\big[G^{-1}\big(\pi^{\text{flat}} + \eta/\Delta_Q\big)\big]_+}{\Delta p_{\min}}.$$

Bonus expenditure accounting. For budget accounting, define $\alpha_{uv} \triangleq \Pr\big(\max\{S_a^u, S_b^v\} \geq \tau\big)$ for $(u,v) \in \{L,H\}^2$. Under continuous scores, exactly one reviewer receives the bonus whenever $\max\{S_a^u, S_b^v\} \geq \tau$. Thus if the population high-effort rate is π, the expected variable bonus expenditure per task is

$$\mathcal{C}_{\text{bonus}}(B,\pi) = B\Big[\pi^2\alpha_{HH} + 2\pi(1-\pi)\alpha_{HL} + (1-\pi)^2\alpha_{LL}\Big] \leq B.$$

Reduced-form whitelist extension. Let z_i denote reviewer i's current whitelist state, $\Omega_i(z_i) \geq 0$ the discounted continuation value of remaining eligible for future tasks, and $q_i(e_i, e_{-i}; z_i) \in [0,1]$ the probability that the current task causes reviewer i to lose whitelist status. Holding fixed the whitelist rule, define

$$\begin{aligned} U_i^{\text{tour+wl}}(e_i; e_{-i}, z_i) &= w_0 + B\,p_i(e_i, e_{-i}) - \psi_i(e_i) \\ &\quad + \big(1 - q_i(e_i, e_{-i}; z_i)\big)\Omega_i(z_i), \\ U_i^{\text{flat+wl}}(e_i; e_{-i}, z_i) &= w_0 - \psi_i(e_i) \\ &\quad + \big(1 - q_i(e_i, e_{-i}; z_i)\big)\Omega_i(z_i). \end{aligned}$$

Let $\Delta q_i(e_{-i}, z_i) \triangleq q_i(L, e_{-i}; z_i) - q_i(H, e_{-i}; z_i)$, and assume $\Delta q_i(e_{-i}, z_i) \geq 0$ for all e_{-i}, z_i. Then

$$\begin{aligned} \Delta U_i^{\text{tour+wl}}(e_{-i}, z_i) &= \Delta U_i^{\text{flat+wl}}(e_{-i}, z_i) + B\,\Delta p_i(e_{-i}), \\ \Delta U_i^{\text{flat+wl}}(e_{-i}, z_i) &= -\kappa_i + \Delta q_i(e_{-i}, z_i)\Omega_i(z_i). \end{aligned}$$

Equivalently,

$$\begin{aligned} \Delta U_i^{\text{tour+wl}}(e_{-i}, z_i) &= B\,\Delta p_i(e_{-i}) - \kappa_i + \Delta q_i(e_{-i}, z_i)\Omega_i(z_i) \\ &\geq \Delta U_i^{\text{tour}}(e_{-i}). \end{aligned}$$

Define $\Lambda_i \triangleq \inf_{e_{-i}, z_i} \Delta q_i(e_{-i}, z_i)\Omega_i(z_i)$, and $\widetilde{\kappa}_i \triangleq \kappa_i - \Lambda_i$. Then

$$\Delta U_i^{\text{tour+wl}}(e_{-i}, z_i) \geq B\,\Delta p_{\min} - \widetilde{\kappa}_i \quad \forall e_{-i}, z_i.$$

Thus all one-shot sufficient conditions in Corollary 1 apply *pointwise* after replacing κ_i by $\widetilde{\kappa}_i$. If $\kappa_i \equiv \Delta C$ and $\Lambda_i \geq \Lambda_{\min}$ for all i, then $B > \dfrac{(\Delta C - \Lambda_{\min})_+}{\Delta p_{\min}}$ is a conservative sufficient condition for high effort to be strictly optimal at every state.

Takeaway. Relative to flat payment, the noisy tournament changes private incentives through $B\,\Delta p_i(e_{-i})$, the bonus-weighted increase in the probability of clearing the quality bar and outperforming the matched reviewer. When this return is large enough relative to reviewer-specific effort cost, the set of high-effort types expands, inducing a first-order stochastic right-shift in per-review latent quality. Under reviewer independence and monotone adjudication, the same right-shift carries over to released-data quality. Whitelist access further strengthens this comparison pointwise by lowering the effective net cost of effort through continuation value.

D Taxonomy of Disciplines

Figure 4: Taxonomy of Disciplines

Table 7: Disciplines and Subfields Statistics

Discipline	Field	Subfield	Count
Agronomy	Animal Husbandry	Animal Nutrition and Feed Science	8
		Animal Rearing and Breeding	7
	Aquaculture	Aquaculture	3
	Crop Science	Crop Science	7
	Forestry	Forest Cultivation and Genetic Breeding	3
		Landscape Plants and Ornamental Horticulture	4
	Veterinary Medicine	Veterinary Medicine	8
Economics	Applied Economics	Economic Statistics	3
		Finance	5
		Industrial Economics	3
		International Trade	3
		Labor Economics	5
		Public Finance	3
		Quantitative Economics	3
	Theoretical Economics	Economic History	1
		Political Economy	3
		Western Economics	5
Education	Education	Educational Technology and Principles	4
		Preschool Education	6
		Special Education	3
		Theory of Curriculum and Instruction	2
	Physical Education	Physical Education and Training	6
		Sports Science and Medicine	3
	Psychology	Psychology	7
Engineering	Aeronautical and Astronautical Science and Technology	Aeronautical and Astronautical Science and Technology	3
	Agricultural Engineering	Agricultural Environment and Soil-Water Engineering	6
		Agricultural Mechanization Engineering	3
	Architecture	Architectural Design and Theory	1
		Urban Planning and Design	1
		Architectural History	4
	Chemical Engineering and Technology	Chemical Transport Engineering	4
		Elements of Chemical Reaction Engineering	8
		Fluid Flow and Heat Transfer in Chemical Engineering	4
		Mass Transport and Separation Process in Chemical Engineering	3
	Civil Engineering	Bridge and Tunnel Engineering	2
		Geotechnical Engineering	3
		Structural Engineering	4
		Urban Infrastructure Engineering	3
	Computer Science and Technology	Advanced Programming Languages	5
		Computer Architecture	3
		Computer Networks	2
		Computer Software and Theory	3
		Data Structures	2
		Databases	4
		Formal Languages	4
		Operating Systems	3
		Pattern Recognition	3
		Principles of Computer Organization	2

Continued on next page...

Table 8 Disciplines and Subfields Statistics

Discipline	Field	Subfield	Count
	Control Science and Engineering	Control Theory and Control Engineering	10
		Guidance, Navigation and Control	2
		Operations Research and Cybernetics	2
	Electrical Engineering	Electrical Theory and New Technologies	3
		High Voltage and Insulation Technology	2
		Power Electronics and Electrical Drives	8
		Power Systems and Automation	3
	Electronic Science and Technology	Circuits and Systems	4
		Electromagnetic Field and Microwave Technology	3
		Microelectronics and Solid-State Electronics	3
	Environmental Science and Engineering	Environmental Engineering	4
		Environmental Science	2
		Environmental and Resource Protection	4
	Food Science and Engineering	Food Biochemistry	3
		Food Processing and Storage Engineering	3
	Forestry Engineering	Forest Engineering	4
		Wood Science and Technology	4
	Geological Resources and Geological Engineering	Geological Resources and Geological Engineering	5
	Hydraulic Engineering	Hydraulics and Hydrology	5
		Water conservancy and Hydropower Engineering	2
	Information and Communication Engineering	Antenna and Radio Communication	5
		Communication Principles	3
		Communication and Information Systems	3
		Optical Fiber Communication	3
		Signal and Information Processing	3
	Instrument Science and Technology	Instrument Science and Technology	8
	Materials Science and Engineering	Materials Physics and Chemistry	3
		Materials Processing Engineering	5
	Mechanical Engineering	Manufacturing Automation	7
		Mechatronic Engineering	3
	Mechanics	Rigid Body Mechanics	3
		Theoretical Fluid Mechanics	5
		Theoretical Mechanics	3
	Metallurgical Engineering	Iron and Steel Metallurgy	3
		Non-ferrous Metallurgy	3
		Physical Chemistry of Metallurgical Process	3
		Principles of Metallurgy	3
	Mining Engineering	Mineral Processing Engineering	3
	Naval Architecture and Ocean Engineering	Marine Engineering	3
		Ship Mechanics and Design Principles	3
	Nuclear Science and Technology	Nuclear Energy and Reactor Technology	2
		Radiation Protection and Nuclear Technology Applications	3
	Optical Engineering	Applied Optics	3
		Laser Technology	4
		Optoelectronic Technology	3
		Theoretical Optics	5
	Petroleum and Natural Gas Engineering	Oil and Gas Field Development and Storage & Transportation Engineering	4
		Poromechanics and Reservoir Physics	2

Continued on next page...

Table 8 Disciplines and Subfields Statistics

Discipline	Field	Subfield	Count
	Power Engineering and Engineering Thermophysics	Engineering Fluid Mechanics	2
		Engineering Thermophysics	3
		Fluid Machinery and Engineering	2
		Heat Transfer	5
		Internal Combustion Engineering	3
		Power Machinery and Engineering	2
		Refrigeration and Cryogenic Engineering	3
		Thermal Energy Engineering	3
	Surveying and Mapping Science and Technology	Cartography and Geographic Information Engineering	2
		Geodesy and Surveying Engineering	3
	Textile Science and Engineering	Textile Chemistry and Dyeing Engineering	3
		Textile Materials Science	3
	Transportation Engineering	Road and Railway Engineering	3
		Traffic Information Engineering and Control	3
		Transportation Planning and Management	3
		Vehicle Operation Engineering	4
	Weapon Science and Technology	Military Chemistry and Pyrotechnics	1
		Weapon Systems Science and Engineering	3
History	History	Archaeology and Museology	9
		Historical Geography	2
		World History	2
Law	Law	Civil and Commercial Law	5
		Constitutional and Administrative Law	6
		Contract Law	13
		Criminal Law	5
		International Law	14
		Law and Social Governance	5
		Legal Theory and Legal History	6
		Procedural Law	7
	Political Science	Political Science	4
Literature and Arts	Art Studies	Broadcasting and Television Art	3
		Dance Studies	3
		Design Arts	4
		Drama and Opera Studies	6
		Film Studies	3
		Fine Arts	3
	Journalism and Communication	Communication and Broadcasting	3
		History and Theory of Journalism and Media Management	5
		Journalism and News Practice	3
	Language and Literature	Classical Asian Literature	3
		French Language and Literature	4
		Linguistics and Applied Linguistics	2
		Literary History	3
		Literary Theory	3
		Modern and Contemporary Literature	1
		Philology and Bibliography	3
		Russian Language and Literature	3
	Musicology	Composition	3
		Harmony	1
		Instrumentation and Performance	1
		Music History, Education, and Technology	2
		Musical Forms and Analysis	3
Management	Business Administration	Business and Accounting Management	6
	Library, Information and Archival Management	Information Management Science	5

Continued on next page...

Table 8 Disciplines and Subfields Statistics

Discipline	Field	Subfield	Count
		Library and Archival Science	9
	Management Science and Engineering	Management Science and Engineering	1
	Public Administration	Education Economics, Management and Social Security	3
		Land Resource Management and Administrative Management	4
		Social Medicine and Health Management	1
Medicine	Basic Medicine	Forensic Medicine	3
		Human Anatomy and Histology-Embryology	3
		Immunology	3
		Pathogen Biology	1
		Pathology and Pathophysiology	2
	Clinical Medicine	Anesthesiology	2
		Clinical Laboratory Diagnostics	8
		Dermatology and Venereology	1
		Emergency Medicine	5
		Geriatric Medicine	3
		Imaging and Nuclear Medicine	3
		Internal Medicine	4
		Neurology	1
		Nursing and Rehabilitation Medicine	7
		Obstetrics and Gynecology	2
		Oncology	1
		Ophthalmology	1
		Otorhinolaryngology	3
		Pediatrics	1
		Psychiatry and Mental Health	2
		Surgery	1
	Pharmacy	Medicinal Chemistry	3
		Microbiology and Biochemical Pharmacy	3
		Pharmaceutical Analysis	3
		Pharmaceutics	2
		Pharmacology	1
	Public Health and Preventive Medicine	Epidemiology and Health Statistics	1
		Health Toxicology and Environmental Health	1
	Stomatology	Basic Stomatology	3
		Clinical Stomatology	2
	Traditional Medicine	Traditional Health Preservation	3
		Traditional Medicine Theory	2
Philosophy	Philosophy	Ethics	2
		Logic	1
		Philosophical Aesthetics	1
		Philosophy of Science and Technology	7
		Religious Studies	2
Science	Astronomy	Astronomical Observation and Technology	2
		Astrophysics	4
		Cosmology	1
		Solar System Science	5
		Stellar and Interstellar Evolution	4
	Atmospheric Science	Atmospheric Physics and Atmospheric Environment	2
		Meteorology	2
	Biology	Biochemistry and Molecular Biology	3
		Biophysics	2
		Botany	3
		Cell Biology	2
		Ecology	3

Continued on next page...

Table 8 Disciplines and Subfields Statistics

Discipline	Field	Subfield	Count
		Genetics	3
		Microbiology	3
		Physiology	3
		Zoology	3
	Chemistry	Analytical Chemistry	3
		Electrochemistry	3
		Inorganic Chemistry	2
		Organic Chemistry	2
		Physical Chemistry	3
		Polymer Chemistry and Physics	4
		Radiochemistry	6
	Geography	Human Geography	3
		Physical Geography	5
	Geology	Geochemistry	8
		Mineralogy, Petrology, and Economic Geology	3
		Paleontology and Stratigraphy	2
		Principles of Seismic Exploration	3
		Structural Geology	3
	Geophysics	Solid Earth Geophysics	2
	Mathematics	Advanced Algebra	4
		Combinatorial Mathematics	2
		Computational Mathematics	3
		Cryptography	8
		Discrete Mathematics	2
		Functions of Complex Variables	4
		Functions of Real Variables	1
		Fundamental Mathematics	3
		Fuzzy Mathematics	8
		Geometry and Topology	3
		Graph Theory	2
		Group Theory	3
		Mathematical Analysis	3
		Number Theory	3
		Numerical Analysis	3
		Ordinary Differential Equations	3
		Polynomials and Series Expansions	1
		Probability and Statistics	2
		Special Number Theory	3
		Stochastic Processes	3
	Oceanography	Hydrogeology	4
		Marine Biology	2
		Marine Chemistry	3
		Underwater Acoustics	2
	Physical Oceanography	Physical Oceanography	3
	Physics	Acoustics	2
		Atomic and Molecular Physics	3
		Electrodynamics	3
		Fluid Physics	3
		Particle and Nuclear Physics	1
		Polymer Physics	3
		Quantum Mechanics	3
		Semiconductor Physics	3
		Subatomic and Atomic Physics	1
		Thermodynamics	2
		Thermodynamics and Statistical Physics	2
Sociology	Sociology	Demography and Anthropology	10
		Social and Folklore Studies	9

E Data Collection Details

E.1 Annotator Recruitment Test

1. **Academic Background**
   - University (Full Name in English):
   - Highest Academic Degree (Bachelor's / Master's / Doctoral / Currently Enrolled):
   - Major Name (Official Title):
   - Intended Disciplinary Fields for Question Design (Limit: 1–3):

2. **Authoritative Domain Knowledge**

   Please list **3** of the most authoritative publications, conferences or monographs (titles included) in your professional field. We intend to evaluate your professional insight through your understanding of the authoritative sources in the industry.

   Format requirement: Name + Type (Journal/-Conference/Monograph) + Rationale

3. **Understanding of Disciplinary Competency Assessment**

   If you were asked to design **3–6 multiple-choice questions** to assess the advanced literacy and cognitive abilities of LLMs in your chosen field, what content would each question target? Please briefly elaborate on the question content and explain the rationale for designing each question.

   Please ensure that the overall design logic of this set of questions is clearly demonstrated: specifically, how these questions, covering different dimensions within the discipline, can be integrated to outline a comprehensive cognitive framework of the subject.

   Format requirement: Content + Rationale

4. **Stress Tolerance and Work Attitude**

   (a) **Attitude Towards Review and Revision (Single Choice)**
   If your questions are returned due to incomplete citations or non-standard formatting, and you are required to supplement supporting evidence item by item, what would your response be?
   - A. The rules are reasonable and necessary; I will supplement the required information accordingly
   - B. I feel a bit annoyed, but I will revise the questions as required
   - C. I think the requirements are overly detailed and somewhat confusing
   - D. I am not willing to continue with this process

   (b) **Tendency of Responsibility Attribution (Single Choice)**
   If your questions fail to pass the review multiple times, what do you think is the primary cause?
   - A. My understanding of the standards and requirements is not thorough enough
   - B. The questions themselves still have room for improvement
   - C. The review standards are overly subjective
   - D. Bad luck

5. **Rule Comprehension Verification**

   (a) **Understanding of Question Types (Multiple Choice)**
   Which of the following questions do **not** meet the requirements for high-quality disciplinary assessment questions?
   - A. Questions testing the reasoning and application of core disciplinary theories in specific scenarios
   - B. Questions requiring the model to memorize specific numerical values or factual information
   - C. Combined logical reasoning questions based on multiple statements
   - D. Static judgment questions involving only extremely niche or rare subjects
   - E. Questions developed based on individual scholars' viewpoints that lack widespread consensus

   (b) **Understanding of Question Design (Single Choice)**
   What is the main source of difficulty in high-quality assessment questions?
   - A. Numerous calculation steps and complex numerical values
   - B. Obscure knowledge points and high memory difficulty
   - C. Interactions between multiple restrictive conditions

- D. Long question length and large information volume

6. **Workflow and Standard Confirmation (Multiple Choice)**
    - Develop all questions in English (including punctuation marks) throughout the entire process
    - Each option (whether correct or incorrect) must be supported by evidence or citations
    - Questions may undergo multiple rounds of review and revision
    - Questions must demonstrate disciplinary representativeness, rather than being esoteric, biased, or tricky

E.2 Annotation Manual

E.2.1 Project Background

Current LLMs have demonstrated strong capabilities in programming, mathematics, and reasoning, but their performance still varies widely across different disciplines. To enable researchers and practitioners in various disciplinary fields to select the most appropriate large AI model for their work, domain experts are needed to design questions that evaluate models' disciplinary knowledge and competence in their respective fields. The questions you provide will serve as critical data for evaluating LLMs.

- Your role: **Question Design Expert**
- Your task: Create high-difficulty multiple-choice questions that at least three LLMs will answer incorrectly, and provide explanations and source materials.
    - Question: Consists of a stem and options. Questions must be **representative and unique** to the discipline, and able to assess high-level knowledge and literacy in the field. Questions may be original, adapted, or directly excerpted; **AI-generated questions are strictly prohibited.**
    - Explanation: Clearly explain why each option is correct or incorrect.
    - Source Materials: The origin of the question, which may be a book or high-quality academic paper.
- Discipline List: Please check whether the list includes your specialized discipline. Ensuring the accuracy and rationality of the questions is our top priority.

E.2.2 Core Criteria for Question Design

What We Require:

1. **Disciplinary Representativeness:** Instances must act as highly representative probes. If a discipline were to be evaluated using only 3 to 5 questions, the submitted instance must be fundamental and comprehensive enough to be one of them.
2. **High-Order Knowledge Application:**
    - Questions must construct a logically complete closed system where the solution is the inevitable product of general deductive reasoning applied to domain-specific axioms.
    - Difficulty should stem from the complex coupling of constraints and multi-hop reasoning, not merely computational heavy-lifting.
    - For instance, physics questions should demand the integration of axioms and skills to solve a specific state transition, rather than asking for the formula of momentum conservation.
3. **Static Knowledge Breadth:** For memory-intensive disciplines (e.g., Education), questions should maximize knowledge coverage (e.g., utilizing $10+$ statements to encompass major pedagogical theories).

What We Reject:

1. **Narrow or Trivial Memorization:** Questions testing pure rote memory devoid of core disciplinary literacy (e.g., "What is the 100th digit of π?") or focusing on hyper-niche, obscure sub-entities.
2. **Idiosyncratic or Tricky Trivia:** Questions universally recognized as flawed or unreasonable even in human examinations.
3. **Weak Epistemological Consensus:** Hypotheses proposed by individual scholars that lack widespread academic consensus or are highly volatile (e.g., highly debated legal interpretations).

4. **Non-Disciplinary Failure Modes:** Instances where LLMs fail due to semantic traps, ambiguous phrasing, or floating-point calculation errors rather than a deficit in disciplinary literacy.

E.2.3 Standardized Annotation Workflow

The question authoring process is strictly compartmentalized into six components. The specifications for each component are detailed in Table 9.

E.2.4 Accepted and Rejected Cases

Table 10 illustrates the dichotomy between accepted and rejected submissions based on our core criteria.

The accepted case provided here is for reference only in terms of content and format. Please note that question types are not limited to calculation problems.

This question is selected as an excellent example for the following reasons:

1. It targets marine chemistry and assesses advanced reasoning and computational abilities, rather than only testing static knowledge.

2. Even though it is a calculation question, the analysis of incorrect options is briefly summarized instead of being copied and pasted repetitively.

E.3 Review Manual

E.3.1 Core Philosophy and General Workflow

The primary objective of the review process is to ensure that each curated instance serves as a highly representative probe for evaluating LLMs on specific fine-grained disciplines. Reviewers must adopt a **zero-tolerance policy** toward ambiguous, peripheral, or substandard data.

- **Comprehensive Feedback:** Unless an instance requires a complete topic overhaul, reviewers must identify and aggregate all logical, factual, and formatting issues into a single comprehensive feedback report before returning it for revision.

- **AI-Generation Rejection:** Submissions exhibiting clear patterns of unedited, large-scale AI generation must be categorically rejected.

- **Escalation Mechanism:** For borderline cases or interdisciplinary disputes, reviewers are required to escalate the instance to the Core Committee for final arbitration.

E.3.2 Disciplinary Alignment and Factuality

- **Disciplinary Matching:** Reviewers must first verify whether the question strictly aligns with the declared subfield. Mismatched instances must be returned with a mandate to reclassify or rewrite.

- **Epistemological Rigor (Especially in Humanities & Social Sciences):** Questions must test established scientific truths, foundational theories, or widely recognized academic consensus. Volatile opinions, transient scholarly debates, or highly subjective hypotheses without proper contextualization must be rejected.

- **Memory Error Rejection:** If an instance causes a model to fail solely due to a superficial memory error (e.g., misremembering an obscure date, a peripheral name, or a raw statistic) rather than a deficit in high-order deductive reasoning or conceptual understanding, the instance must be rejected. Questions must evaluate the mastery of causal mechanisms, not data retrieval.

E.3.3 Source Verification and Material Authenticity

Every claim must be fully traceable to authoritative human knowledge sources.

- **Authoritative Sources:** Referenced literature must be identifiable via valid URLs, DOIs, or standard APA citations. Acceptable sources include high quality peer-reviewed journals (e.g., SCI/SSCI Q1, CSSCI), authoritative monographs, or official databases. Preprints (e.g., arXiv) are only acceptable if they demonstrate high citation counts or widespread community validation.

- **Material Consistency:** Any accompanying materials (e.g., figures, charts) must be intrinsically necessary to solve the question and perfectly correspond to the cited source.

E.3.4 Option Rigor and Structural Integrity

To prevent LLMs from exploiting logical loopholes, the 10-option structure must adhere to strict constraints:

- **Pseudo-Multi-Choice Constraints:** For combination-based questions, the question stem must present a minimum of **6 distinct foundational statements**.

Table 9: Standardized Annotation Workflow and Specifications.

Component	Specifications and Constraints
Question	Must be pure English text. Image uploads in the stem are prohibited. For pseudo-multi-choice formats, the stem must contain ≥ 6 foundational statements. The question stem should not include instructions such as "You are an expert". Must be structurally self-contained; references like "in this paper" are strictly forbidden. All necessary premises must be explicitly stated. Sources must be highly authoritative: e.g., established textbooks, Q1 SCI/SSCI journals, or CCF-A/B proceedings. Submissions citing controversial publishers (e.g., MDPI, Hindawi) or user-generated forums are categorically rejected.
Options	Must provide exactly 10 distinct, pure-text English options. For combinatorial options, the distribution must reflect statistical normality (avoiding extreme imbalance where one option has 1 statement and another has 8). Options should not begin with labels like "A.". **Strict Prohibition of Proper Subsets:** Options must not exhibit subset relationships (e.g., if A is $\{1, 2, 4\}$, B cannot be $\{1\}$) to prevent logical leakage.
Explanation	Must explicitly diagnose why the correct option is valid and why *each* distractor is flawed (avoiding generic copy-pasting of theoretical background). Unique explanations per option are mandatory, except for deterministic calculation questions.
Sources	**Visual Evidence:** Must upload screenshots of the source material with precise highlights and annotations corresponding to each option/statement. **Traceability:** Must provide accessible URLs or rigorous APA-formatted citations (including specific page numbers).
Material	Reference images necessary for solving the question must be uploaded and annotated.
Formatting	All mathematical formulas and symbols must be written in standard LaTeX and render flawlessly in the Markdown environment.

Table 10: Accepted and Rejected Cases

Case 1: Accepted Case	
Discipline	Marine Science - Marine Chemistry
Question	It is known that Modified Circumpolar Deep Water (MCDW) is defined as the water mass in the Ross Sea located between neutral density surfaces of 28 kg m^{-3} and 28.27 kg m^{-3}. The average salinity of MCDW is 34.72 ± 0.01, and the average salinity of Antarctic Bottom Water (AASW) is 33.7 ± 0.2. The dissolved organic carbon (DOC) concentration in MCDW is 40.0 μmol kg^{-1}, in AASW is 45.1 μM, and in Deep Shelf Water (DSW) is 48.4 μM. During the formation of DSW, how much does the DOC concentration increase in the Ross Sea, and how much excess DOC is exported from the deep water to the shelf area?
Options	**A.** 7 μM; 4 ± 0.6 Tg DOC yr^{-1} **B.** 8 μM; 4 ± 0.8 TgC yr^{-1}
Explanation	**Option A (Correct):** MCDW is the primary water source for the Ross Ice Shelf, formed by the mixing of Circumpolar Deep Water (CDW) and AASW on the continental slope. Given the average salinity of CDW (34.72 ± 0.01) and AASW (33.7 ± 0.2), assuming salinity on the continental shelf is conservative (i.e., unaffected by processes like brine rejection), we calculate the proportions required to form MCDW (salinity = 34.47 ± 0.05). The calculation is: $f(\text{CDW}) = \frac{S(\text{AASW}) - S(\text{mCDW})}{S(\text{AASW}) - S(\text{CDW})}$ $f(\text{AASW}) = 1 - f(\text{CDW})$ Where S represents average salinity and f represents the proportion. The proportions of CDW and AASW forming MCDW are 0.75 and 0.25, respectively. With DOC(CDW) = 40.0 and DOC(AASW) = 45.1: DOC(MCDW) $=$ DOC(CDW) $\cdot f(\text{CDW}) +$ DOC(AASW) $\cdot f(\text{AASW}) = 41.0\mu$M. DOC(enriched) $=$ DOC(DSW) $-$ DOC(mCDW) $= 48.4 - 41.0 = 7.4\mu$M. The DOC export off the continental shelf is calculated as: 7μM export$\times 6$ years residence time $= 4$ TgC yr^{-1}. **Option B (Incorrect):** The enriched DOC content is overestimated. The estimated DOC content in DSW is 1μM higher than the correct value. Consequently, the exported DOC is also overestimated: $8\mu\text{M} \times 6$ years $=$ 4.8 TgC yr^{-1}.
Source	Bercovici, S. K., Huber, B. A., DeJong, H. B., Dunbar, R. B., & Hansell, D. A. (2017). Dissolved organic carbon in the Ross Sea: Deep enrichment and export. *Limnology and Oceanography*, 62(6), 2593-2603.

Table 10: Accepted and Rejected Cases (continued)

Case 2: Rejected Case)	
Discipline	Biology - Zoology
Question	Based on the content of Chapter 1 in the attachment, which of the following statements about the genus *Acanthorhodeus* are correct: 1. Body elongated, slightly compressed, large head, flat snout. 2. Lateral line scales are at least weakly ctenoid, two dorsal fins, separated. 3. Posterior margin of the caudal fin is rounded. 4. The first dorsal fin is short, consisting of 3-4 spines. 5. Upper jaw is slightly shorter than the lower jaw.
Options	**A.** 1, 2 **B.** 1, 2, 3 (Correct Option) **C.** 1, 2, 3, 4, 5 **D.** 1 **E.** 2 **F.** 1, 2, 4 **G.** 1, 3, 5 **H.** 3, 5 **J.** 4, 5
Explanation	**Option A:** Statement 1 is correct (Body elongated, slightly compressed, large head, flat snout); Statement 2 is correct (Lateral line scales are at least weakly ctenoid, two dorsal fins, separated). **Option J:** Statement 4 is incorrect (The first dorsal fin is short, consisting of 4-6 spines); Statement 5 is incorrect (Upper jaw is slightly longer than the lower jaw).
Rejection Rationale	**1. Weak Disciplinary Representativeness:** The question fails to engage high-order disciplinary literacy. Statements 4 and 5 merely test pure static memory of isolated trivia, lacking sufficient cognitive coverage to serve as a representative evaluation probe. **2. Structural Flaw (Proper Subsets):** Options A, D, and E are proper subsets of the correct Option B. Models can exploit this logical leakage (selecting A, D, or E would technically be correct). Furthermore, the option distribution is severely imbalanced, violating statistical normality constraints. **3. Statement Count Deficit:** The question stem only contains 5 statements, failing the strict pseudo-multi-choice minimum requirement (≥ 6 statements). **4. Context Drift:** The stem explicitly references an unprovided context ("Chapter 1"), violating the requirement that all questions must be logically self-contained.

- **No Proper Subset Relationships:** To avoid logical leakage, options must not have proper subset relationships. For example, if Option A is $\{1, 2\}$ and Option B is $\{1, 2, 3\}$, proving B correct implicitly validates A, introducing ambiguity. Options must be mutually exclusive in their truth values.

- **Explanation Completeness:** The explanations provided for the options must comprehensively address why the correct option is uniquely valid and why all distractors are flawed. For deterministic questions (e.g., mathematical calculations), overlapping explanations across options are permissible provided the core proof is sound.

E.4 Formatting and Rendering

Reviewers must ensure that all text, mathematical formulas, and symbolic logic within the Question, Options, and Explanations strictly conform to standard LaTeX syntax and render flawlessly in Markdown environments. Formatting failures constitute immediate grounds for rejection.

E.5 Appeal Handling and Quality Enforcement

- **Annotator Appeals:** Recognizing the domain expertise of annotators, the pipeline supports a formal appeal process. Reviewers are required to objectively re-evaluate contested instances based on newly provided academic evidence.

- **Accountability:** Reviewers who consistently exhibit superficial auditing ("lazy consensus"), approve logically flawed instances, or violate the double-blind protocols will be permanently removed from the reviewer pool and forfeit their compensation.

E.6 LLM-based Filtering framework

Feature Extraction Prompt

```
# Role
You are a domain expert in {{field_name}}. Your
    task is to analyze a single question and extract
    a structured summary.

# Input
Question {{question_number}}: {{question_text}}
Correct Answer: {{correct_answer}} =
    {{correct_answer_content}}

# Analysis Task
Extract the following attributes:
| Field | Description | Values |
|-------|-------------|--------|
| `knowledge_point` | The core concept being
    tested. Be specific-identify the exact
    principle, mechanism, or rule. | 10-20 words |
| `question_type` | How the question tests
    understanding | See definitions below |
| `key_topics` | Main topics/keywords for clustering
    | 2-4 terms |
| `real_world_impact` | Does this knowledge have
    significant practical consequences? | HIGH /
    MEDIUM / LOW |

## Question Type Definitions
- **CASE_BASED**: Presents a scenario
    (patient, legal case, experiment) requiring
    analysis and judgment
- **APPLIED_REASONING**: Tests ability to
    apply principles to specific situations without
    a narrative scenario

## Real-World Impact Levels

- **HIGH**: Errors in this knowledge area could
    cause significant harm, financial loss, or
    policy failures (e.g., drug interactions, legal
    precedents, safety protocols)
- **MEDIUM**: Knowledge is professionally
    important but errors have limited immediate
    consequences
- **LOW**: Primarily academic interest; minimal
    direct real-world application

# Output
Return ONLY a JSON object:
{
  "question_number": {{question_number}},
  "knowledge_point": "<specific concept>",
  "question_type": "<type>",
  "key_topics": ["<topic1>", "<topic2>"],
  "real_world_impact": "<impact>"
}

Do not include any text outside the JSON object.
```

Failure Pattern Analysis Prompt

```
# Role
You are evaluating how well frontier LLMs
    performed on a domain-specific question to
```

assess its difficulty and discriminative value.

Input

Question {{question_number}} (Preview)
{{question_preview}}

Correct Answer
{{correct_answer}}

LLM Results
{{llm_results_text}}

Analysis Task

Analyze each model's performance individually, then synthesize aggregate metrics.

Per-Model Analysis

For each model, determine:

Field	Description
`model`	Model name
`answer`	What the model answered (letter or "abstained")
`correct`	Whether the answer matches the correct answer
`error_type`	If incorrect, categorize the likely error (see categories below)
`error_hypothesis`	If incorrect, brief hypothesis for why the model failed

Error Type Categories

- **KNOWLEDGE_GAP**: Model lacks the domain-specific knowledge required
- **MISINTERPRETATION**: Model misread or misunderstood the question/options
- **REASONING_ERROR**: Model had the knowledge but applied it incorrectly
- **DISTRACTOR_TRAP**: Model fell for a plausible-sounding wrong answer
- **PARTIAL_KNOWLEDGE**: Model knew some relevant facts but not enough to discriminate
- **ABSTAINED**: Model declined to answer
- **UNKNOWN**: Cannot determine error pattern from available information

Aggregate Metrics

Field	Description
`difficulty_assessment`	Overall difficulty based on accuracy rate
`accuracy_rate`	Fraction of models that answered correctly (e.g., "2/5")
`dominant_error`	Most common wrong answer, if any (e.g., "E (3/5 incorrect)") or "varied"
`error_pattern_summary`	One-sentence description of why models failed
`discriminative_power`	Does this question separate stronger from weaker models?

Difficulty Assessment Criteria

Level	Accuracy Rate	Description
EASY	>80% correct	Most models answered correctly
MEDIUM	50-80% correct	Mixed results
HARD	20-50% correct	Most models failed
VERY_HARD	<20% correct	Nearly all models failed

Discriminative Power Criteria

- **HIGH**: Clear separation-top-tier models succeeded while weaker models failed
- **MEDIUM**: Some separation, but inconsistent patterns
- **LOW**: No meaningful separation-success/failure appears random across model tiers

Output

Return ONLY a JSON object:
```json
{
  "question_number": {{question_number}},
  "correct_answer": "{{correct_answer}}",
  "model_results": [
    {
      "model": "<model_name>",
      "answer": "<letter | 'abstained'>",
      "correct": <true | false>,
      "error_type": "<category if incorrect, else null>",
      "error_hypothesis": "<brief explanation if incorrect, else null>"
    }
  ],
  "aggregate": {
    "accuracy_rate": "<n/total>",
    "difficulty_assessment": "<EASY | MEDIUM | HARD | VERY_HARD>",
    "dominant_error": "<most common wrong answer or 'varied'>",
    "error_pattern_summary": "<one sentence explaining the failure mode, or 'N/A' if most correct>",
    "discriminative_power": "<HIGH | MEDIUM | LOW>"
  }
}
```

Do not include any text outside the JSON object.

Consensus Voting Prompt

Role
You are an expert academic assessor selecting the most representative questions for a benchmark test set.

Input

Field
{{field_name}}

Number to Select
{{num_to_select}}

Question Summaries
The following are combined summaries from question analysis (Worker 1) and LLM performance analysis (Worker 2):

{{combined_summaries}}

Selection Criteria

Criterion	Priority	Description
Knowledge Coverage	HIGH	Select questions covering diverse, distinct knowledge points. Avoid overlap.
Subject Uniqueness	HIGH	Prefer questions requiring genuine domain expertise that cannot be answered with general knowledge
Discriminative Power	HIGH	Prioritize questions that separate experts from non-experts (HIGH discriminative_power)
Practical Relevance	MEDIUM	Prefer CASE_BASED and APPLIED_REASONING over FACTUAL_RECALL
Difficulty Balance	MEDIUM	Mix of HARD and VERY_HARD; avoid EASY questions
Real-World Impact	MEDIUM	Consider questions where errors have significant consequences

Selection Rules

1. **NO OVERLAP**: Two selected questions must NOT test the same knowledge point
2. **DISCRIMINATIVE PRIORITY**: Always prefer HIGH discriminative_power questions over LOW
3. **TYPE PREFERENCE**: CASE_BASED > APPLIED_REASONING > CONCEPTUAL > FACTUAL_RECALL
4. **TOPIC DIVERSITY**: Ensure selected questions span different key_topics
5. **ERROR PATTERN VALUE**: Questions with clear, consistent error patterns are valuable for understanding model weaknesses

Output

Return ONLY a JSON object:
```json
{
  "selected_questions": [
    {
      "question_number": <int>,
      "knowledge_point": "<the knowledge point being tested>",
      "selection_reason": "<why this question was selected, referencing criteria>"
    }
  ],
  "coverage_summary": "<1-2 sentences on how selections cover the field's core knowledge>",
  "rejected_notable": [
    {
      "question_number": <int>,
      "reason": "<why this notable question was excluded>"
    }
  ]
}
```

Notes:
- ‘selected_questions‘ must have exactly {{num_to_select}} items
- ‘rejected_notable‘ should list 2-4 questions that were strong candidates but excluded (e.g., due to overlap or lower discriminative power)
- Do not include any text outside the JSON object

E.7 Agentic Workflow Verification

Diagnosis Agent Prompt

Role
You are a Critical Academic Reviewer and Logic Auditor. Your goal is to rigor-check high-difficulty benchmark questions for Large Language Models.
You have access to a DeepSearch tool.

Task
Analyze the provided [Question], [Options], [Correct Answer], and [Source Material].
You must verify the scientific validity, logical rigorousness, and linguistic clarity.

Input Data
<Question_Data>
{{JSON_DATA_HERE}}
</Question_Data>

Audit Checklist (5 Dimensions)

1. Source Confidence & Boundary Check (DeepSearch Required)
* **Objective:** Verify if the "Correct Answer" holds true *only* under specific unstated conditions.
* **Search Strategy:** Search for the core finding + keywords like "contradictory results", "limitations", "in vivo vs in vitro", "boundary conditions".
* **Failure Mode:** If another paper refutes the conclusion under a different setting (which is not excluded in the question), mark as **[CONFIDENCE_ISSUE]**.
* You MUST record the **citation (APA)** or **URL** of the paper/source that provides this contradictory evidence.

2. Logical Necessity & Uniqueness Check (DeepSearch Required)
* **Objective:** Check for "Affirming the Consequent" or "Non-Exclusivity". If the question asks for the "Cause of C", is "A" the *only* cause? Or could "B" also cause "C"?
* **Search Strategy:** Search for "causes of [Phenomenon C]" or "does [Option B] cause

[Phenomenon C]?".
* **Failure Mode:** If an incorrect option (Distractor) could also theoretically be correct under current wording, mark as **[NON_UNIQUE_ISSUE]**.

3. Ambiguity & Clarity Check
* **Objective:** Identify wording that allows for multiple valid interpretations.
* **Failure Mode:** If a key term is polysemous or the sentence structure causes scope ambiguity, mark as **[AMBIGUITY_ISSUE]**.

4. Option Value Check
* **Objective:** Identify options that are trivial, nonsensical, or fail to test the core concept (e.g., simple typo−based distractors).
* **Failure Mode:** Mark as **[LOW_VALUE_OPTION]**.

Output Format
Return a JSON object. If the question is perfect, set "has_issues" to false.

```json
{
  "has_issues": true,
  "issue_type": [
    "CONFIDENCE_ISSUE",
      "NON_UNIQUE_ISSUE",
    "AMBIGUITY_ISSUE",
      "LOW_VALUE_OPTION"
  ],
  "analysis": {
    "step_1_search_query": "What did you search for?",
    "step_2_findings": "What contradictory evidence or alternative logic did you find?",
    "evidence_sources": [
      "Author, A. A. (Year). Title. Journal. (For contradictory evidence found in Step 1)",
      "[https://valid−url.com/resource]"
    ],
    "step_2_search_query": "<What did you search to check distractors?>",
    "reasoning": "Detailed explanation of why the current question fails."
  },
  "suggestion_directive": "Specific instruction on what needs to be fixed (e.g., 'Add a constraint that the experiment is in vitro', or 'Clarify that we are asking for the primary mechanism')."
}
```

Refinement Agent Prompt

Role
You are an Expert Item Writer for a specialized scientific benchmark. Your task is to refine questions based on an Audit Report to ensure they are rigorous, logically sound, and unambiguously correct.

Input Data
1. **Original Question Data**: {{ORIGINAL_JSON}}
2. **Audit Report**: {{AUDIT_JSON_FROM_AGENT_A}}

Refinement Strategy Matrix

Apply the following strategies based on the 'issue_type' identified:

Strategy A: Fix "Source Confidence" (Premise Injection)
* **Action:** Add specific boundary conditions to the Question Stem.
* **Example:** Change "Does X cause Y?" to "In the context of [Specific Cell Line/Condition] as reported by [Author et al.], does X cause Y?" OR "Under [Assumption Z], which outcome is observed?"
* **Goal:** Make the original source's conclusion the *only* valid answer by narrowing the scope.

Strategy B: Fix "Non−Unique / Logical Necessity" (Exclusion & Precision)
* **Action:** 1. Modifiers: Use "primary cause," "most direct mechanism," or "initial trigger" instead of generic "cause."
  2. Exclusion: Add "Assuming no [Confounding Factor B]..."
* **Goal:** Eliminate the validity of the competing distractor found by the auditor.

Strategy C: Fix "Ambiguity" (Disambiguation)
* **Action:** Rewrite the confusing phrase using standard academic terminology. Ensure the syntax (e.g., modifier attachment) is singular in meaning.

Strategy D: Fix "Low Value Option" (Distractor Replacement)
* **Action:** Replace the weak option with a "plausible but incorrect" option (a common misconception or a related but distinct concept).

Strategy E: Fix "WEAK_CAUSALITY" (Assumption−based Reasoning)
* **Diagnosis:** It is hard to prove "A causes B" definitively, but "If A occurs, B follows" is a valid deduction under a specific theory.
* **Action:** Reframe the question to **assume the conclusion/premise is true**, and ask for the logical consequence or necessary condition.

Execution Rules
1. **Explanation Sync:** If you modify ANY Option or the Correct Answer, you **MUST** rewrite the corresponding 'explanation' to align with the new logic.
2. **Source Addition:** If your modification introduces new facts or constraints not supported by the original source,
you **MUST** provide a 'new_source' (APA format or accessible URL).

```
* **Example:**
    * *Before:* "Is Theory X correct?" (Hard to
        prove)
    * *After:* "**Assuming Theory X is
        correct**, which of the following
        observations would be expected?"
    OR "**Based on the standard model of [Field
        Y]**, what is the implication of Z?"

# Constraints
1. **Do NOT make the question easier.** The goal
    is rigor, not simplification.
2. **Maintain the Markdown/LaTeX format.**
3. **Evidence update:** If you add a new premise
    based on the audit, ensure the
    'question_content' reflects this (you may add
    a note "Assuming setting X").

# Output
Return a JSON object containing ONLY the fields
    that require modification.
Do not return unchanged fields (e.g., if the answer
    is unchanged, do not include it).
Example:
{
  "revision_summary":
  "Applied Strategy A: Added 'in vitro' constraint
      to the question stem; synced explanation.",
  "new_sources": [ // Include ONLY if Strategy
      A/E requires new evidence
      "Smith, J. (2024). Title. Journal.",
      "https://example.com/evidence"
  ],
  "modifications": {
      // 1. Only include keys that have changed.
      "question": "In the context of HeLa cells,
          does X cause Y?",
      "options": [
          "B":[
            "answer":"1, 3",
            "explanation":Statement1 is wrong
                because ..."
          ]
          "C":[...],
          "..."
      ],
      "source": "..."
  }
}
```

F Experiment Details

F.1 List of Models Evaluated

F.2 Evaluation Prompt

Evaluation Prompt

```
Answer the following multiple choice question.
    There is only one correct answer.
The last line of your response should be in the
    format 'Answer: $LETTER' (without quotes),
where LETTER is one of {letters}. Think step by
    step before answering.
Question:
{question}
Options:
{options}
```

F.3 Result Details

G Datasheet for KINA

We follow the structure of Gebru et al. (2021).

Motivation

For what purpose was the dataset created? KINA was created to evaluate frontier large language models on *disciplinary-representative* knowledge tasks. Existing benchmarks either prioritize scale (SuperGPQA), extreme difficulty (HLE), or narrow expert depth (GPQA), but none combine wide disciplinary coverage with an explicit representativeness criterion and an incentive-aligned annotation pipeline.

Who created the dataset and on behalf of which entity? The dataset was created by the KINA team.

Who funded the creation? [Funding source]

Composition

What do the instances represent? Each instance is a pseudo-multiple-choice question with 10 combinatorial options. An instance includes the stem, 10 options, an option-level explanation with cited sources, and the originating reference material.

How many instances are there in total? 899 items, distributed across 12 top-level disciplines, 70 fields, and 261 fine-grained subfields.

Does the dataset contain all possible instances or is it a sample? A sample. The selection process is described in §3.1 and §4.

Does the dataset identify any subpopulations? The instances are organized by discipline, field, and subfield following the CIP-aligned taxonomy (Appendix D).

Does the dataset contain data that might be considered confidential? No. All source materials are drawn from publicly published academic literature.

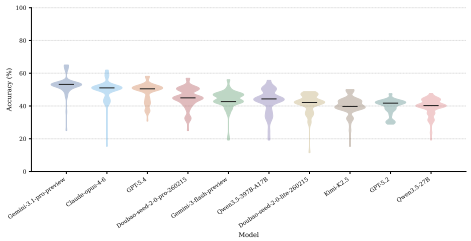
Does the dataset contain data that might be offensive, insulting, threatening, or anxiety-inducing? No.

Table 11: Overview of Evaluated Models on KINA Benchmark.

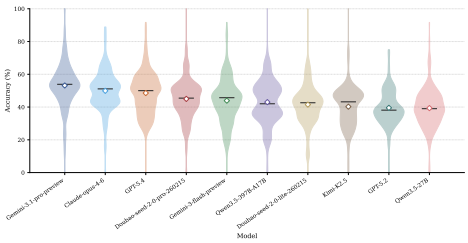
Provider / Family	Specific Models
OpenAI	GPT-5.4 (openai, 2026), GPT-5.2 (openai, 2025)
Google	Gemini-3.1-Pro-Preview , Gemini-3-Flash-Preview (Deepmind, 2026)
Anthropic	Claude-Opus-4-6 (anthropic, 2026a), Claude-Sonnet-4-6 (anthropic, 2026b)
ByteDance	Doubao-Seed-2.0-Pro-260215, Doubao-Seed-2.0-Lite-260215 (Seed, 2026)
Tongyi (Qwen)	Qwen3.5-397B-A17B, Qwen3.5-122B-A10B-FP8, Qwen3.5-35B-A3B-FP8, Qwen3.5-27B (Team, 2026) Qwen3-Max-250923 (Team, 2025), Qwen3-Next-80B-A3B-Instruct, Qwen3-Next-80B-A3B-Thinking-FP8 (Yang et al., 2025) Qwen3-235A-A22B, Qwen3-235B-A22B-Thinking-2507, Qwen3-30B-A3B, Qwen3-30B-A3B-Thinking-2507, Qwen3-32B, Qwen3-14B, Qwen3-8B, Qwen3-4B, Qwen3-4B-Thinking-2507, Qwen3-1.7B, Qwen3-0.6B (Yang et al., 2025) Qwen2.5-72B-Instruct (Yang et al., 2024a), Qwen2-72B-Instruct (Yang et al., 2024b)
Meta (Llama)	Llama-4-Maverick-17B-Instruct-FP8 (by arXiv, 2026), Llama-3.1-405B-Instruct, Meta-Llama-3-70B-Instruct (Grattafiori et al., 2024)
DeepSeek	DeepSeek-v3.2-Thinking (DeepSeek-AI et al., 2025)
Moonshot (Kimi)	Kimi-k2.5 (Team et al., 2026)
xAI	Grok-4.1-Fast-Reasoning (xAI Team, 2025)
Zhipu AI	GLM-5 (Zeng et al., 2026)
MiniMax	Minimax-M2.5 (MiniMax, 2026)
StepFun	Step-3.5-Flash (Huang et al., 2026)
Mistral	Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024)

Table 12: Evaluation Stability of KINA.

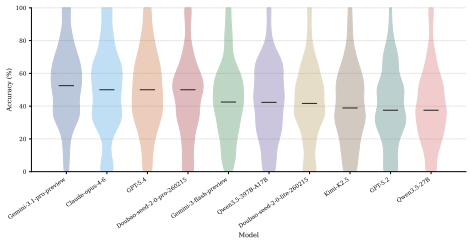
Model	Avg@4	Std	Model	Avg@4	Std
Gemini-3.1-Pro-Preview	53.17	0.72	Claude-Opus-4.6	49.92	0.32
GPT-5.4	48.55	0.56	Doubao-Seed-2.0-Pro-260215	44.99	0.80
Gemini-3-Flash-Preview	43.91	0.71	Qwen3.5-397B-A17B	42.99	0.78
Doubao-Seed-2.0-lite-260215	41.49	0.61	Kimi-K2.5	40.24	0.66
GPT-5.2	39.52	0.99	Qwen3.5-27B	39.35	0.55
Qwen3.5-122B-A10B	38.88	0.90	Deepseek-V3.2-Thinking	38.01	0.57
Qwen3-Max-2025-09-23	35.90	0.44	GLM-5	35.85	0.55
Qwen3.5-35B-A3B	35.43	0.69	Grok-4.1-Fast-Reasoning	33.73	0.64
Qwen3-235B-A22B-Thinking-2507	32.15	0.52	Llama-4-Maverick-17B-Instruct	31.62	0.59
Minimax-M2.5	30.28	0.46	Qwen3.5-9B	30.09	0.91
Qwen3-Next-80B-A3B-Instruct	30.09	0.62	Claude-Sonnet-4.6	30.01	0.56
Llama-3.1-405B-Instruct	29.59	0.62	Qwen3-235B-A22B	29.37	0.83
Step-3.5-Flash	29.12	0.52	Qwen3.5-4B	28.50	0.72
Qwen3-Next-80B-A3B-Thinking	28.28	1.33	Qwen3-32B	27.41	0.38
Qwen3-30B-A3B-Thinking-2507	27.03	0.42	Meta-Llama-3-70B-Instruct	26.61	0.39
Qwen3-14B	26.00	0.64	Qwen2-72B-Instruct	24.92	1.24
Qwen3-4B-Thinking-2507	24.83	0.37	Qwen3-30B-A3B	24.28	0.45
Qwen2.5-72B-Instruct	23.03	1.13	Qwen3-8B	22.27	1.62
Qwen3-4B	21.50	0.81	Qwen3.5-2B	20.52	0.46
Qwen3-1.7B	18.33	1.16	Mixtral-8x7B-Instruct-v0.1	17.83	0.55
Qwen3.5-0.8B	16.66	0.80	Qwen3-0.6B	14.49	0.78



(a) Top-10 Models: Discipline-Level Weighted Score Distribution



(b) Top-10 Models: Field-Level Weighted Score Distribution



(c) Top-10 Models: Subfield-Level Weighted Score Distribution

Figure 5: Top-10 Model Performance Distributions Across Three Aggregation Levels (continued)

Note. Each violin shows the distribution of weighted model scores, with weights proportional to sample size. Panels differ only in aggregation level: discipline, field, and subfield. Wider sections indicate higher score density, and the central marker indicates the median.

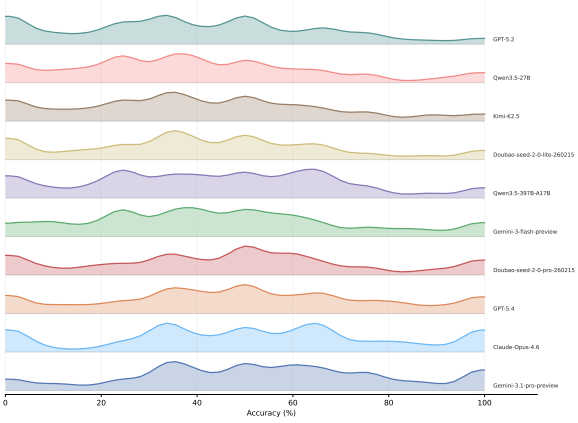


Figure 6: Subject-Level Score Distribution Across Top-10 Models

Note. Top-10 models are ranked by overall discipline score. Each ridge shows the smoothed density of one model's subject-level scores (0–100). Right-shifted and narrower ridgelines indicate higher and more stable performance across subjects.

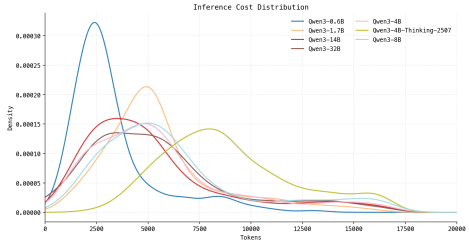


Figure 7: Inference Cost Distribution of Qwen3 Dense Models.

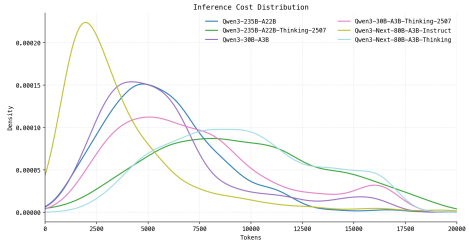


Figure 8: Inference Cost Distribution of Qwen3 Moe Models.

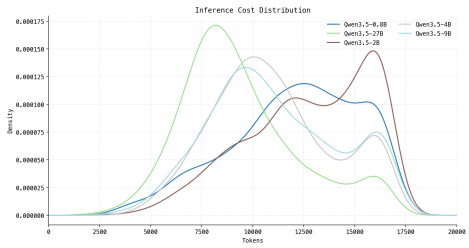


Figure 9: Inference Cost Distribution of Qwen3.5 Dense Models.

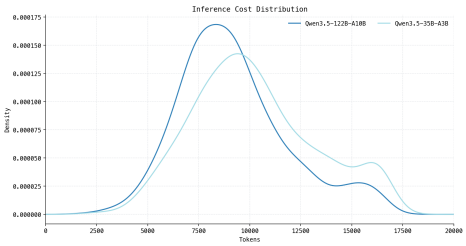


Figure 10: Inference Cost Distribution of Qwen3.5 Moe Models.

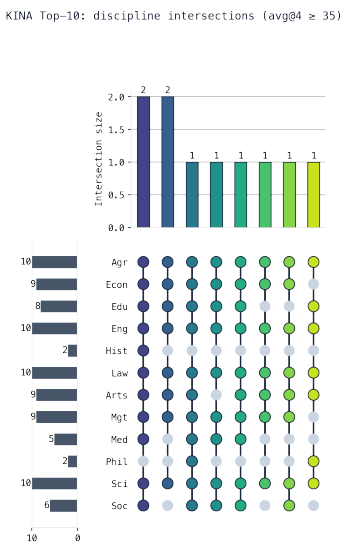


Figure 11: UpSet plot visualizing intersections of high-performing data points (average score ≥ 35) among the KINA Top-10 models. The matrix at the bottom denotes set inclusion for distinct intersection combinations, paired with horizontal bars on the left showing total set sizes per discipline. The vertical bar chart at the top specifies the sizes of these intersections.

Collection process

How was the data collected? Three-stage pipeline: rule-based screening, double-blind expert review under a bonus-on-bar tournament mechanism, and three-judge LLM consensus (§4).

Who was involved in the data collection process and how were they compensated? Annotators were graduate students from top-tier global universities and senior industry experts, recruited via a two-round examination (Appendix E.1). Compensation followed the bonus-on-bar tournament described in §3.2, calibrated per discipline.

Over what timeframe was the data collected? [Collection start month] to [collection end month].

Was an ethical review process conducted? [Yes/No, IRB details if applicable]

Preprocessing, cleaning, labeling

Was any preprocessing/cleaning of the data done? Yes. Stage 1 enforces uniqueness (cosine similarity < 0.8), formatting, and a 3-of-5 flagship-LLM-failure difficulty filter. Stage 3 LLM-as-judge applied multidimensional feature scoring. An agentic refinement loop modified $\sim\frac{1}{3}$ of the candidate pool to remove residual boundary defects.

Uses

Has the dataset been used for any tasks already? Yes: zero-shot evaluation of 42 frontier LLMs (§5), including web-search-augmented evaluation and parameter scaling analysis.

What other tasks could the dataset be used for? Calibration of LLM confidence, retrieval-augmented generation evaluation, prompt-engineering benchmarking, multilingual transfer studies (after translation), and meta-evaluation of LLM-as-judge frameworks.

Are there tasks for which the dataset should not be used? KINA should not be used as a primary signal for clinical, legal, or high-stakes deployment decisions. The benchmark is designed to differentiate *relative* model capability under controlled conditions, not to certify absolute correctness on real-world tasks.

Distribution

Will the dataset be distributed to third parties? Yes. The dataset is released under **CC-BY-4.0** and the evaluation code under **MIT license**

at https://anonymous.4open.science/r/kina-benchmark. To mitigate contamination, the test split is distributed as an encrypted archive with a canary string; the development split is distributed openly.

Will there be a leaderboard? Yes. A continuously updated public leaderboard is maintained at [leaderboardURL]. Submissions follow a hold-out protocol.

Will the dataset be updated? KINA will be refreshed (versioned as KINA-v$x.y$) when the strongest evaluated model exceeds 70% overall accuracy, to prevent saturation.

Maintenance

Who is supporting/hosting/maintaining the dataset? The KINA team commits to maintaining the dataset and leaderboard for at least 36 months from initial release. Issue tracking and contribution guidelines are at the project repository.

If others want to contribute to the dataset, is there a mechanism for them to do so? Yes. We accept community contributions of subfield-specific items via a standardized submission form, subject to the same three-stage verification pipeline described in §4.