LIVEMATHEMATICIANBENCH: A Live Benchmark for Research-Level Mathematical Reasoning with Proof Sketches

Linyang He[1,*] **Qiyao Yu**[1,*] **Hanze Dong**[2,*] **Baohao Liao**[2,3]
Xinxing Xu[2] **Micah Goldblum**[1,†] **Jiang Bian**[2,†] **Nima Mesgarani**[1,†]

[1]Columbia University [2]Microsoft Research [3]University of Amsterdam
[*]Equal contribution. [†]Senior authors.
linyang.he@columbia.edu qy2266@columbia.edu hanzedong@microsoft.com

For the latest benchmark updates, visit https://LiveMathematicianBench.github.io/

Abstract

Mathematical reasoning is widely regarded as a hallmark of human intelligence, and determining whether large language models (LLMs) can meaningfully perform it remains a central question in artificial intelligence and cognitive science. As LLMs are increasingly integrated into scientific workflows, rigorous evaluation of their mathematical capabilities becomes a practical necessity. Existing mathematical reasoning benchmarks, however, are often limited by synthetic problem settings and growing contamination from widely circulated datasets.

We present **LIVEMATHEMATICIANBENCH**, a dynamic multiple-choice benchmark for evaluating research-level mathematical reasoning using recent arXiv papers published after model training cutoffs. By grounding evaluation in newly published theorem statements, the benchmark provides a more realistic testbed for assessing whether models can reason about natural mathematical claims beyond memorized benchmark patterns. LIVEMATHEMATICIANBENCH introduces a taxonomy of thirteen problem categories based on the logical form of theorem statements, enabling fine-grained evaluation across reasoning types such as implication, equivalence, existence, and uniqueness. It further introduces a proof-sketch-guided distractor generation pipeline, in which proof sketches are used to construct plausible but invalid answer choices that reflect misleading proof directions. This design makes the benchmark more sensitive to genuine mathematical understanding rather than surface-level answer matching. We additionally introduce a substitution-resistant evaluation mechanism designed to distinguish answer recognition from substantive mathematical reasoning.

Evaluation of frontier models shows that the benchmark is far from saturated: the best-performing model, Gemini-3.1-pro-preview, achieves only 43.5% accuracy in the standard setting. Under substitution-resistant evaluation, accuracy drops sharply across all models: GPT-5.4 achieves the highest score at 30.6%, while Gemini-3.1-pro-preview falls to 17.6%, below the 20% random baseline. A dual-mode protocol comparing performance with and without proof-sketch access reveals that sketches yield consistent gains (e.g., +13.6 pp for Gemini-3.1-pro-preview), suggesting that models can leverage high-level proof strategies to improve reasoning.

Overall, LIVEMATHEMATICIANBENCH offers a scalable and contamination-resistant testbed for studying research-level mathematical reasoning in large language models.

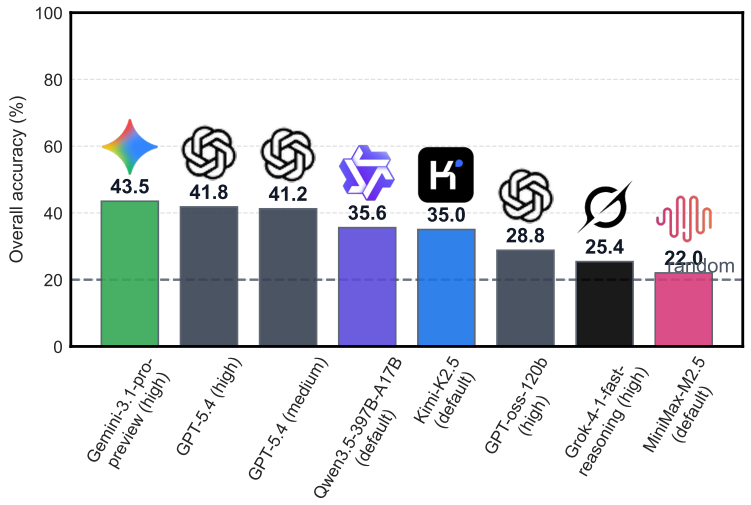

Figure 1. Overall accuracy on LIVEMATHEMATICIANBENCH across different LLMs. Even recent frontier models remain far from saturation. Random: 20%.

1 Introduction

As large language models (LLMs) become increasingly integrated into scientific workflows, evaluating their reasoning ability has become a central challenge. Mathematics is a natural domain for this evaluation because it combines precise logical structure with verifiable ground truth. Yet existing benchmarks, including GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), AIME, MathVista (Lu et al., 2023), Omni-Math (Gao et al., 2024), and OlympiadBench (He et al., 2024), are largely synthetic, competition-based, or historically sourced, and are increasingly susceptible to contamination. These limitations reduce their value for assessing research-level mathematical reasoning in frontier models.

The core limitation lies in the fundamental mismatch between "Olympiad-style" competition mathematics and authentic mathematical research. Existing benchmarks predominantly feature calculation-heavy problems that reward heuristic pattern matching and the use of specialized tricks to obtain a closed-form answer. Although challenging, such tasks do not reflect the true nature of mathematics research.

By *research-level mathematics*, we mean open-ended, non-routine mathematical reasoning that is characterized by abstraction, sensitivity to technical hypotheses, and the lack of a pre-specified solution strategy. Such work often requires one to reason through layers of definitions, infer subtle logical dependencies, and identify the right conceptual framework in which a problem can even be meaningfully posed. A central difficulty of research lies not only in proving statements, but also in determining what the right question is to ask and what potential statement might be true in the first place. Unlike competition problems, naturally arising research questions do not come with a contrived setup, an intended solution path, or even a clear indication of which tools are relevant. Progress therefore often requires exploration under substantial uncertainty, familiarity with prior literature, and occasional conceptual innovations, which are sometimes sparked by ideas from seemingly unrelated fields of mathematics.

This gap is visible even at the level of theorem comprehension. A model that can solve routine calculus or algebra problems may still fail to track the precise hypotheses of a modularity lifting theorem, a stability result in geometric invariant theory, or a compactness theorem in nonlinear PDE, where small changes in local conditions, regularity assumptions, or finiteness hypotheses can fundamentally alter the statement. Therefore, to truly assess an LLM's potential as a scientific research assistant, evaluation must shift from measuring the ability to "solve exams" toward measuring the ability to comprehend, interpret, and reason about mathematical hypotheses and statements.

This conceptual mismatch is compounded by a second problem: data contamination. State-of-the-Art (SOTA) models have reached saturation on standard benchmarks, often exceeding 90% accuracy, yet this performance is brittle. Studies suggest that many problems have leaked into pre-training corpora, allowing models to memorize solution templates rather than derive answers from first principles (Glazer et al., 2024). Consequently, performance drops precipitously when questions are rephrased or numerical values altered. This reliance on static, potentially contaminated datasets necessitates a shift toward continuous and refreshable evaluations grounded in authentic research artifacts.

Notably, Zhang et al. (2025) introduced RealMath, a benchmark derived directly from arXiv papers to reduce contamination risk. While RealMath reduces contamination risk and demonstrates that LLMs can engage with research-derived material, its construction prioritizes theorem types that admit straightforward automated verification. As a result, it covers only a limited subset of research-level mathematical reasoning, with less support for tasks involving inequalities, asymptotic relations, or non-constructive arguments, excluding common research tasks such as establishing asymptotic bounds in analysis, deriving classification results in algebraic geometry, or determining modes of convergence in probability. These limitations suggest that the central challenge is not only to control contamination risk in benchmarks, but to do so without collapsing research mathematics into a narrowly verifiable subset of problems of limited width and depth.

To address these limitations, we introduce LIVEMATHEMATICIANBENCH, a benchmark framework that expands evaluation along two dimensions: the diversity of mathematical reasoning it captures and the use of high-level proof strategies, enabling evaluation beyond statement-level recognition to strategy-level mathematical reasoning.

For the first dimension, we introduce a taxonomy of thirteen distinct problem types defined by the underlying logical forms and tailor question design to each type. Unlike RealMath, whose questions are restricted to settings with a single uniquely verifiable answer, this framework enables evaluation of mathematical skills that are central to research practice but often considered too ambiguous for automated metrics. These include **Logical Structure** (distinguishing necessary vs. sufficient conditions via Equivalence and Implication), **Qualitative Analysis** (establishing Inequalities and Asymptotic bounds), and **Generalization** (validating Universal quantifiers and structural Bijections).

For the second dimension, we incorporate proof sketches into the benchmark. A proof sketch is a high-level outline of the main ideas and strategic steps of a proof, omitting most technical details. Access to this proof structure makes question construction more targeted and more diagnostic: instead of merely testing whether a model can retrieve a theorem statement, it allows us to design questions about the same result that probe deeper structural understanding. Without such proof-aware design, questions are more likely to be too easy or to reward memorization and surface-level pattern matching. Proof sketches also enable a dual-mode evaluation protocol, where models are assessed both with and without access to high-level proof guidance. This makes it possible to measure whether strategic hints improve mathematical reasoning, and how effectively current models can use them.

Building on these two design principles, we introduce the **LIVEMATHEMATICIANBENCH**, a dynamic benchmark designed to assess LLMs on research-level mathematical reasoning. Each item is presented as a multiple-choice question (*MCQ*) with five answer options, exactly one of which is correct. Our contributions fall into three categories: benchmark construction, methodological novelty, and scientific questions enabled by the benchmark.

Our contributions are threefold. **1)** We present a contamination-resistant benchmark for research-level mathematical reasoning, built from dynamically sourced post-cutoff arXiv theorems, organized by a logic-based taxonomy, and augmented with curated proof sketches. **2)** We introduce a proof-aware benchmark construction methodology, including category-specific question design, sketch-adversarial distractor generation, and dual-mode evaluation with and without sketch access. **3)** We establish a new testbed for studying higher-level mathematical reasoning in LLMs, including logical-form-specific failure modes and the use of proof-level guidance.

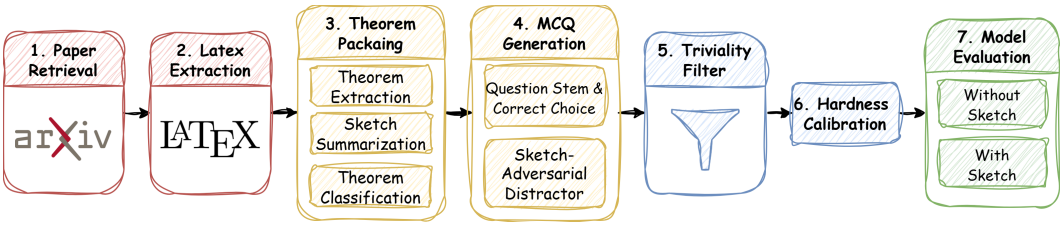

Figure 2. Our benchmark construction includes 7 stages.

2 Methods

Our benchmark construction pipeline transforms raw arXiv papers into calibrated research-level multiple-choice questions (MCQs). The pipeline contains seven stages: (1) paper retrieval, (2) LaTeX source extraction, (3) theorem extraction and classification, (4) MCQ generation, (5) stem-only triviality filtering, (6) hardness calibration, and (7) model evaluation. Each stage is designed to preserve mathematical fidelity while increasing question discriminativeness and benchmark difficulty. Human quality control combines a complete review of the 177 final benchmark items with stratified audits of intermediate pipeline outputs. We additionally conduct a small source-author validation study; see Appendix B for details.

Stage 1: Paper Retrieval. We query the arXiv API for papers submitted within a target month under the `math.*` category. The retrieval window is incrementally widened until a configurable maximum number of papers is collected. For each paper, we record its arXiv identifier, submission link, source link, and title. By restricting retrieval to papers published strictly after a model's training cutoff, we ensure that the resulting benchmark is contamination-free by construction.

Stage 2: LaTeX Source Extraction. For each retrieved paper, we download its e-print source archive and extract the constituent `.tex` files. The extractor resolves `\input` and `\include` directives to reconstruct the complete document in its logical order, strips LaTeX comments and comment environments, and concatenates the result into a single full-text representation. This produces a clean, self-contained LaTeX corpus suitable for downstream parsing.

Stage 3: Theorem Extraction and Classification.

Stage 3a: Hybrid Agentic Extraction. Extracting well-formed theorem statements from raw LaTeX is nontrivial due to the heterogeneity of document styles, custom macros, and cross-referencing conventions. We adopt a *hybrid agentic* extraction architecture that balances throughput with robustness:

1. **Rule-based fast path.** A rule-based extractor restricts its search to the *Introduction* section, exploiting the convention that a paper's main theorem is formally stated there. Within the introduction, it identifies all theorem-like environments declared via `\newtheorem` in the preamble (excluding `remark`-type environments) and applies a *priority ranking*: the standard `theorem` environment is selected first; only if no `theorem` block is found does the extractor fall back to custom environments. This design ensures that auxiliary results such as lemmas, propositions, or corollaries, which may appear elsewhere in the introduction, are not mistakenly extracted in place of the main theorem.
2. **Agentic fallback.** When the rule-based path fails (e.g., due to an absent Introduction heading or unconventional formatting), an LLM-based agentic extractor is invoked over a larger document window. The extraction prompt explicitly instructs the model to return at most *one* primary theorem-level claim and to prefer `theorem` environments over `lemma`, `proposition`, or `corollary` when multiple candidates appear, maintaining the same main-theorem priority as the rule-based path.
3. **LaTeX normalization.** Custom commands (`\newcommand`, `\def`) are expanded in-line so that downstream modules receive standard mathematical notation.

Category	Canonical Form
Algorithmic / Constructive	There exists an algorithm A that computes $f(x)$
Asymptotic / Limit	As $n \to \infty, \ldots$
Biconditional / Equivalence	$A \iff B$
Classification / Bijection	There is a bijection between X and Y
Existence	$\exists\, x$ such that $P(x)$
Existential–Universal	$\exists\, x\, \forall\, y,\ P(x, y)$
Implication	$A \implies B$
Independence / Consistency	$T \nvdash P$ and $T \nvdash \neg P$
Inequality / Bound	$f(x) \leq g(x)$
Nonexistence	$\nexists\, x$ such that $P(x)$
Uniqueness	$\exists!\, x$ such that $P(x)$
Universal	$\forall\, x,\ P(x)$
Universal–Existential	$\forall\, x\, \exists\, y,\ P(x, y)$

Table 1. Logical taxonomy of theorem categories. Each extracted theorem is classified into one of these thirteen types, enabling category-specific question design and evaluation.

4. **Reference resolution.** Internal cross-references (`\ref`, `\eqref`) are resolved by collecting all labeled environments and substituting their content, producing an *Expanded Theorem* that is semantically self-contained.
5. **Context recovery.** A two-layer retrieval mechanism assembles the notational and definitional context needed to make the theorem self-contained. *First*, within the Introduction, paragraph blocks preceding the theorem are scored by (i) token overlap with the theorem statement, (ii) the presence of definitional cues (e.g., "let," "denote," "suppose"), (iii) mathematical content density, and (iv) positional proximity, with the two blocks immediately before the theorem always retained. The top-scoring blocks are selected and concatenated in their original order. *Second*, a full-paper pass extracts high-value *anchor terms* from the theorem and scans the entire document for blocks containing definitions, setup, or preliminaries, scoring them by anchor-term hits, section-heading relevance (e.g., `Setup`, `Preliminary`), and distance to the theorem. Finally, any content referenced via `\ref` or `\eqref` within the selected blocks is resolved and appended, and the combined context is trimmed to a character budget.

Stage 3b: Proof-sketch summarization. A large language model extracts a concise proof sketch for each theorem, capturing the high-level strategy without full formal detail. These sketches serve a dual purpose: they inform adversarial distractor generation (Stage 4) and enable sketch-aware evaluation (Stage 7).

Stage 3c: Logical Taxonomy of Theorems. Each extracted theorem is classified into one of **thirteen** logical categories, listed in Table 2. Classification is performed by a specialized LLM-based classifier. This taxonomy enables fine-grained diagnosis of model reasoning failures across distinct logical forms.

Stage 4: Question–Answer Pair Generation. Stage 4 synthesizes adversarial MCQs via a **two-stage generative protocol**, each informed by the theorem's logical category and the proof sketch.

Stage 4a: Question Stem and Correct Choice. A key design principle is that every question should read as a *genuine research-type question* about a specific mathematical object or property, rather than a generic request such as "which is the correct result below." To this end, the generator selects one of **thirteen category-specific system prompts**, one for each logical type in Table 2, each prescribing a tailored stem formulation that exercises the distinguishing logical feature of the category. For example, an equivalence-type theorem is transformed into a question that presents one side of the biconditional and asks "which statement is equivalent to A?"; an implication-type theorem embeds all hypotheses in the stem and asks "which is the *strongest* conclusion about … that holds under these assumptions?", etc. With the *strongest* quantifier, we ensure there is a unique answer for questions of all logical types.

Each prompt further instructs the model to (i) isolate a single *decisive mathematical feature* of the conclusion, such as quantifier scope, sharp bound, exact dependence, existence versus uniqueness, or asymptotic regime, rather than restating the theorem verbatim, and (ii) define any paper-specific objects or notation inline so that the question is self-contained. Red-flag checks automatically reject stems containing trivializing phrases (e.g., "which of the following is the strongest result") or correct options that bundle multiple theorem clauses, triggering a regeneration pass. Finally, a *repair pass* re-examines the draft stem and, operating in `theorem_only_repair` mode, injects any missing definitions or notation from the recovered paper context (Stage 3, Step 5) to guarantee self-containedness without leaking answer-relevant information.

Stage 4b: Sketch-Adversarial Distractor Generation. The model receives the theorem, proof sketch, question stem, and correct choice, and generates **four distractors** (Options B–E). The proof sketch plays a critical role: by exposing the "load-bearing" logical steps of the argument (e.g., the dependence of constants on hidden parameters, a critical case split, or a uniformity condition that distinguishes effective from non-effective bounds), the generator is guided to engineer distractors that exploit specific mathematical constraints or misconceptions. Each distractor is constructed by one of four general perturbation strategies:

- **Controlled perturbation.** A small surface-level modification to the theorem's conclusion, such as an altered constant, exponent, inequality direction, or parameter range.
- **Semantic weakening.** A statement that is plausible but strictly weaker than the theorem's conclusion, e.g., dropping a uniformity requirement or restricting the domain.
- **Semantic strengthening.** A statement that overstates the conclusion beyond what is proven, e.g., adding uniqueness, higher regularity, or optimality not asserted by the theorem.
- **Property confusion.** A statement concerning a closely related but non-equivalent mathematical object or property, e.g., swapping an L^2 norm for an L^∞ norm, or confusing pointwise convergence with uniform convergence.

The concrete realization of each strategy is *category-dependent*: the distractor prompt is tailored to the theorem's logical type (Table 2). For instance, for an equivalence-type theorem the controlled perturbation modifies a quantifier within one side of the biconditional, whereas for a bound-type theorem it alters the exponent or direction of the inequality; for an existence-type theorem the semantic strengthening adds uniqueness not stated in the theorem, whereas for a classification-type theorem it replaces the bijection with a mere surjection. This category-aware design ensures that distractors probe the *decisive logical feature* specific to each theorem type, rather than relying on generic, one-size-fits-all perturbations. An *adversarial revision pass* subsequently re-examines the generated distractors, auditing them for surface-level distinguishability from the correct option and rewriting any that can be rejected without genuine mathematical inspection.

We evaluate the contributions of sketch access, category-aware distractor guidance, and the adversarial revision pass in Appendix E.

Substitution-Resistant Option Design. A common shortcut exploited by strong models is *option substitution*: systematically plugging each candidate answer back into the question stem and verifying consistency, rather than deriving the answer through genuine mathematical reasoning. Process-of-elimination strategies similarly bypass comprehension by narrowing choices through superficial cues. To counteract both shortcuts, we introduce a *substitution-resistant* mechanism: for a configurable fraction of generated items, the correct option is replaced with the meta-option *"One of the remaining options is correct, but a stronger result can be proven."* Answering correctly now requires the model to (i) identify the valid option among the distractors and (ii) independently determine whether a strictly stronger result holds, a task that demands substantive mathematical reasoning about the relative strength of claims and cannot be resolved by mechanical substitution or elimination alone.

Quality Rubric. Each generated MCQ undergoes a quality evaluation scored on a 0–8 rubric comprising four dimensions:

- **Answer Leakage Score (ALS, 0–2):** Measures whether the correct answer is inadvertently revealed by the question stem or distractor phrasing.
- **Tautology Avoidance Score (TAS, 0–2):** Assesses whether the correct option is trivially true or self-evident without mathematical knowledge.
- **Generative Pressure Score (GPS, 0–2):** Evaluates whether distractors exert sufficient "pressure" to discriminate genuine understanding from guessing.
- **Distractor Quality Score (DQS, 0–2):** Rates the mathematical plausibility, precision, and diversity of the distractors.

MCQs that do not meet a minimum aggregate threshold (5) are discarded.

Stage 5: Rote-Recall Prevented Triviality Filter. Stage 5 ensures that the remaining questions are not trivially solvable from the question stem alone, without reference to the answer choices. A judge model receives *only* the question stem (with all options withheld) and generates a free-form response. A second judge then determines whether this response matches the correct answer (Option A). Questions for which the stem alone reveals the answer are classified as *stem-trivial* and excluded from the final candidate pool. This filter prevents the benchmark from rewarding rote recall and ensures that genuine option-level reasoning is required.

Stage 6: Hardness Calibration. Stage 6 ensures that the final benchmark is genuinely difficult for state-of-the-art models through a three-step calibration pipeline:

Step 1: Overgenerate Hard Pool. For each stem-nontrivial item, we retain both the original distractor set and one regenerated alternative set, thereby doubling the candidate pool per source theorem. This overgeneration increases the likelihood of retaining at least one maximally challenging variant.

Step 2: First-Pass Accuracy Test. A frontier model is evaluated on the entire candidate pool at moderate reasoning effort. Items that the model answers correctly are flagged as potentially too easy.

Step 3: Source-Level Hardest Selection. For each source theorem, the pipeline compares all surviving candidate variants and selects the **hardest**, defined as the variant the calibration model answered incorrectly while achieving the highest quality rubric score. Source groups in which all candidates were solved are dropped entirely. This procedure yields the final benchmark subset.

Stage 7: Evaluation Protocol. We propose a dual-mode evaluation protocol: **Mode 1: Selection.** The model receives the question stem and all five options, then selects the answer that it judges to be the correct theorem statement. This mode evaluates the model's baseline mathematical comprehension. **Mode 2: Sketch-Aware Selection.** The model additionally receives the proof sketch as a hint alongside the stem and options. By comparing performance across modes, we isolate the model's ability to synthesize and apply high-level proof strategies, a proxy for *mathematical intuition*.

To ensure a fair rendering of notations, all questions undergo LaTeX validation via compilation and, when necessary, automated LaTeX repair before presentation to the model.

3 Benchmark

We summarize the logical-form composition of the hard benchmark split in Figure 3. The hard split contains 177 theorems in total, and a single theorem may belong to multiple logical categories. The benchmark therefore covers a broad range of theorem structures rather than concentrating on a single template: implication and universal statements are the most common categories, followed by existence and inequality/bound problems. The month-level breakdown further shows that benchmark composition changes over time

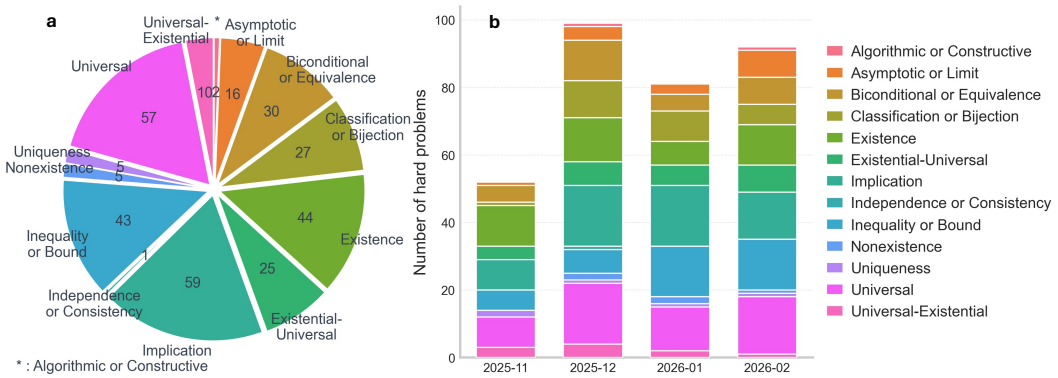

Figure 3. Category statistics for the hard split of LIVEMATHEMATICIANBENCH. **(a)** aggregate distribution of logical categories across all months. **(b)** month-by-month category breakdown from November 2025 to February 2026. Counts correspond to category memberships rather than unique questions, since a single theorem can receive multiple logical labels.

while preserving category diversity, with all slices drawn from papers released after the evaluated models' knowledge cutoffs. Examples can be found in the Appendix C.

4 Evaluation Results

Even frontier models suffer. Figure 1 shows overall accuracy on LIVEMATHEMATICIANBENCH. The strongest performance is achieved by Gemini-3.1-pro-preview at 43.5% overall, followed by GPT-5.4 (high) at 41.8%. Even the strongest systems remain far from saturation.

Category-wise pattern and monthly trajectories. Figure 4-(a) shows that no single model dominates every reasoning type. Gemini-3.1-pro-preview achieves the strongest overall profile and is particularly competitive on biconditional/equivalence and classification/bijection problems, whereas GPT-5.4 variants are stronger on implication, universal, and inequality/bound items. These categories combined constitute a substantial portion of the benchmark (Figure 3). This heterogeneity suggests that research-level mathematical reasoning is not a one-dimensional capability: models with similar aggregate performance can still exhibit sharp weaknesses on particular logical forms. The monthly trajectories in Figure 4-(b) further show that performance is unstable across benchmark slices. Several models do best on the 2026/01 subset but perform worse again on the 2026/02 subset.

Substitution-resistant options sharply reduce accuracy. Figure 5-(a) compares model accuracy on the original-choice subset and the substitution-resistant subset, where the correct answer is replaced by a stronger meta-option intended to frustrate shortcut strategies such as direct substitution and surface-level elimination. Across all evaluated systems, the substitution-resistant setting is substantially harder than the original-choice setting. The strongest original-choice performance is achieved by Gemini-3.1-pro-preview at 67.4%, but its accuracy drops to only 17.6% on substitution-resistant items, below the 20% random-guess baseline. GPT-5.4 provides a useful contrast: although it ranks second overall, it reaches 52.2% on original items under both medium and high reasoning effort and still retains 29.4%–30.6% on the substitution-resistant subset. GPT-5.4 is therefore more stable under the harder choice design, remaining above the random-guess baseline even when the benchmark removes the direct theorem statement as an option, whereas most other evaluated models fall to chance level or below. More broadly, the remaining difficulty in LIVEMATHEMATICIANBENCH lies not only in understanding the theorem itself, but also in reasoning among nearby mathematical claims once surface matching and answer substitution become ineffective.

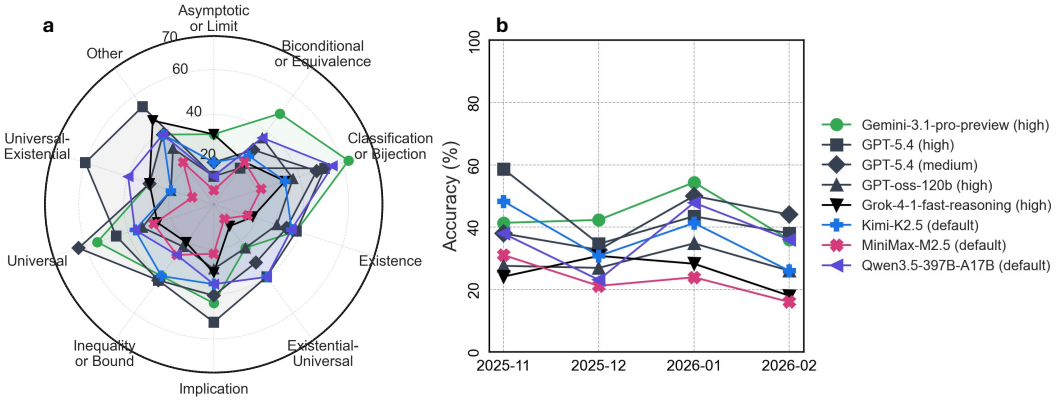

Figure 4. **(a)** accuracy by logical category aggregated across benchmark months; rare categories with fewer than ten examples are grouped into *Other*. **(b)** monthly overall accuracy from November 2025 to February 2026.

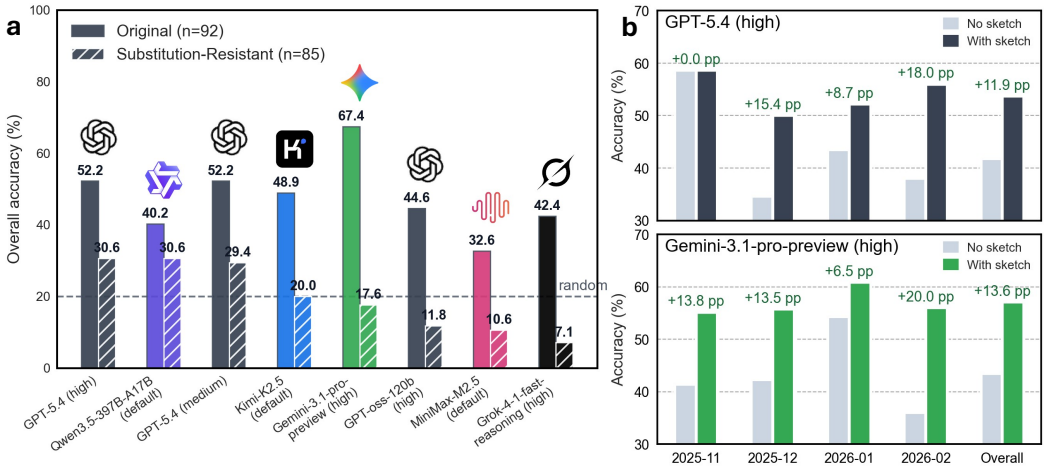

Figure 5. **(a)** Accuracy by answer-choice format on LIVEMATHEMATICIANBENCH. Each model is evaluated separately on the original-choice subset and the substitution-resistant subset. Ranked by substitution-resistant performance. **(b)** Accuracy gain from adding proof sketches. Each panel compares the same model setting with and without sketch access across benchmark months.

Proof sketches yields a consistent improvement. As shown in Figure 5-(b), GPT-5.4 (high) improves from 41.8% to 53.7% overall, a gain of 11.9% points, while Gemini-3.1-pro-preview (high) rises from 43.5% to 57.1%, a gain of 13.6% points. This pattern suggests that proof sketches are not merely redundant restatements of the theorem. Instead, they provide useful strategic information that helps models disambiguate closely related answer choices and identify the load-bearing structure of the argument.

5 Endnote and Discussion

From aggregate accuracy alone, Gemini-3.1-pro-preview (high) achieves the highest score, with GPT-5.4 variants close behind by less than 3%. However, a category-level analysis in Section 4 (Fig. 4(a)) reveals additional structure that supports the first central design choice of LIVEMATHEMATICIANBENCH: organizing problems by logical type. Performance varies meaningfully across logical forms in ways that are not visible from aggregate scores alone. This suggests that logical structure may be an important axis along which current models differ, and it raises a broader interpretability question: why are some models systematically stronger on certain forms of reasoning than others? If such patterns persist at

scale, they may also have practical implications for mathematical workflows, since different models may prove more useful for different kinds of reasoning tasks. In this sense, a logic-based taxonomy complements traditional field-based categorization by capturing structural similarities that cut across mathematical areas. More broadly, this suggests that LIVEMATHEMATICIANBENCH can serve not only as a leaderboard benchmark, but also as a diagnostic testbed for studying structure-specific strengths and weaknesses in mathematical reasoning. This diagnostic value is especially relevant because the benchmark is constructed from recent post-cutoff arXiv papers, making it less vulnerable to rapid saturation and contamination than static mathematics benchmarks. In this sense, LIVEMATHEMATICIANBENCH contributes not only a new dataset, but also a more durable evaluation setting for tracking progress in research-level mathematical reasoning.

Substitution-resistant options (Stage 4b) are designed to test whether a model can recognize that the *semantically weakened* option is true while also inferring that a strictly stronger statement follows from the hypotheses. Success on such questions therefore requires an additional deductive step. In this sense, substitution-resistant questions provide a useful probe of whether a model can move beyond recognition toward a limited form of mathematical conjecturing. This is precisely the kind of ability mathematicians hope AI systems may leverage on when facilitating research, as illustrated by the work of Georgiev et al. (2025) using AlphaEvolve for mathematical exploration and discovery. As highlighted in Section 4, all tested models perform substantially worse with substitution-resistant options (Fig. 5(a)). This gap suggests that current models rely heavily on option substitution in standard multiple-choice settings. More importantly, it underscores the importance of adversarial option design for isolating genuine reasoning ability. At its core, it shows that benchmark design can materially affect the capabilities we appear to measure. From considerations of both overall benchmark performance and performance under substitution-resistant evaluation, GPT-5.4 (high) and Gemini-3.1-pro-preview (high) emerge as the two strongest models on LIVEMATHEMATICIANBENCH.

Finally, we have observed earlier that providing proof sketches at inference time induces significant performance improvements in the two highest-scoring models (Fig. 5(b)). Since proof sketches are extracted from the source papers and typically summarize the central ideas of the proof at a high level, this result suggests that models can often use strategic mathematical guidance without full derivational detail, demonstrating some similarities to how expert researchers work. These findings further support the second design philosophy of LIVEMATHEMATICIANBENCH: incorporating proof strategy into evaluation rather than restricting assessment to statement-level recognition alone. More broadly, they indicate that the benchmark can distinguish not only whether a model reaches the correct answer, but also how its performance changes when supplied with information closer to the level of mathematical strategy used by human researchers. Collectively, these features make LIVEMATHEMATICIANBENCH useful not only for comparing current models, but also for future work on proof-guided reasoning, tool use, and interactive mathematical exploration.

References

Mohammed Abouzaid, Andrew J. Blumberg, Martin Hairer, Joe Kileel, Tamara G. Kolda, Paul D. Nelson, Daniel Spielman, Nikhil Srivastava, Rachel Ward, Shmuel Weinberger, and Lauren Williams. First proof, 2026. URL https://arxiv.org/abs/2602.05192.

Yuchen Bi and Jie Zhou. Quantitative stability of the clifford torus as a willmore minimizer, 2025. URL https://arxiv.org/abs/2511.19681.

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

Simon Frieder. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024.

Bogdan Georgiev, Javier Gómez-Serrano, Terence Tao, and Adam Zsolt Wagner. Mathematical exploration and discovery at scale, 2025. URL https://arxiv.org/abs/2511.02864.

Julius Giesler. Jacobian rings and the infinitesimal torelli theorem, 2026. URL https://arxiv.org/abs/2601.17765.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*, 2024.

Torben Grabbel, Gebhard Martin, Giacomo Mezzedimi, Maia Raitz von Frentz, and Paul Jakob Schmidt. On symmetries of hyperbolic lattices of large rank, 2026. URL https://arxiv.org/abs/2602.05652.

Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno Dumont, Elyas Obbad, and Sanmi Koyejo. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning in llms. *arXiv preprint arXiv:2508.08292*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

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

Vitezslav Kala and Om Prakash. There are consecutive cubic fields with large class numbers, when ordered by discriminant, 2026. URL https://arxiv.org/abs/2601.03984.

Eyal Kaplan. A note on the ketonen order and lipschitz reducibility between ultrafilters, 2025a. URL https://arxiv.org/abs/2512.12835.

Eyal Kaplan. The number of normal measures, revisited. *arXiv preprint arXiv:2507.20466*, 2025b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Diego Marques and Pavel Trojovsky. Geometric progressions meet zeckendorf representations, 2025. URL https://arxiv.org/abs/2512.19586.

Juan Felipe Ariza Mejía, Ionuţ Chifan, Denis Osin, and Bin Sun. Mcduff superrigidity for group ii_1 factors, 2025. URL https://arxiv.org/abs/2511.23123.

Walner Mendonça, Meysam Miralaei, and Guilherme O. Mota. Graphs with asymmetric ramsey properties, 2025. URL https://arxiv.org/abs/2511.02963.

Joaquín Moraga and Juan Pablo Zúñiga. Degenerations of cluster type varieties, 2026. URL https://arxiv.org/abs/2511.14959.

Yanmin Mu and Dehua Wang. Global existence for the relativistic vlasov-poisson system in a two-dimensional bounded domain, 2025. URL https://arxiv.org/abs/2511.06595.

Makoto Ozawa. Crossing numbers of knots on closed surfaces, 2026. URL https://arxiv.org/abs/2602.21659.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J Nay, Kshitij Gupta, and Aran Komatsuzaki. Arb: Advanced reasoning benchmark for large language models. *arXiv preprint arXiv:2307.13692*, 2023.

Quanyu Tang and Shengtong Zhang. Harmonic lcm patterns and sunflower-free capacity, 2025. URL https://arxiv.org/abs/2512.20055.

George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. Putnambench: A multilingual competition-mathematics benchmark for formal theorem-proving. In *AI for Math Workshop@ ICML 2024*, 2024.

Jinmin Wang. Non-negative scalar curvature on spin surgeries and novikov conjecture, 2025. URL https://arxiv.org/abs/2512.16535.

XTX Markets. The ai mathematical olympiad (aimo) prize, 2024. URL https://aimoprize.com/. Accessed: 2024-05-01.

Jie Zhang, Cezara Petrui, Kristina Nikolić, and Florian Tramèr. Realmath: A continuous benchmark for evaluating language models on research-level mathematics. *arXiv preprint arXiv:2505.12575*, 2025.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*, 2021.

A Related Work

Standard and Advanced Mathematical Benchmarks The evaluation of mathematical reasoning in large language models has traditionally relied on static datasets spanning elementary to undergraduate-level problems. Early benchmarks such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) helped establish common evaluation standards for quantitative and symbolic reasoning. As model performance has improved, however, these benchmarks have become less discriminative for frontier systems.

Subsequent work has therefore shifted toward more challenging settings. The Advanced Reasoning Benchmark (ARB) (Sawada et al., 2023) and GHOSTS (Frieder, 2023) extend evaluation toward graduate-level mathematics, with GHOSTS further emphasizing natural language interaction and expert assessment. Other efforts focus on competition-level mathematics, including OlympiadBench (He et al., 2024), PutnamBench (Tsoukalas et al., 2024), and Omni-Math (Gao et al., 2024). These benchmarks substantially increase difficulty, but because they are derived from historical competition archives, they may be more susceptible to contamination from pretraining data and benchmark-specific memorization.

Contamination Mitigation, Formal Mathematics, and Frontier-Level Evaluation Several recent benchmarks attempt to mitigate contamination more directly. Putnam-AXIOM (Gulati et al., 2025) perturbs existing problems through programmatic modifications such as renaming variables or altering constants. While this strategy can reduce exact overlap with known solutions, it preserves much of the original problem structure. The AI Mathematical Olympiad (AIMO) (XTX Markets, 2024) instead introduces newly created olympiad-style problems, and MiniF2F (Zheng et al., 2021) focuses on formally verified mathematics across theorem proving systems such as Lean and Isabelle. These directions address important aspects of evaluation reliability, but they remain centered on competition-style or formalized problem settings.

Moving beyond competition settings, First Proof (Abouzaid et al. (2026)) curates a collection of ten original, research-level problems drawn directly from the authors' active work. While this approach effectively guarantees the absence of data contamination, the extremely small sample size limits its utility as a comprehensive evaluation benchmark.

FrontierMath (Glazer et al., 2024) is more closely related to our setting in that it targets highly difficult mathematics under reduced contamination risk through a larger set of expert-authored problems. However, similar to First Proof, tailored problem authoring requires substantial dedicated human effort, which inherently limits scalability.

LIVEMATHEMATICIANBENCH differs in both construction and evaluation emphasis. Rather than relying on manually authored problems, it draws from a continuously growing stream of recent arXiv papers, enabling a scalable and refreshable benchmark construction pipeline. In addition, whereas prior benchmarks largely emphasize answer derivation, our framework explicitly evaluates structural comprehension through theorem-level logical categories and proof-sketch-guided distractors that probe sensitivity to plausible but invalid reasoning directions.

B Human Validation Protocol

All prompts used in theorem extraction, classification, question generation, and filtering were authored by a research-level mathematician. For quality control, the final released set of 177 benchmark items was reviewed for theorem fidelity, self-containedness, answer uniqueness, and distractor quality. To audit upstream pipeline quality, we additionally performed stratified reviews of intermediate outputs, sampling approximately 10 cases per logical category for theorem extraction, category assignment, proof-sketch quality, and MCQ generation. Given the unusual breadth and depth of the mathematical content covered by the pipeline, these reviews were intended as benchmark quality control, assessing whether outputs were mathematically sound, internally consistent, and benchmark-ready, rather than as full peer-review-level sense of independent verification of every source-paper claim.

B.1 Source-Author Validation

To complement the internal review, we conducted a validation study with authors of the source papers.

For each item included in the study, we sent the source-paper authors the extracted theorem, question stem, proof sketch, and multiple-choice options.

We asked them to assess:

(i) whether the item provided sufficient definitions and context;

(ii) whether it captured the theorem's load-bearing mathematical content; and

(iii) whether the question stem and selected answer accurately rephrased the original theorem statement.

Across 10 valid responses, the mean scores were 7.8 out of 10 for context sufficiency, 7.4 out of 10 for capturing the theorem's load-bearing content, and 7.3 out of 10 for rephrasing accuracy.

These results provide preliminary expert evidence that the generated items are aligned with their source theorems.

The study is limited by its small sample size and does not constitute independent verification of every source-paper theorem or every released benchmark item.

Scope of Correctness Validation. LIVEMATHEMATICIANBENCH evaluates whether generated items faithfully represent claims stated in recent source papers.

It does not independently certify that every theorem in those papers is mathematically correct.

Requiring peer-review-level verification of every source claim would substantially delay benchmark construction and weaken its freshness and contamination resistance.

The pipeline mitigates this risk by selecting theorem statements with accompanying proof context, reviewing final items for fidelity and answer uniqueness, and obtaining feedback from source-paper authors.

These safeguards reduce, but do not eliminate, errors introduced during extraction or inherited from the source papers.

We now provide an example demonstrating how the combination of prompt design and human validation ensures that all distractors are demonstrably false, thereby guaranteeing the overall validity of the MCQs.

In Appendix C, Example 6 is generated from an Existential-Universal type theorem, and the correct option generated from the theorem is replaced by the correct substitution-resistant option. When generating distractors, the system prompt (see Appendix H, Stage 4b) forces the LLM to include a meta description for each distractor indicating the type of perturbation strategies used in its construction. For Example 6, the meta description reads:

```
"meta": {
     "weaker_true_label": "C",
     "false_labels": [
       "B",
       "D",
       "E"
     ],
     "wildcard_false_label": "B"
   },
   "sketch_usage_meta": [
     {
       "label": "B",
```

```
      "sketch_hook_type": "counting_estimate",
      "tampered_component": "counts signed cubic discriminants instead of
      cubic fields",
      "template_used": "wildcard"
    },
    {
      "label": "C",
      "sketch_hook_type": "finiteness",
      "tampered_component": "dropped the lower bound guaranteeing
      at least k/2 cubic fields in the interval",
      "template_used": "weaker_true"
    },
    {
      "label": "D",
      "sketch_hook_type": "counting_estimate",
      "tampered_component": "threshold k/2 replaced by (k+1)/2",
      "template_used": "stronger_trap"
    },
    {
      "label": "E",
      "sketch_hook_type": "regularity",
      "tampered_component": "removed the epsilon-loss in the exponent",
      "template_used": "stronger_trap"
    }
  ]
```


We first notice that option **C** is the "weaker_true_label", meaning that it is strictly weaker than the theorem's assertion (which is always true if we assume the theorem itself is true). However, since the question asks for the 'strongest' true statement, option **C** is not correct.

Option **B** is the "wildcard_false_label", meaning that it is designed to be false but with a question specific tempering. It counts the number of integers within that range that are determinants of cubic fields. This is clearly not the same as counting the number of cubic fields, as multiple non-isomorphic fields can share the same discriminant. Therefore, it is not realistic for a smaller set to have the same lower bound. This reasoning aligns with the "tempered_component" field provided in the meta description. Moreover, this option critically uses the proof sketch.

Option **D** and **E** are false labels tempered with "stronger_trap", meaning that they are strengthened versions of the theorem statement in different ways. Option **D** increases the lower bound of the set being counted, making the statement stronger. However, the lower bound is obtained by explicit construction and a bigger lower bound cannot be achieved. Option **E** modifies the count for the specific integers d by removing the epsilon factor in the exponent. The epsilon-dependence in the original theorem comes from deep conjectures in the subject of counting number fields, which cannot be removed.

To conclude, taking the original theorem as the ground truth, we can see from a combination of meta description and human expertise that the distractors are false and the MCQ is valid. In particular, this example demonstrates that our prompt design alone has guaranteed the mathematical falsity of the distractors to a great extent.

C Detailed Logical Taxonomy of Theorem Categories

This appendix provides extended descriptions of the thirteen logical theorem categories listed in Table 2. The taxonomy focuses exclusively on the *logical form* of a theorem statement, independent of its mathematical topic or context. Each category is accompanied by its canonical logical form and a representative example.

1. Algorithmic / Constructive. *Canonical form:* There exists an algorithm A that computes $f(x)$, or "object X can be explicitly constructed." This category covers theorems providing an

explicit or computable method to obtain an object. *Example:* "There exists a polynomial-time algorithm that computes a maximum matching in any bipartite graph."

LIVEMATHEMATICIANBENCH Example

Original Theorem (Marques & Trojovsky (2025))i@

Let $u \geq 1$, $q \geq 2$ be integers, and let $\mathcal{F}$ be a finite family of non-empty binary words. Let $L = \max\{|f| : f \in \mathcal{F}\}$ and fix $M \geq L$. Then the set

$$S_u^{(M)} = \{n \geq 0 : \ uq^n \in \mathcal{K}_{\mathcal{F}}^{(M)}\}$$

is either finite or ultimately periodic. That is, there exist integers $n_0 \geq 0$ and $p \geq 1$ such that

$$n \in S_u^{(M)} \iff n + p \in S_u^{(M)} \qquad (n \geq n_0).$$

Moreover, one can effectively compute such a pair (n_0, p) and decide whether $S_u^{(M)}$ is finite or infinite. In the finite case, the same procedure outputs all elements of $S_u^{(M)}$.

Proof Sketch

To prove the main theorem, the argument is "entirely finite-state". The key input is that "multiplication by a fixed integer q in the Fibonacci numeration system is realizable by a finite sequential transducer", and when "reading digits from the least significant side" this yields (for each M) "a deterministic update rule on the set of length-M windows $\{0, 1, \#\}^M$." Composing this with "the finite automaton that checks avoidance of $\mathcal{F}$ inside the window" produces "a deterministic dynamical system on a finite state space." The orbit corresponding to $(uq^n)_{n\geq 0}$ is thus "eventually periodic", and "which yields ultimate periodicity of $S_u^{(M)}$ as an immediate corollary." Finally, "the same finite-state construction gives an explicit algorithm to compute the eventual period" (hence to compute (n_0, p) and decide finiteness, and list elements in the finite case).

LIVEMATHEMATICIANBENCH Example 1 (Question & Answers)

Generated Question

Let $(F_n)_{n\geq 0}$ be the Fibonacci numbers, defined by $F_0 = 0$, $F_1 = 1$, and $F_{n+2} = F_{n+1} + F_n$. Every integer $N \geq 1$ has a unique Zeckendorf representation

$$N = \sum_{i=2}^{k} \varepsilon_i F_i,$$

where $\varepsilon_i \in \{0,1\}$, $\varepsilon_k = 1$, and $\varepsilon_i \varepsilon_{i+1} = 0$. Write $\mathrm{rev}(Z(N)) = \varepsilon_2 \varepsilon_3 \cdots \varepsilon_k$ for the Zeckendorf digits read from least significant to most significant, and let $\widetilde{Z}(N) = \mathrm{rev}(Z(N))\,\#^{\omega}$, where $\# \notin \{0,1\}$ is a padding symbol.
Let $u \geq 1$ and $q \geq 2$ be integers, let $\mathcal{F}$ be a finite family of non-empty binary words, let $L = \max\{|f| : f \in \mathcal{F}\}$, and fix $M \geq L$. Define

$$\mathcal{K}_{\mathcal{F}}^{(M)} = \{N \geq 1 : \text{ the prefix of length } M \text{ of } \widetilde{Z}(N) \text{ contains no factor from } \mathcal{F}\},$$

and

$$S_u^{(M)} = \{n \geq 0 : uq^n \in \mathcal{K}_{\mathcal{F}}^{(M)}\}.$$

Which is the strongest conclusion that can be justified about $S_u^{(M)}$?

Answer Choices

A. $S_u^{(M)}$ is either finite or ultimately periodic: there exist integers $n_0 \geq 0$ and $p \geq 1$ such that

$$n \in S_u^{(M)} \iff n + p \in S_u^{(M)} \qquad (n \geq n_0).$$

Moreover, one can effectively compute such a pair (n_0, p) and decide whether $S_u^{(M)}$ is finite or infinite; in the finite case, the same procedure outputs all elements of $S_u^{(M)}$. ✓

B. $S_u^{(M)}$ is either finite or ultimately periodic: for each fixed q, $\mathcal{F}$, and M, there exist integers $n_0 \geq 0$ and $p \geq 1$, independent of u, such that for every $u \geq 1$ one has

$$n \in S_u^{(M)} \iff n + p \in S_u^{(M)} \qquad (n \geq n_0).$$

Moreover, one can effectively compute such a universal pair (n_0, p) and decide for each u whether $S_u^{(M)}$ is finite or infinite; in the finite case, the same procedure outputs all elements of $S_u^{(M)}$.

C. $S_u^{(M)}$ is ultimately periodic: there exist integers $n_0 \geq 0$ and $p \geq 1$ such that

$$n \in S_u^{(M)} \iff n + p \in S_u^{(M)} \qquad (n \geq n_0).$$

Moreover, one can effectively compute such a pair (n_0, p).

D. $S_u^{(M)}$ is either finite or ultimately periodic: there exist integers $n_0 \geq 0$ and $p \geq 1$ such that

$$n \in S_u^{(M)} \iff n + p \in S_u^{(M)} \qquad (n \geq n_0).$$

Moreover, one can always choose these with $n_0 < 3^M$ and $1 \leq p \leq 3^M$, and effectively compute such a pair together with a decision procedure for whether $S_u^{(M)}$ is finite or infinite; in the finite case, the same procedure outputs all elements of $S_u^{(M)}$.

E. $S_u^{(M)}$ is either finite or ultimately periodic: there exist integers $n_0 \geq 0$ and $p \geq 1$ such that

$$n \in S_u^{(M)} \iff n + p \in S_u^{(M)} \qquad (n \geq n_0).$$

Moreover, one can effectively decide whether $S_u^{(M)}$ is finite or infinite and, in the infinite case, compute an eventual period p; however, in general there is no procedure that always computes a valid preperiod n_0 or lists all elements when $S_u^{(M)}$ is finite.

2. Asymptotic / Limit. *Canonical form:* As $n \to \infty$ (or another limiting regime), a quantity behaves as described. This category covers theorems describing limiting or asymptotic behavior, including exact limits, asymptotic equivalences ($\sim$), and big-O/little-o estimates. *Example:* "The number of primes less than x satisfies $\pi(x) \sim x/\ln x$ as $x \to \infty$."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Tang & Zhang (2025))
Fix an integer $k \geq 3$. Then, as $N \to \infty$,

$$f_k(N) \geq (\log N)^{c_k - o(1)},$$

where

$$c_k := \frac{k-2}{e((k-2)!)^{1/(k-2)}}.$$

Proof Sketch
To prove the main theorem, they use a "prime bucketing" construction: partition primes in a suitable interval into many disjoint blocks of comparable harmonic sum, and form squarefree integers by choosing exactly $k-2$ primes from each block. A short combinatorial lemma then forces LCM-k-freeness, while the harmonic sum factorizes and can be optimized over a free parameter.

LIVEMATHEMATICIANBENCH Example 2 (Question & Answers)

Generated Question
Fix an integer $k \geq 3$. For $N \geq 1$, write $[N] := \{1, 2, \ldots, N\}$, and define

$$f_k(N) := \max\left\{\sum_{a \in A} \frac{1}{a} : A \subseteq [N] \text{ is LCM-}k\text{-free}\right\},$$

where A is called LCM-k-free if there do not exist distinct $a_1, \ldots, a_k \in A$ such that all pairwise least common multiples are equal, i.e.

$$\operatorname{lcm}(a_i, a_j) = \operatorname{lcm}(a_1, a_2) \qquad (1 \leq i < j \leq k).$$

Define

$$c_k := \frac{k-2}{e\,((k-2)!)^{1/(k-2)}}.$$

As $N \to \infty$, which is the strongest asymptotic lower bound that holds for the quantity $f_k(N)$?

Answer Choices
A. One of the remaining options is correct, but a stronger result can be proven. ✓
B.

$$f_k(N) \geq (\log N)^{c_k + o(1)} \qquad (N \to \infty),$$

with $c_k = \dfrac{k-2}{e\,((k-2)!)^{1/(k-2)}}$.
C.

$$f_k(N) \geq (\log N)^{c_k - o(1)} \qquad (N \to \infty),$$

for some positive constant c_k depending only on k.
D.

$$f_k(N) \geq (\log N)^{c_k - o(1)}$$

for every sufficiently large N, where now $c_k = \dfrac{k-2}{((k-2)!)^{1/(k-2)}}$.
E.

$$f_k(N) \geq (\log N)^{\frac{k-1}{e\,((k-1)!)^{1/(k-1)}} - o(1)} \qquad (N \to \infty).$$

3. Biconditional / Equivalence. *Canonical form:* $A \iff B$, or "the following are equivalent." This category covers theorems asserting that two (or more) conditions are logically equivalent. The theorem may present the equivalence as a single biconditional or as a list of mutually equivalent statements. *Example:* "A ring R is Noetherian if and only if every ideal of R is finitely generated."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Bi & Zhou (2025))

Let $\Sigma \subset S^3$ be a smooth embedded surface of genus at least 1. Then its canonical five-parameter family

$$\{\Sigma_{(v,t)}\}_{(v,t)\in \mathring{B}^4\times[-\pi,\pi]}$$

satisfies

$$\max_{(v,t)\in \mathring{B}^4\times[-\pi,\pi]} \operatorname{Area}\big(\Sigma_{(v,t)}\big) \;\geqslant\; 2\pi^2.$$

Moreover, equality holds if and only if Σ is the Clifford torus.

Proof Sketch

Using the Marques–Neves five-parameter family $\{\Sigma_{(v,t)}\}$, the key point is to show that if a genus–one surface Σ satisfies $\mathcal{W}(\Sigma) \leqslant 2\pi^2 + \delta^2$, then after a *conformal* transformation one gets a surface Σ' with $\operatorname{Area}(\Sigma') \geqslant 2\pi^2 - C\delta^2$.

To prove the main theorem, we use that the full (v,t)-family contains a slice of area at least $2\pi^2$, and since v encodes conformal transformations and t is signed distance, the aim is to realize such a high-area slice *within the conformal* four-parameter subfamily; for this one must control t in terms of the energy gap δ, so the maximizing slice lies close to $t = 0$.

A closer examination of the Heintze–Karcher inequality used by Marques–Neves shows that, once the family is built from the quantitatively normalized surface, the problem reduces to proving the maximizing parameter (v_0, t_0) stays a definite distance away from the boundary of the conformal group, i.e. there exists $\eta_1 > 0$ such that $|v_0| \leqslant 1 - \eta_1$ (reduced later; technical proof deferred to an appendix).

Geometric intuition for this uniform bound: as $|v| \to 1$, $\Sigma_{(v,t)}$ becomes quantitatively close to a geodesic sphere, forcing area strictly below $2\pi^2$, so such v cannot maximize. More precisely, for $t = 0$, $\Sigma_v = \Sigma_{(v,0)}$ is Hausdorff close to a geodesic sphere S, so $\operatorname{Area}(\Sigma_v)$ is close to $\operatorname{Area}(S) \leq 4\pi$, and for small $|t|$ one uses

$$\big|\operatorname{Area}(\Sigma_{(v,t)}) - \operatorname{Area}(\Sigma_v)\big| \lesssim |t|\, \mathcal{W}(\Sigma_v)$$

so the area stays $< 5\pi$. For larger $|t|$, either (i) if Σ_v stays away from the poles, an Euclidean-model argument shows $\Sigma_{(v,t)}$ is a small Lipschitz graph over a geodesic sphere hence has area $< 5\pi$, or (ii) if Σ_v lies close to a pole, comparison geometry gives $\operatorname{Area}(\Sigma_{(v,t)}) < 5\pi$. Thus for $|v|$ sufficiently close to 1, $\operatorname{Area}(\Sigma_{(v,t)}) < 5\pi < 2\pi^2$, proving $|v_0| \leq 1 - \eta_1$.

With this uniform control of v, the remaining task is to derive a quantitative stability estimate for minimal surfaces; the paper then uses the earlier area lower bound to reduce Willmore stability to minimal-surface stability, and later performs linearization/stability analysis to obtain the quantitative bounds (conformal factor, conformal structure, and $W^{2,2}$-distance to the Clifford torus), thereby proving the theorem.

LIVEMATHEMATICIANBENCH Example 3 (Question & Answers)

Generated Question

Let $\Sigma \subset \mathbb{S}^3 \subset \mathbb{R}^4$ be a smooth embedded surface in the round 3-sphere, with genus at least 1. For $v \in \mathring{B}^4 = \{v \in \mathbb{R}^4 : |v| < 1\}$, let

$$F_v(x) = (1 - |v|^2)\frac{x - v}{|x - v|^2} - v$$

be the associated conformal transformation of $\mathbb{S}^3$, and set $\Sigma_v := F_v(\Sigma)$. If $\mathbb{S}^3 \setminus \Sigma = A \cup A^*$ is the decomposition into connected components, let $A_v := F_v(A)$, let d_v be the signed distance to Σ_v (positive on A_v, negative on the other component), and define

$$\Sigma_{(v,t)} = \partial\{x \mid d_v(x) < t\}, \qquad (v,t) \in \mathring{B}^4 \times [-\pi, \pi].$$

Thus $\{\Sigma_{(v,t)}\}_{(v,t)\in\mathring{B}^4\times[-\pi,\pi]}$ is the canonical five-parameter family of Σ. Which quantitative area estimate holds for this family, and in which case does equality occur?

Answer Choices

A.

$$\max_{(v,t)\in\mathring{B}^4\times[-\pi,\pi]} \mathrm{Area}(\Sigma_{(v,t)}) \geq 2\pi^2,$$

and equality holds if and only if Σ is the Clifford torus. ✓

B.

$$\max_{(v,t)\in\mathring{B}^4\times[-\pi,\pi]} \mathrm{Area}(\Sigma_{(v,t)}) \leq 2\pi^2,$$

and equality holds if and only if Σ is the Clifford torus.

C.

$$\max_{(v,t)\in\mathring{B}^4\times[-\pi,\pi]} \mathrm{Area}(\Sigma_{(v,t)}) \geq 2\pi^2.$$

D.

$$\max_{v\in\mathring{B}^4} \mathrm{Area}(\Sigma_v) \geq 2\pi^2,$$

and equality holds if and only if Σ is the Clifford torus.

E.

$$\max_{(v,t)\in\mathring{B}^4\times[-\pi,\pi]} \mathrm{Area}(\Sigma_{(v,t)}) \geq 2\pi^2,$$

and equality holds whenever Σ is conformally equivalent to the Clifford torus.

4. Classification / Bijection. *Canonical form:* There is a bijection between X and Y, or "objects of type A are completely classified by objects of type B." This category covers theorems establishing a one-to-one correspondence or a complete classification between two classes of mathematical objects. *Example:* "Finite simple groups are classified into cyclic groups of prime order, alternating groups, groups of Lie type, and 26 sporadic groups."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Giesler (2026))
Let Δ be an n-dimensional lattice polytope with $\mathrm{Int}(\Delta) \cap M \nsubseteq$ hyperplane. Then

$$\begin{aligned}\ker(d\mathcal{P}^k_{B,f}) \overset{\mathrm{mod}\,\mathrm{Lie}(T)}{=} \Big\langle g_\Gamma(f) \cdot x^w \in L(\Delta)/\mathbb{C} \cdot f \quad | \quad & \Gamma \leq \Delta \text{ a facet}, \\ & w + v \in \big(\mathrm{Int}((k+1)\cdot\Delta) \cup \mathrm{Int}((k+1)\cdot\Gamma)\big) \cap M, \\ & \forall\, v \in \mathrm{Int}(k \cdot \Delta) \cap M \Big\rangle.\end{aligned}$$

where $\langle\rangle$ denotes the span as $\mathbb{C}$-vector space.

Proof Sketch
This direct approach to the diagram of Griffiths

$$\begin{tikzcd} T_{B,f} \arrow[r, "\kappa_f"] \arrow[dr, "d\mathcal{P}^k_{B,f}"'] & H^1(Y_f, T_{Y_f}) \arrow[d, "\Phi^k_f"] \\ & \mathrm{Hom}(H^k(Y_f, \Omega^{n-1-k}_{Y_f}), H^{k+1}(Y_f, \Omega^{n-2-k}_{Y_f})) \end{tikzcd}$$

"yields a both computational and much explicit perspective." Under the assumptions of the main theorem, one "easily deduce[s]

$$\ker(d\mathcal{P}^k_{B,f}) = \ker(d\mathcal{P}^1_{B,f}), \quad k \geq 1$$

We record the precise statement as follows.
Corollary . *Under the assumption of theorem we have*

$$\ker(d\phi_f) = \ker(d\phi^1_f).$$

For an element $g = g_\Gamma(f) \cdot x^w \in \ker(d\mathcal{P}^1_f)$ one obtains the distinction

$$\langle w, n_\Gamma \rangle = \begin{cases} 0 & \Leftrightarrow\ g \equiv 0, \\ -1 & \Leftrightarrow\ g \in \ker(\kappa_f), \\ \leq -2 & \text{exceptional cases}, \end{cases}$$

and also "$\langle w, n_{\Gamma'} \rangle \geq 0$ for $\Gamma' \neq \Gamma$ by the main theorem." The "last case" $\langle w, n_\Gamma \rangle \leq -2$ "causes $\Phi_{f|\mathrm{Im}\,\kappa_f}$ to be not injective, which implies in particular that the infinitesimal Torelli theorem (ITT) fails."
For $n = 3$, to rule out the exceptional case and conclude injectivity, the text sketches: "we construct a 3-dimensional empty polytope Q with 6 vertices," namely "the convex span of an empty triangle with vertex $(0,0,0)$ and this triangle dilated by w," and then "use a theorem of White on empty 3-simplices to deduce $|Q \cap M| \geq 7$ under the assumption $\langle w, n_\Gamma \rangle \leq -2$, contradiction."

LIVEMATHEMATICIANBENCH Example 4 (Question & Answers)

Generated Question

Let $M \cong \mathbb{Z}^n$ be a lattice, $T = \operatorname{Spec} \mathbb{C}[M] \cong (\mathbb{C}^*)^n$, and let $f = \sum_{m \in \Delta \cap M} a_m x^m$ be a nondegenerate Laurent polynomial with n-dimensional Newton polytope Δ. Write $L(\Delta) := \bigoplus_{m \in \Delta \cap M} \mathbb{C}\, x^m$, and let $\operatorname{Int}(\cdot)$ denote interior (relative interior on faces such as facets Γ). Assume that $\operatorname{Int}(\Delta) \cap M$ is not contained in any hyperplane. In the notation where $d\mathcal{P}^k_{B,f}$ is the differential under consideration and $g_\Gamma(f)$ is the element associated to a facet $\Gamma \leq \Delta$, which subspace gives $\ker(d\mathcal{P}^k_{B,f})$ modulo $\operatorname{Lie}(T)$?

Answer Choices

A. Modulo $\operatorname{Lie}(T)$, it is exactly the $\mathbb{C}$-span of all elements $g_\Gamma(f)\, x^w \in L(\Delta)/\mathbb{C} \cdot f$ such that $\Gamma \leq \Delta$ is a facet and $w + v \in \big(\operatorname{Int}((k+1)\Delta) \cup \operatorname{Int}((k+1)\Gamma)\big) \cap M$ for every $v \in \operatorname{Int}(k\Delta) \cap M$. ✓

B. Modulo $\operatorname{Lie}(T)$, it is exactly the $\mathbb{C}$-span of all elements $g_\Gamma(f)\, x^w \in L(\Delta)/\mathbb{C} \cdot f$ such that $\Gamma \leq \Delta$ is a facet and there exists $v \in \operatorname{Int}(k\Delta) \cap M$ with $w + v \in \big(\operatorname{Int}((k+1)\Delta) \cup \operatorname{Int}((k+1)\Gamma)\big) \cap M$.

C. Modulo $\operatorname{Lie}(T)$, it is contained in the $\mathbb{C}$-span of all elements $g_\Gamma(f)\, x^w \in L(\Delta)/\mathbb{C} \cdot f$ such that $\Gamma \leq \Delta$ is a facet and $w + v \in \big(\operatorname{Int}((k+1)\Delta) \cup ((k+1)\Gamma)\big) \cap M$ for every $v \in \operatorname{Int}(k\Delta) \cap M$.

D. Modulo $\operatorname{Lie}(T)$, it is exactly the $\mathbb{C}$-span of all elements $g_\Gamma(f)\, x^w \in L(\Delta)/\mathbb{C} \cdot f$ such that $\Gamma \leq \Delta$ is a facet and $w + v \in \big(\operatorname{Int}((k+1)\Delta) \cup \operatorname{Int}((k+1)\Gamma)\big) \cap M$ for every $v \in k\Delta \cap M$.

E. Modulo $\operatorname{Lie}(T)$, it is exactly the $\mathbb{C}$-span of all elements $g_\Gamma(f)\, x^w \in L(\Delta)/\mathbb{C} \cdot f$ such that $\Gamma \leq \Delta$ is a facet and $w + v \in \big(\operatorname{Int}((k+1)\Delta) \cup \operatorname{Int}((k+1)\Gamma)\big) \cap M$ for every $v \in \operatorname{Int}(k\Gamma) \cap M$.

5. Existence. *Canonical form:* $\exists\, x$ such that $P(x)$. This category covers theorems guaranteeing that at least one object with a given property exists, without claiming uniqueness. *Example:* "There exists a continuous, nowhere-differentiable function on $[0, 1]$."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Mejía et al. (2025))

There exists a family of countable W^*-McDuff groups $\{G_i\}_{i \in I}$ such that $|I| = 2^{\aleph_0}$, $G_i \ncong G_j$ whenever $i \neq j$, and every G_i is McDuff superrigid.

Proof Sketch

The groups G_i are described as "infinite direct sums of W^*-superrigid groups with property (T) constructed in I. Chifan, A. Ioana, A. O. Sasyk, and D. Stefănescu, CIOS21." Beyond this, the text only states that "[t]he proof of Theorem To prove the main theorem, incorporates several new ideas on both the group-theoretic and analytic sides" without outlining any steps or structure.

LIVEMATHEMATICIANBENCH Example 5 (Question & Answers)

Generated Question

A countable discrete group G is called a W^*-McDuff group if its group von Neumann algebra $L(G)$ is a McDuff II_1 factor. Here, G is called McDuff superrigid in the strongest sense if whenever $L(G) \cong L(H)$ for some group H, one has $H \cong G \times A$ for some ICC amenable group A (ICC means every nontrivial conjugacy class is infinite). Which is the strongest existence statement that holds for countable W^*-McDuff groups?

Answer Choices

A. There exists a family $\{G_i\}_{i \in I}$ of countable W^*-McDuff groups with $|I| = 2^{\aleph_0}$ such that $G_i \ncong G_j$ whenever $i \neq j$, and every G_i is McDuff superrigid. ✓

B. There exists a family $\{G_i\}_{i \in I}$ of countable W^*-McDuff groups with $|I| = 2^{\aleph_0}$ such that $G_i \ncong G_j$ whenever $i \neq j$, and every G_i is McDuff superrigid in the strongest sense.

C. There exists a countably infinite family $\{G_i\}_{i \in I}$ of pairwise non-isomorphic countable W^*-McDuff groups such that every G_i is McDuff superrigid.

D. There exists a family $\{G_i\}_{i \in I}$ of countable ICC groups with $|I| = 2^{\aleph_0}$ such that $G_i \ncong G_j$ whenever $i \neq j$, and every G_i is McDuff superrigid.

E. For every countable W^*-McDuff group G, there exists a family $\{G_i\}_{i \in I}$ of countable W^*-McDuff groups with $|I| = 2^{\aleph_0}$ such that $G_i \ncong G_j$ whenever $i \neq j$, every G_i is McDuff superrigid, and G is isomorphic to one of the G_i.

6. Existential–Universal. *Canonical form:* $\exists\, x\ \forall\, y,\ P(x, y)$. This category covers theorems asserting the existence of an object (often a constant, function, or parameter) that "works" uniformly for all elements of a given class. The existential quantifier precedes the universal one. *Example:* "There exists a constant $C > 0$ such that for all $f \in L^2(\mathbb{R})$, $\|f\|_\infty \leq C\,\|f\|_{H^1}$."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Kala & Prakash (2026))

Fix a sign choice $\pm$, $\varepsilon > 0$, and positive integers k, H. There exists a positive constant c (depending on $\pm, \varepsilon, k, H$) such that for all sufficiently large X, there are at least $cX^{1-\kappa-\varepsilon}$ positive integers $d \leq X$ such that the following hold:

1. For each $1 \leq i \leq k$, if there exists a cubic field K with $\pm\Delta_K = d + i$, then $h(K) > H$.
2. We have
$$\#\{K \text{ cubic field } \mid \pm\Delta_K \in \{d+1, d+2, \ldots, d+k\}\} \geq \frac{k}{2}.$$

Proof Sketch

We will prove this result in **Theorem .** *Assume Setting (S). There exists a positive constant c (depending on $\pm, \varepsilon, k, H$) such that for all sufficiently large X, there are at least $cX^{1-\kappa-\varepsilon}$ positive integers $t \leq X$ such that the following hold:*

(1) For each $1 \leq i \leq k$, if there exists a cubic field K with $\pm\Delta_K = a + tm + i$, then $h(K) > H$.

(2) We have
$$\#\{K \text{ cubic field } \mid \pm\Delta_K \in \{a+tm+1, a+tm+2, \ldots, a+tm+k\}\} \geq \frac{k}{2}.$$

, and the proof is described as constructive. The approach is:

- First, "construct integers $0 \leq a < m$ such that whenever there exists a cubic field K with $\pm\Delta_K = d + i$, where $d \equiv a \pmod{m}$, then $h(K) > H$." This is obtained "by controlling the *genus number* which gives a lower bound for the size of the 2-torsion in the class group of K." The motivation given is that this "sidesteps the difficult issue of estimating the regulator" (so there is no need to "restrict ourselves to a suitable family of cubic fields"), though consequently "we only show that the class numbers are larger than some prescribed bound."
- Then, "show that indeed there are many cubic fields with $\pm\Delta_K \in \{d+1, d+2, \ldots, d+k\}$, which also ensures that our first assertion is not vacuous."

It is also noted that, in counting the integers $d \leq X$ in establishing the main theorem, one must "switch between counting cubic fields and cubic discriminants," and that "non-isomorphic cubic fields can have the same discriminant," i.e. discriminants are counted with multiplicity when counting fields.

LIVEMATHEMATICIANBENCH Example 6 (Question & Answers)

Generated Question

For a cubic field (that is, a degree-3 number field) K, let Δ_K denote its discriminant and $h(K)$ its class number. Fix once and for all either the plus sign or the minus sign, so that $\pm\Delta_K$ means $+\Delta_K$ throughout or $-\Delta_K$ throughout according to that choice; also fix a real number $\varepsilon > 0$ and positive integers k and H, and let κ be the fixed constant appearing in the exponent below. Which is the strongest uniform existence statement that holds for all sufficiently large X?

Answer Choices

A. One of the remaining options is correct, but a stronger result can be proven. ✓

B. There exists a positive constant c depending on $\pm, \varepsilon, k, H$ such that for all sufficiently large X, there are at least $cX^{1-\kappa-\varepsilon}$ positive integers $d \leq X$ such that the following hold: for each $1 \leq i \leq k$, if there exists a cubic field K with $\pm\Delta_K = d + i$, then $h(K) > H$; and moreover

$$\#\{n \in \{d+1, d+2, \ldots, d+k\} \mid n \text{ is a cubic discriminant with sign } \pm\} \geq \frac{k}{2}.$$

C. There exists a positive constant c depending on $\pm, \varepsilon, k, H$ such that for all sufficiently large X, there are at least $cX^{1-\kappa-\varepsilon}$ positive integers $d \leq X$ for which, for each $1 \leq i \leq k$, if there exists a cubic field K with $\pm\Delta_K = d + i$, then $h(K) > H$.

D. There exists a positive constant c depending on $\pm, \varepsilon, k, H$ such that for all sufficiently large X, there are at least $cX^{1-\kappa-\varepsilon}$ positive integers $d \leq X$ such that the following hold: for each $1 \leq i \leq k$, if there exists a cubic field K with $\pm\Delta_K = d + i$, then $h(K) > H$; and moreover

$$\#\{K \text{ cubic field} \mid \pm\Delta_K \in \{d+1, d+2, \ldots, d+k\}\} \geq \frac{k+1}{2}.$$

E. There exists a positive constant c depending on $\pm, \varepsilon, k, H$ such that for all sufficiently large X, there are at least $cX^{1-\kappa}$ positive integers $d \leq X$ such that the following hold: for each $1 \leq i \leq k$, if there exists a cubic field K with $\pm\Delta_K = d + i$, then $h(K) > H$; and moreover

$$\#\{K \text{ cubic field} \mid \pm\Delta_K \in \{d+1, d+2, \ldots, d+k\}\} \geq \frac{k}{2}.$$

7. Implication. *Canonical form:* $A \implies B$, or "if P holds, then Q holds." This category covers theorems in which a set of hypotheses implies a specific conclusion, without asserting the converse. *Example:* "If f is a continuous function on a closed interval $[a, b]$, then f is uniformly continuous on $[a, b]$."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Moraga & Zúñiga (2026))

Let $\pi\colon \mathcal{X} \to \mathbb{D}$ be a projective fibration. Let $(\mathcal{X}, \mathcal{X}_0 + \mathcal{B})$ be a log Calabi–Yau pair over $\mathbb{D}$ for which $(\mathcal{X}, \mathcal{X}_0)$ is plt. If $(\mathcal{X}_t, \mathcal{B}_t)$ is a toric pair[a] for $t \in \mathbb{D}^*$, then $(\mathcal{X}_0, \mathcal{B}_0)$ is a finite quotient of a toric pair.

Proof Sketch

In the proof of the main theorem, we will see that $(\mathcal{X}, \mathcal{B}) \to \mathbb{D}$ is indeed a finite quotient of an isotrivial toric family. Thus, in the toric case, there are not many interesting degenerations for which the degeneration of the toric boundary still has lc singularities.

[a] Meaning that $\mathcal{X}_t$ is a projective toric variety and $\mathcal{B}_t$ is the reduced torus invariant boundary.

LIVEMATHEMATICIANBENCH Example 7 (Question & Answers)

Generated Question

Let $\pi\colon \mathcal{X} \to \mathbb{D}$ be a projective morphism to the unit disk $\mathbb{D}$, with central fiber $\mathcal{X}_0 = \pi^{-1}(0)$. Let $\mathcal{B}$ be a divisor such that $(\mathcal{X}, \mathcal{X}_0 + \mathcal{B})$ is a log Calabi–Yau pair over $\mathbb{D}$ (so $K_{\mathcal{X}} + \mathcal{X}_0 + \mathcal{B}$ is trivial over $\mathbb{D}$), and assume that $(\mathcal{X}, \mathcal{X}_0)$ is plt (purely log terminal). For each $t \in \mathbb{D}^* = \mathbb{D} \setminus \{0\}$, write $\mathcal{X}_t = \pi^{-1}(t)$ and $\mathcal{B}_t = \mathcal{B}|_{\mathcal{X}_t}$. Suppose that for every $t \in \mathbb{D}^*$, the fiber pair $(\mathcal{X}_t, \mathcal{B}_t)$ is toric, meaning that $\mathcal{X}_t$ is a projective toric variety and $\mathcal{B}_t$ is its reduced torus-invariant boundary. Under these assumptions, which is the strongest statement about the special fiber pair $(\mathcal{X}_0, \mathcal{B}_0)$ that holds?

Answer Choices

A. The special fiber pair $(\mathcal{X}_0, \mathcal{B}_0)$ is a finite quotient of a toric pair. ✓
B. The special fiber pair $(\mathcal{X}_0, \mathcal{B}_0)$ becomes a toric pair after normalizing $\mathcal{X}_0$ and replacing $\mathcal{B}_0$ by the induced reduced boundary on the normalization.
C. The special fiber pair $(\mathcal{X}_0, \mathcal{B}_0)$ is birational to a toric pair.
D. After a finite base change of the disk, the pulled-back degeneration has toric central fiber; equivalently, $(\mathcal{X}_0, \mathcal{B}_0)$ becomes toric after such a base change.
E. There exists a finite morphism of pairs $(\mathcal{X}_0, \mathcal{B}_0) \to (Y, D)$ onto a toric pair (Y, D).

8. Independence / Consistency. *Canonical form:* $T \nvdash P$ and $T \nvdash \neg P$, or "statement P is independent of axiom system T." This category covers theorems expressing a meta-mathematical statement that a proposition cannot be proved or refuted from given axioms. *Example:* "The Continuum Hypothesis is independent of ZFC."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Kaplan (2025a))

Consistently from a measurable cardinal, there are two σ-complete ultrafilters $\mathcal{V}, \mathcal{W}$ such that $\mathcal{V}$ is Lipschitz below $\mathcal{W}$, but $\mathcal{V}$ and $\mathcal{W}$ are Ketonen-incomparable. Furthermore, the same is also consistent with the assumption that the Weak Ultrapower Axiom holds.

Proof Sketch

We will rely on the construction from Kaplan (2025b) in the proof of the main theorem. The paper then proceeds by: (1) defining the Ketonen and Lipschitz orders and outlining their basic properties; (2) showing, consistently, that the Ketonen and Lipschitz orders do not coincide (not in a model of the Weak UA). We first prove the following theorem.

Theorem . *Assume $V = L[U]$, κ is the unique measurable cardinal, $\mathbb{P}$ is the forcing notion described above and G, G^*, U_0, U_1 are as above. Denote $\mathcal{V} = U_0$ and $\mathcal{W}$ the measure derived from $j^{V[G]}_{(U_1)^2}$ using the ordinal $j_U(\kappa) + \kappa$ as a seed. Then $\mathcal{V} <_L \mathcal{W}$ but $\mathcal{V}, \mathcal{W}$ are Ketonen incomparable.*

(3) sketching basic properties of forcing with nonstationary support products; and (4) separating the Ketonen and Lipschitz orders in a model of the Weak UA, thereby completing the proof of the main theorem.

LIVEMATHEMATICIANBENCH Example 8 (Question & Answers)

Generated Question

Assume the consistency strength of a measurable cardinal. A σ-complete ultrafilter is an ultrafilter closed under countable intersections. Using the standard comparison notions $<_L$ (the Lipschitz order) and $<_k$ (the Ketonen order) on σ-complete ultrafilters, where "Ketonen-incomparable" means that neither ultrafilter is below the other in $<_k$, and writing WUA for the Weak Ultrapower Axiom, which is the strongest consistency statement that holds about σ-complete ultrafilters?

Answer Choices

A. Relative to a measurable cardinal, it is consistent that there exist two σ-complete ultrafilters $\mathcal{V}$ and $\mathcal{W}$ such that $\mathcal{V}$ is Lipschitz below $\mathcal{W}$, while $\mathcal{V}$ and $\mathcal{W}$ are Ketonen-incomparable; moreover, it is also consistent that such a situation occurs together with the Weak Ultrapower Axiom. ✓
B. Relative to a measurable cardinal, it is consistent that there exist two σ-complete ultrafilters $\mathcal{V}$ and $\mathcal{W}$ such that $\mathcal{V}$ is Lipschitz below $\mathcal{W}$, while $\mathcal{V}$ and $\mathcal{W}$ are Ketonen-incomparable; however, no model satisfying the Weak Ultrapower Axiom can contain such a pair.
C. Relative to a measurable cardinal, it is consistent that there exist two σ-complete ultrafilters $\mathcal{V}$ and $\mathcal{W}$ such that $\mathcal{V}$ is Lipschitz below $\mathcal{W}$, while $\mathcal{V}$ and $\mathcal{W}$ are Ketonen-incomparable.
D. Relative to a measurable cardinal, it is consistent that there exist two σ-complete ultrafilters $\mathcal{V}$ and $\mathcal{W}$ such that $\mathcal{V}$ is Lipschitz below $\mathcal{W}$ and the Lipschitz and Ketonen orders do not coincide on σ-complete ultrafilters; nevertheless, in every model of the Weak Ultrapower Axiom, any relation $\mathcal{V} <_L \mathcal{W}$ already implies that $\mathcal{V}$ and $\mathcal{W}$ are comparable in the Ketonen order.
E. Relative to a measurable cardinal, it is consistent that the Lipschitz and Ketonen orders on σ-complete ultrafilters can fail to coincide, but only outside the Weak Ultrapower Axiom; under WUA, the Ketonen order is always linear on σ-complete ultrafilters.

9. Inequality / Bound. *Canonical form:* $f(x) \leq g(x)$ (or $\geq, <, >$), holding universally. This category covers theorems providing a quantitative estimate, such as a uniform bound, norm inequality, or comparison between quantities. *Example:* "For all $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, $\|\hat{f}\|_2 = \|f\|_2$."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Ozawa (2026))

Theorem (Fundamental inequality)**.** Let $F \subset S^3$ be a closed surface. Then every knot $K \subset S^3$ satisfies:

$$c(K;F) \geq 2\big(t(K) - \delta(F)\big) + 1,$$

where $t(K)$ is the tunnel number of K, and $\delta(F) = g(M_1) + g(M_2) - g(F)$ is the Heegaard deficiency of F.

Proof Sketch

To prove the main theorem, the argument "relies on a chain of inequalities relating the surface crossing number $c(K;F)$, the surface ascending number $a(K;F)$, and the surface bridge number $b(K;F)$":

$$\frac{c(K;F)-1}{2} \geq a(K;F) \geq b(K;F) - 1 \geq t(K) - \delta(F).$$

This sequence yields the stated lower bound for $c(K;F)$ in terms of "a combination of the knot exterior's complexity, $t(K)$, and the ambient surface's Heegaard deficiency, $\delta(F)$."
For optimality, the introduction says they "prove that the linear lower bound in the main theorem is asymptotically optimal" by taking iterated connected sums K_m of a prime knot and obtaining

$$c(K_m;F) \geq 2m - 2\delta(F) + 1.$$

"Pairing this with a general upper bound derived from the subadditivity of the planar crossing number yields $c(K_m;F) = \Theta(t(K_m))$," so "no general lower bound of higher order exists," and $c(K;F)$ grows linearly with tunnel number.

LIVEMATHEMATICIANBENCH Example 9 (Question & Answers)

Generated Question

Let $F \subset S^3$ be a closed surface, and let M_1 and M_2 be the two components of $S^3 \setminus F$. Define the Heegaard deficiency of F by

$$\delta(F) = g(M_1) + g(M_2) - g(F),$$

where $g(M_i)$ denotes the Heegaard genus of M_i and $g(F)$ denotes the genus of F. For a knot $K \subset S^3$, let $c(K;F)$ denote the surface crossing number of K with respect to F, i.e. the minimal number of crossings among all regular diagrams of K obtained by isotoping K into a regular neighborhood of F. If $t(K)$ is the tunnel number of K, which is the strongest quantitative estimate that holds for $c(K;F)$?

Answer Choices

A. For every knot $K \subset S^3$,

$$c(K;F) \geq 2\big(t(K) - \delta(F)\big) + 1. \checkmark$$

B. For every knot $K \subset S^3$,

$$c(K;F) \geq 2\big(t(K) - \delta(F) + 1\big) + 1.$$

C. For every knot $K \subset S^3$,

$$c(K;F) \geq 2\big(t(K) - \delta(F)\big).$$

D. For every knot $K \subset S^3$,

$$c(K;F) \geq 2t(K) - \delta(F) + 1.$$

E. For every knot $K \subset S^3$,

$$c(K;F) \geq 2\big(t(K) - \delta(F)\big) + 2.$$

10. Nonexistence. *Canonical form:* $\nexists\, x$ such that $P(x)$ (or equivalently, $\forall\, x,\ \neg P(x)$). This category covers theorems asserting that no object satisfies a given property. *Example:* "There does not exist a rational number whose square is 2."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Wang (2025))

Let M^n be a closed aspherical manifold and N a spin manifold. Assume that the rational strong Novikov conjecture holds for $\pi_1(M)$. Then the connected sum $M\#N$ does not carry any complete metric with positive scalar curvature. Furthermore, any complete metric with non-negative scalar curvature on $M\#N$ is Ricci-flat.

Proof Sketch

Our proofs rely on Dirac operator methods. A fundamental ingredient is the relative index theorem for higher indices due to Xie and Yu (Theorem A in Xie–Yu, *[title not provided in ref_dict]* (2014)). Inspired by the approaches in Su–Zhang, *[title not provided in ref_dict]*, and Wang–Xie–Linf, *[title not provided in ref_dict]*, we construct an explicit representative of the relative higher index and compute it in the K-theory of the group C^*-algebra. We also provide a Dirac-operator-based proof of the Ricci-flatness statement, following ideas similar to those in Wang–Zhu, *[title not provided in ref_dict]*. (In establishing the main theorem, we treat it as a special case of

Theorem . *Let M^n be a closed aspherical manifold and $\Omega \subset M$ a smooth region. Assume that*

$$\pi_1(\Omega) \to \pi_1(M) \text{ is the trivial map.}$$

Let N^n be a spin manifold with boundary, and $i\colon \partial N \to \partial\Omega$ a diffeomorphism preserving the spin structure[a]*. Set $Z := (M\backslash\Omega) \cup_{\partial\Omega} N$. If the rational strong Novikov conjecture holds for $\pi_1(M)$, then Z does not carry any complete metric with positive scalar curvature. Furthermore, any complete metric with non-negative scalar curvature on Z is Ricci-flat.*

by choosing Ω to be a small ball in M.)

---

[a] Although M itself need not be spin, the triviality of $\pi_1(\Omega) \to \pi_1(M)$ implies that Ω and $\partial\Omega$ inherit a canonical spin structure from the spin universal cover $\widetilde{M}$.

LIVEMATHEMATICIANBENCH Example 10 (Question & Answers)

Generated Question

Let M^n be a closed aspherical manifold (that is, a closed manifold whose universal cover is contractible), and let N be a spin n-manifold so that the connected sum $M\#N$ is defined. Write $\Gamma = \pi_1(M)$, and assume that the rational strong Novikov conjecture holds for Γ, i.e. the assembly map $\mu\colon K_*(M) \otimes \mathbb{Q} \to K_*(C_r^*(\Gamma)) \otimes \mathbb{Q}$ is injective. Which is the strongest statement about complete Riemannian metrics on $M\#N$ that is valid under these hypotheses?

Answer Choices

A. The connected sum $M\#N$ admits no complete metric of positive scalar curvature. Moreover, if $M\#N$ carries a complete metric of non-negative scalar curvature, then that metric must be Ricci-flat. ✓

B. The connected sum $M\#N$ admits no complete metric of positive scalar curvature. Moreover, every complete metric on $M\#N$ with non-negative scalar curvature is in fact flat.

C. Every complete metric on $M\#N$ with non-negative scalar curvature has identically zero scalar curvature; in particular, $M\#N$ admits no complete metric of positive scalar curvature.

D. The connected sum $M\#N$ admits no complete metric of non-negative scalar curvature at all; equivalently, $M\#N$ carries neither a complete Ricci-flat metric nor a complete metric of positive scalar curvature.

E. If $M\#N$ carries one complete metric of non-negative scalar curvature, then every complete metric on $M\#N$ must also have non-negative scalar curvature and be Ricci-flat.

11. Uniqueness. *Canonical form:* $\exists!\, x$ such that $P(x)$. This category covers theorems asserting both existence and uniqueness of an object satisfying a given property. *Example:* "For every positive definite matrix A, there exists a unique positive definite matrix B such that $B^2 = A$."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Mu & Wang (2025))

Let $f_0 \in C_0^{1,\mu}(\bar{\Omega} \times \mathbb{R}^2)$, $f_0 \geq 0$ for some $0 < \mu < 1$, satisfying

$$f_0 \in C^{1,\mu}, \quad f_0(x,v) = \text{constant}, \ \text{dist}((x,v),\Gamma) \leq \delta_0, \tag{1}$$

Suppose that $h \in C^{2,\mu}(\partial\Omega)$ satisfies

$$\int_\Omega f_0(x,v)\mathrm{d}x\mathrm{d}v = \int_{\partial\Omega} h(x)\mathrm{d}l, \tag{2}$$

and $h > 0$. Then there exists a unique solution $f \in C^{1;1,\lambda}_{t;(x,v)}\big((0,\infty) \times \bar{\Omega} \times \mathbb{R}^2\big)$, $\varphi \in C^{1;3,\lambda}_{t;x}\big((0,\infty) \times \bar{\Omega}\big)$ for some $0 < \lambda < \mu$, of the relativistic Vlasov-Poisson system

$$\partial_t f + \hat{v} \cdot \nabla_x f + \nabla_x \varphi \cdot \nabla_v f = 0, \tag{3}$$

$$\Delta\varphi = \rho, \tag{4}$$

with compact support in x and v, where the initial boundary conditions of (f, φ) satisfy

$$f(0,x,v) = f_0(x,v), \quad x \in \Omega, \ v \in \mathbb{R}^2, \tag{5}$$

$$f(t,x,v) = f(t,x,v^*), \quad x \in \partial\Omega, \ v \in \mathbb{R}^2, \ t > 0, \tag{6}$$

$$f_0(x,v) \geq 0, \tag{7}$$

$$v^* = v - 2(n_x \cdot v)n_x, \quad (x,v) \in \partial\Omega \times \mathbb{R}^2, \tag{8}$$

and

$$\frac{\partial\varphi}{\partial n_x} = h(x), \ \ x \in \partial\Omega, \ t > 0, \tag{9}$$

respectively.

Proof Sketch

To prove the main results (in particular the main theorem), the introduction describes the following strategy.

1. **Linearized well-posedness via a velocity lemma:** "we first apply the velocity lemma to establish the well-posedness of linearized problems."
2. **Iterative construction and convergence:** "Then, we construct an iterative scheme and show the convergence of the iterative sequences." The "main issues" in this step are "the uniform boundedness in a given function space and the prolongation of uniform estimates for the functions f^n."
3. **Bootstrapping to final regularity/conclusion:** "Finally, we use bootstrapping techniques to reach the desired conclusions."
4. **Control of boundary interactions / collisions (via adapted coordinates):** Near the singular set and boundary, the plan is to "draw on insights from previous research," assuming "the initial data f_0 remains constant near the singular set," and "to analyze the characteristic curves close to the boundary, we will use geometric methods." For the (Neumann) boundary analysis, they "select coordinate variables (α, β)," where "α characterizes the distance from points on the characteristic curve to the singular set" and "β ... describes the number of collisions between particles and the boundary," with the observation that "the number of collisions is inversely proportional to the distance to the singular set." For this purpose (to ensure the collision number is uniformly bounded), they assume

$$f_0 \in C^{1,\mu}, \quad f_0(x,v) = \text{constant}, \ \text{dist}((x,v),\Gamma) \leq \delta_0,$$

5. **Paper roadmap matching the proof:** Sections 2–5 implement this plan: "introduce an iterative system ... define a sequence of functions $\{f^n\}$, the limit ... is the desired global solution," then "establish the well-posedness of the linear problem," then "show ... convergence ... under the condition that $Q(t)$ is bounded," and finally "prove the boundedness of the function $Q(t)$. This concludes the proof of the first theorem."

LIVEMATHEMATICIANBENCH Example 11 (Question & Answers)

Generated Question

Let $0 < \mu < 1$, let $\Omega \subset \mathbb{R}^2$ have smooth boundary $\partial\Omega$, and let n_x denote the outward unit normal at $x \in \partial\Omega$. Write $\Gamma := \{(x, v) \in \partial\Omega \times \mathbb{R}^2 : v \in T_x\partial\Omega\}$, let $C_0^{1,\mu}(\bar{\Omega} \times \mathbb{R}^2)$ denote $C^{1,\mu}$ functions with compact support, and let $\mathrm{d}l$ denote arc-length measure on $\partial\Omega$. Suppose $f_0 \in C_0^{1,\mu}(\bar{\Omega} \times \mathbb{R}^2)$ is nonnegative and satisfies $f_0(x, v) =$ constant whenever $\mathrm{dist}((x, v), \Gamma) \leq \delta_0$, and let $h \in C^{2,\mu}(\partial\Omega)$ satisfy $h > 0$ and $\int_\Omega f_0(x, v)\,\mathrm{d}x\,\mathrm{d}v = \int_{\partial\Omega} h(x)\,\mathrm{d}l$. For the relativistic Vlasov-Poisson system

$$\partial_t f + \hat{v} \cdot \nabla_x f + \nabla_x \varphi \cdot \nabla_v f = 0, \qquad \Delta\varphi = \rho, \qquad \hat{v} = \frac{v}{\sqrt{1 + |v|^2}}, \quad \rho(t, x) = \int_{\mathbb{R}^2} f(t, x, v)\,\mathrm{d}v,$$

with initial condition $f(0, x, v) = f_0(x, v)$, specular reflection boundary condition $f(t, x, v) = f(t, x, v^*)$ on $\partial\Omega \times \mathbb{R}^2 \times (0, \infty)$, where $v^* = v - 2(n_x \cdot v)n_x$, and Neumann boundary condition $\partial\varphi/\partial n_x = h(x)$ on $\partial\Omega \times (0, \infty)$, what is the strongest statement about existence and uniqueness that follows?

Answer Choices

A. There exists a unique pair (f, φ) with $f \in C^{1;1,\lambda}_{t;(x,v)}((0, \infty) \times \bar{\Omega} \times \mathbb{R}^2)$ and $\varphi \in C^{1;3,\lambda}_{t;x}((0, \infty) \times \bar{\Omega})$ for some $0 < \lambda < \mu$, solving the above relativistic Vlasov-Poisson system, satisfying the stated initial, specular-reflection, and Neumann boundary conditions, and having compact support in x and v. ✓

B. There exists a unique pair (f, φ) with $f \in C^{1;1,\lambda}_{t;(x,v)}((0, \infty) \times \bar{\Omega} \times \mathbb{R}^2)$ and $\varphi \in C^{1;3,\lambda}_{t;x}((0, \infty) \times \bar{\Omega})$ for some $0 < \lambda < \mu$, solving the above relativistic Vlasov-Poisson system, satisfying the stated initial, specular-reflection, and Neumann boundary conditions, and having compact support in x and v; moreover one may take $\lambda = \mu$.

C. There exists at least one pair (f, φ) with $f \in C^{1;1,\lambda}_{t;(x,v)}((0, \infty) \times \bar{\Omega} \times \mathbb{R}^2)$ and $\varphi \in C^{1;3,\lambda}_{t;x}((0, \infty) \times \bar{\Omega})$ for some $0 < \lambda < \mu$, solving the above relativistic Vlasov-Poisson system, satisfying the stated initial, specular-reflection, and Neumann boundary conditions, and having compact support in x and v.

D. There exists a unique pair (f, φ) with $f \in C^{1;1,\lambda}_{t;(x,v)}((0, \infty) \times \bar{\Omega} \times \mathbb{R}^2)$ and $\varphi \in C^{1;3,\lambda}_{t;x}((0, \infty) \times \bar{\Omega})$ for some $0 < \lambda < \mu$, solving the above relativistic Vlasov-Poisson system, satisfying the stated initial, specular-reflection, and Neumann boundary conditions, and having compact support in x and v, even without the flatness assumption $f_0(x, v) =$ constant whenever $\mathrm{dist}((x, v), \Gamma) \leq \delta_0$.

E. For every $0 < \lambda < \mu$, there exists a unique pair (f, φ) with $f \in C^{1;1,\lambda}_{t;(x,v)}((0, \infty) \times \bar{\Omega} \times \mathbb{R}^2)$ and $\varphi \in C^{1;3,\lambda}_{t;x}((0, \infty) \times \bar{\Omega})$, solving the above relativistic Vlasov-Poisson system, satisfying the stated initial, specular-reflection, and Neumann boundary conditions, and having compact support in x and v.

12. Universal. *Canonical form:* $\forall\, x,\ P(x)$. This category covers theorems asserting that a property holds for *every* object in a specified class. *Example:* "Every finite-dimensional real vector space admits an inner product."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Grabbel et al. (2026))

Let L be a hyperbolic lattice of rank at least 46. Then, the exceptional lattice of L is trivial.

Proof Sketch

To prove the main theorem, we start from Nikulin's explicit description of the exceptional lattice (Nikulin, "Elliptic fibrations on K3 surfaces", Theorem 4.2), see also the following proposition.

Proposition . *Let L be a lattice such that $\mathcal{E}_\infty(L)$ is non-empty. Then the exceptional lattice satisfies*

$$E(L) = \bigcap_{e \in \mathcal{E}_\infty(L)} (e^\perp)^{(2)}_{\mathrm{pr}}.$$

It implies that $E(L)$ is trivial "as soon as every simple (-2)-root $r \in L$ admits a cusp with infinite stabilizer $e \in \overline{\mathcal{D}}_L$ such that $e.r > 0$."
For a hyperbolic lattice L of rank at least 46, one "tak[es] an arbitrary maximal overlattice of L" to reduce to lattices of the form

$$L = U \oplus E_8^n \oplus M, \quad n \geq 5,\ \mathrm{rk}(M) \leq 11$$

by the following proposition.

Proposition . *If L is a hyperbolic lattice of rank at least 6, then there exist an overlattice L' of L, an integer n and a negative-definite lattice M such that*

$$L' \cong U \oplus E_8^n \oplus M,$$

with $\mathrm{rk}\, M \leq 11$ *and such that the root part of M does not contain any E_8 or an ADE lattice of rank at least 9.*
Using the isometry $U \oplus E_8^3 \cong U \oplus \Lambda$ and that the Leech lattice Λ has no (-2)-root, one deduces that "the exceptional lattice $E(L)$ is contained in M." Since there are "only finitely many possibilities for the root part of M," the argument given later checks all cases and shows "that $E(L) = 0$ for all of them," which "conclud[es] the proof of the main theorem."

LIVEMATHEMATICIANBENCH Example 12 (Question & Answers)

Generated Question

Here, a hyperbolic lattice means an even integral lattice of signature $(1, \mathrm{rk}\, L - 1)$, and the exceptional lattice of L is the sublattice of vectors in L with finite orbit under $\mathrm{O}^+(L)/W(L)$, where $\mathrm{O}^+(L)$ is the group of positive isometries of L and $W(L)$ is the subgroup generated by reflections in (-2)-roots. Which is the strongest statement that holds for every hyperbolic lattice L of rank at least 46?

Answer Choices

A. One of the remaining options is correct, but a stronger result can be proven. ✓
B. For every hyperbolic lattice L of rank at least 46, the lattice L itself is isometric to $U \oplus E_8^n \oplus M$ for some $n \geq 5$ and some negative-definite lattice M of rank at most 11 whose root part contains no E_8 or ADE summand of rank at least 9.
C. The exceptional lattice of every hyperbolic lattice L of rank at least 46 has rank at most 11.
D. For every hyperbolic lattice L of rank at least 46, if L' is a maximal overlattice of L with $L' \cong U \oplus E_8^n \oplus M$ and $n \geq 5$, then the exceptional lattice of L is exactly the root sublattice of M.
E. For every hyperbolic lattice L of rank at least 46, after passing to a maximal overlattice $L' \cong U \oplus E_8^n \oplus M$ with $n \geq 5$, the isometry $U \oplus E_8^3 \cong U \oplus \Lambda$ with the Leech lattice Λ already forces the exceptional lattice of L to be trivial, without any further case-by-case analysis of M.

13. Universal–Existential. *Canonical form:* $\forall\, x\ \exists\, y,\ P(x, y)$. This category covers theorems asserting that for every object in a class, there exists an associated object satisfying a property. The existential witness may depend on the universally quantified variable. *Example:* "For every $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ whenever $|x - a| < \delta$."

LIVEMATHEMATICIANBENCH Example (Theorem & Sketch)

Original Theorem (Mendonça et al. (2025))

Theorem (Nešetřil & Rödl, 1976)**.** For every $k \geq 2$ there is a graph G such that $K_k \nsubseteq G$ and $G \to K_{k-1}$.

Proof Sketch

The post-theorem introduction gives the following proof outline for the main theorem: it "combines probabilistic methods with the hypergraph container framework Balogh–Morris–Samotij, *Independent sets in hypergraphs* (2015) and Saxton–Thomason, *Hypergraph containers* (2015) and is inspired by ideas from Bollobás–Erdős, *Ramsey graphs* (2001)." The paper is then organized into steps: (1) Next, "with high probability ... the graph G obtained in a natural way from every 'dense' subhypergraph of a suitable n-vertex random s-uniform hypergraph satisfies $G \to (K_s, K_{k-1})$ for $s = R(k) - 1$." (2) Then, "with high probability a suitable random hypergraph $\mathcal{H}$ contains a dense subhypergraph $\mathcal{H}_0$ that will allow us to obtain a graph G such that $G \not\to K_k$." (3) "These results are then combined later" to produce a graph G satisfying both $G \not\to K_k$ and $G \to (K_s, K_{k-1})$ (with $s = R(k) - 1$).

LIVEMATHEMATICIANBENCH Example 13 (Question & Answers)

Generated Question

Let K_t denote the complete graph on t vertices. For graphs G and H, write $G \to H$ to mean that every red-blue coloring of the edges of G contains a monochromatic copy of H. Which is the strongest statement that holds for every integer $k \geq 2$?

Answer Choices

A. One of the remaining options is correct, but a stronger result can be proven. ✓
B. For every integer $k \geq 2$ there is a graph G such that $K_k \nsubseteq G$ and $G \to (K_k, K_{k-1})$.
C. For every integer $k \geq 2$ there is a graph G such that $G \to K_{k-1}$.
D. For every integer $k \geq 2$ there is a graph G such that $K_k \nsubseteq G$ and $G \to (K_{R(k)-1}, K_{k-1})$, where $R(k)$ denotes the Ramsey number of K_k.
E. For every integer $k \geq 2$ there is a graph G such that $G \nrightarrow K_k$ and $G \to (K_{R(k)-1}, K_k)$, where $R(k)$ denotes the Ramsey number of K_k.

D Benchmark Composition Across Construction Stages

Figure 6 summarizes how the benchmark size changes across the main construction stages for each month. **Full** denotes the complete post-generation candidate pool, **Quality Rubric**$>$ 5 denotes the subset retained after the rubric-based quality filter, and **Hard** denotes the final released split after stem-nontrivial filtering and hardness calibration. The figure shows that the pipeline consistently reduces the candidate pool in a structured way rather than through ad hoc pruning: each month begins from a substantially larger generated set, is narrowed by quality control, and is then further distilled into a compact hard benchmark.

E Validation of Distractor Generation

To directly evaluate the design choices used in distractor generation, we conduct a fixed-theorem ablation over 98 benchmark items.

Across all conditions, we hold the source theorem, question stem, and correct option fixed, and vary only three components of false-option generation:

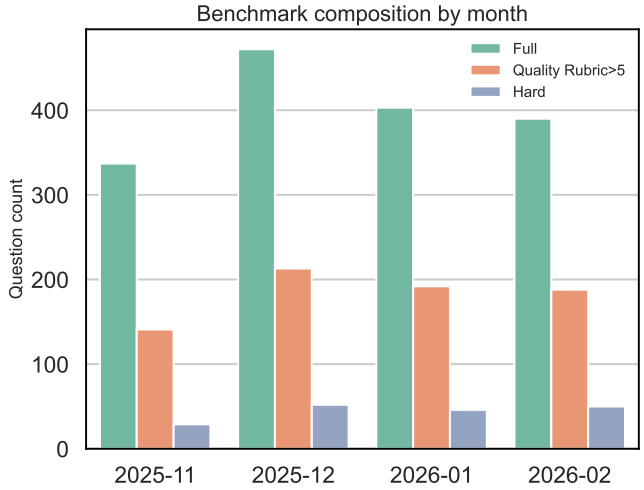

Figure 6. Benchmark composition by month across major construction stages. **Full** is the complete generated candidate pool, **Quality Rubric**> 5 is the subset passing the rubric threshold, and **Hard** is the final released split after the remaining filtering and hardness-selection steps.

Setting	Sketch	Category	Revision	Acc.	Hardness	n
Minimal baseline	No	No	No	72.4%	27.6%	98
w/o sketch	No	Yes	Yes	75.5%	24.5%	98
w/o category guidance	Yes	No	Yes	63.3%	36.7%	98
w/o revision	Yes	Yes	No	62.2%	37.8%	98
Full pipeline	Yes	Yes	Yes	**61.2%**	**38.8%**	98

Table 2. Fixed-theorem ablation of distractor generation. Sketch denotes access to the proof sketch; Category denotes category-aware distractor guidance; and Revision denotes the adversarial revision pass. Hardness is defined as one minus solver accuracy.

(i) access to the proof sketch, (ii) category-aware distractor guidance, and (iii) the adversarial revision pass.

We report solver accuracy, where lower accuracy indicates more challenging distractors.

The full pipeline produces the most challenging distractors, reducing solver accuracy from 72.4% under the minimal baseline to 61.2%, an 11.2 percentage-point decrease.

Removing access to the proof sketch increases accuracy to 75.5%, indicating that sketch access makes the largest contribution to distractor difficulty.

Removing category-aware guidance increases accuracy by 2.1 percentage points, while removing the adversarial revision pass increases accuracy by 1.0 percentage point.

These results support the use of proof sketches and category-aware guidance for constructing more challenging distractors, while the revision pass provides a smaller additional contribution.

F Evaluation Implementation Details

This section documents the released evaluation implementation in `LiveMathematicianBench/eval/`. The current repository provides three backend-specific entry points: `eval.py` for Azure OpenAI models, `eval_claude.py` for Claude models accessed through the Anthropic API, and `eval_vllm.py` for models served through an OpenAI-compatible vLLM endpoint. Despite backend-specific client code and usage accounting, the three scripts share the same hard-set loader, prompt template, deterministic choice shuffling, answer parser, correctness rule, and output JSON structure.

Benchmark file format. For each evaluation month, the scripts load the hard benchmark split from `data/<month>/hard/qaEval_<month>_ge5_hard.json`. Each JSON item contains a unique identifier and an `mcq` object with a question stem, one `correct_choice`, four distractors in `choices`, and auxiliary metadata such as `mcq.meta.score`. During evaluation, the released scripts consume the finalized `mcq` fields only; source theorem context, sketches, and other record-level annotations remain in the benchmark file but are not inserted into the released prompts.

Prompt construction. Evaluation is implemented as a five-way multiple-choice selection task. For each item, the scripts combine the stored correct choice with the four distractors, then apply a deterministic shuffle using a per-item seed equal to the global seed plus the item's dataset index. The shuffled options are relabeled as `A`–`E`. The model then receives a fixed system prompt instructing it to act as an expert mathematician, reason step by step, and place its final answer inside `\boxed{}`. The user prompt contains only two blocks:

1. the question stem from `mcq.question`; and
2. the five relabeled answer choices.

The choices are formatted as "`(A) ...`", "`(B) ...`", and so on, separated by blank lines. All models were evaluated in the without-sketch setting. Due to budget constraints, we additionally ran sketch-aware evaluation only for the few models that performed best in the without-sketch setting.

Answer extraction and scoring. After generation, the scripts recover the model's predicted label by first searching for the last occurrence of `\boxed{A}` through `\boxed{E}`. If no boxed answer is found, the parser falls back to either a one-character response in `A`–`E` or, failing that, the last standalone capital letter in `A`–`E`. In the Azure and vLLM evaluators, answer extraction is applied directly to the returned response text. In the Claude evaluator, the Anthropic response is split into text and thinking blocks; the final prediction is extracted from the text block, while the thinking block is recorded separately. An item is marked correct iff the extracted label matches the post-shuffle label of the ground-truth option. Monthly accuracy is then computed as `correct/total` over the evaluated hard split.

Execution details. All evaluators parallelize item-level inference with a `ThreadPoolExecutor` controlled by `--concurrency`. They also support repeated sampling through `--n`; in that case each question is answered multiple times, a full per-sample list is stored under `samples`, and the top-level prediction fields intentionally mirror the first sample for backward compatibility. For every sample, the scripts record elapsed time, prompt tokens, completion tokens, total tokens, and reasoning-token counts whenever the backend exposes them.

The backend-specific command-line options differ as follows. The Azure evaluator requires `--endpoint` and `--api-key`, accepts `--api-version`, exposes both a client-wide `--timeout` and a per-sample `--request-timeout`, and optionally switches from chat completions to the Responses API through `--use-responses-api`; this switch is also enabled automatically when the model name contains `gpt-5.4`. The Claude evaluator requires `--base-url`, accepts `--api-key`, supports `--thinking-budget` to request extended thinking, and exposes `--debug`, `--timeout`, and `--request-timeout`. The vLLM evaluator requires `--base-url`, accepts an optional `--api-key`, exposes `--temperature`, `--top-p`, `--debug`, `--timeout`, and `--request-timeout`, and also supports optional `--use-responses-api` with automatic activation for model names containing `gpt-5.4`.

Token accounting is normalized across backends. Azure records usage directly from the provider response. Claude and vLLM additionally normalize `completion_tokens` so that, when reasoning-token metadata is available, the reported completion count includes both visible output tokens and reasoning tokens, matching the OpenAI-style accounting used elsewhere in the benchmark.

Resuming and result serialization. All scripts support fault-tolerant resumption through `--resume`. If an output file already exists, the evaluator loads the pre-

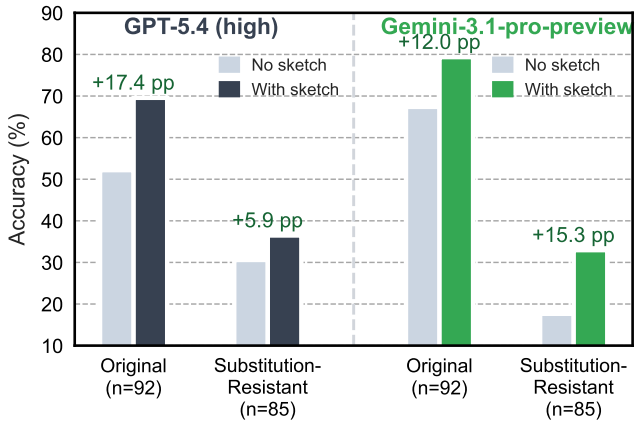

Figure 7. Sketch-aware accuracy gains broken down by choice style. For each model, we compare accuracy without sketch access and with sketch access on the original-choice subset and the substitution-resistant subset. Proof sketches improve accuracy in both regimes, but the relative gain differs across models and is particularly pronounced on the harder substitution-resistant items for Gemini-3.1-pro-preview.

vious JSON, skips items whose prior record has a non-None `model_answer` and no recorded error, and re-runs unanswered or failed items only. Results are written to `results/<month>/accuracy_test_<model>_<month>_<effort>.json`, where model names are sanitized for filesystem safety and `<effort>` defaults to `default` when no reasoning level is specified. Each output file contains (i) `test_info` metadata such as model name, month, seed, maximum token budget, number of samples, and timestamps; (ii) duplicated aggregate accuracy summaries under `summary.all` and `overall`; and (iii) `detailed_results` records containing the item identifier, extracted `model_answer`, shuffled `correct_answer`, `raw_response`, correctness flag, latency, token usage, `reasoning_tokens`, reasoning effort, `n_samples`, optional per-sample `samples`, auxiliary `mcq.meta.score`, and any backend error message. For Claude runs, each record additionally stores `raw_thinking` extracted from the Anthropic thinking block.

G Additional Evaluation Results

G.1 Sketch-Aware Accuracy with Substitution-Resistant Settings

Built upon the sketch-aware evaluation results in Figure 5(b), we further analyze how the presence of proof sketches interacts with and without substitution-resistant question formulations. We break down sketch-aware gains, separating standard multiple-choice questions from substitution-resistant items. Figure 7 shows that proof sketches improve performance in both settings, but the magnitude of the gain depends on the model and the choice style. GPT-5.4 benefits more strongly on the original-choice subset, whereas Gemini-3.1-pro-preview shows especially large gains on the substitution-resistant subset. This suggests that proof-level guidance can help models not only identify the correct theorem statement, but also reason through adversarial answer formulations designed to resist shortcut strategies.

G.2 Cost Analysis

Accuracy alone does not capture how practical a model is for benchmark-scale evaluation. Figure 8 shows that the accuracy-cost frontier is led by GPT-5.4 rather than by the highest-token models. Gemini-3.1-pro-preview achieves the top accuracy at 43.5%, but it requires about 15.3k average completion tokens. By contrast, GPT-5.4 (high) reaches a similar 41.8% accuracy with only about 7.0k tokens, and GPT-5.4 (medium) remains close at 41.2% with about 3.8k tokens. In other words, GPT-5.4 combines high accuracy with materially lower

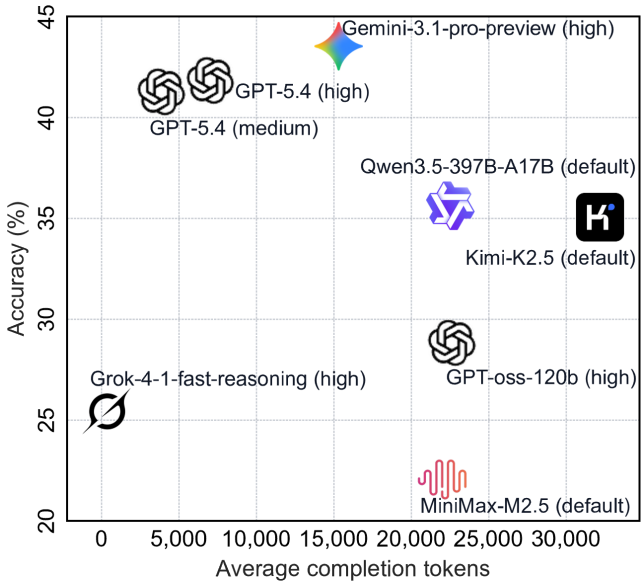

Figure 8. Accuracy versus average completion tokens on LIVEMATHEMATICIANBENCH. Each point denotes one model setting. The most favorable region is the upper-left, which corresponds to higher accuracy with fewer generated tokens. GPT-5.4 lies on a particularly strong frontier: it remains close to Gemini-3.1-pro-preview in accuracy while using substantially fewer completion tokens.

token cost, while several other long-generation models, such as Qwen and Kimi, use far more tokens without matching GPT-5.4 or Gemini in accuracy.

At the same time, several open-weight or alternative frontier systems occupy distinct efficiency niches. Qwen3.5-397B-A17B and Kimi-K2.5 achieve mid-tier accuracy at higher token budgets, while GPT-oss-120b generates even more tokens without matching the top closed-model accuracy band. This pattern suggests that improvements in research-level mathematical reasoning cannot be reduced to simply allocating more inference budget: architectural and training differences still materially affect how efficiently models convert tokens into correct mathematical judgments.

H Prompt Details

We provide below some examples of system prompts used in our construction pipeline. We will focus on prompts used in Stage 4 (Question-Answer Pair Generation) as it best demonstrates the methodological novelty of LIVEMATHEMATICIANBENCH.

To show case our *category-specific* system prompts, we include prompts from two categories: *biconditional/equivalence* and *implication* as they are representative of our central design philosophy.

Stage 4a: Question Stem and Correct Choice Generation.

System Prompt for Biconditional/Equivalence Theorems

You are an expert in mathematics. You are given one input:

- `"main_theorem"`: the full LaTeX code of a theorem-like environment.

The `"main_theorem"` is an equivalence-type theorem: statements asserting "A if and only if B" or "the following are equivalent."
Task: Generate a multiple-choice question stem and the single correct answer option that exactly and faithfully represent the full mathematical content of the theorem.

1. **Detect the equivalence structure:** Read the theorem and identify any "if and only if" or "the following are equivalent" pattern. List each distinct statement or property that appears in the equivalence. Rewrite them as short, self-contained statements such as A, B, C, etc.
2. **Decide which side(s) will appear in the question.** If there are two statements ($A \Leftrightarrow B$), design the question as "Which of the following statements is equivalent to A?". i.e. A appears in the stem and B becomes the correct answer. If there are three or more equivalent statements ($A \Leftrightarrow B \Leftrightarrow C \ldots$), focus on identifying any new side of the equivalence introduced by the theorem, and write the question as "Which of the following statements is equivalent to both A and B?".
3. **Write the question stem.** Use only the side(s) named in Step 2. Do not include the correct equivalent statement in the stem. Phrase the question clearly, using formal but concise mathematical language. Avoid referencing the theorem itself ("according to the theorem", etc.).

   Do not use phrases such as:
   - "according to the theorem/statement"
   - "the theorem shows"
   - "as proved above"
4. **Determine the correctness criterion for the correct option.** The correct option must satisfy all of the following:
   - It captures the full conclusion of the theorem, up to harmless rephrasing;
   - It includes all essential aspects of the result (structure, existence, bounds, conditional clauses, arithmetic restrictions, etc.);
   - It does not omit essential components, strengthen or weaken the statement, change quantifiers or scope, or add auxiliary conclusions not present in the theorem.
5. **Output requirements.**
   - Do NOT generate distractors.
   - Do NOT mention any proof or proof sketch.
   - Produce a JSON object with exactly the following fields:

```
{
  "question": "<question stem written from main_theorem>",
 "correct_choice": {"label": "A", "text": "<the correct option based on main_theorem>"}
}
```

System Prompt for Implication Theorems

You are an expert in mathematics. You are given one input:

- "main_theorem": the full LaTeX code of a theorem-like environment.

The "main_theorem" is an implication-type theorem, i.e. it asserts that certain assumptions imply a specific conclusion (for example: "If A holds, then B holds").
Task: Generate a multiple-choice question stem and the single correct answer option that exactly and faithfully represent the full mathematical content of the theorem.
Instruction:

1. **Detect the implication structure:**
   Read the theorem and identify:
   - the full set of assumptions (A), including parameter ranges, structural hypotheses, and problem setup;
   - the conclusion (B).

   Rewrite these internally as self-contained statements, but do not simplify or strengthen them.
2. **Write the question stem**
   The question stem must:
   - explicitly include all essential assumptions of the theorem (including equations, boundary conditions, parameter ranges, and structural hypotheses);
   - ask the solver to identify the strongest statement that can be proved under these assumptions;
   - avoid referencing the theorem itself ("according to the theorem", etc.).

   Do not use phrases such as:
   - "according to the theorem/statement"
   - "the theorem shows"
   - "as proved above"

   Acceptable stem formulations include:
   - "What is the strongest statement that can be proved about . . . ?"
   - "Which of the following is the strongest consequence of the assumptions above?"

   The stem should be general enough that weaker consequences, stronger but unjustified conclusions, and conclusions about closely related properties all appear plausible.
3. **Determine the correctness criterion for the correct option**
   The correct option must satisfy all of the following:
   - It captures the full conclusion of the theorem, up to harmless rephrasing;
   - It includes all essential aspects of the result (structure, existence, bounds, conditional clauses, arithmetic restrictions, etc.);
   - It does not omit essential components, strengthen or weaken the statement, change quantifiers or scope, or add auxiliary conclusions not present in the theorem.
4. **Output requirements**
   - Do NOT generate distractors.
   - Do NOT mention any proof or proof sketch.
   - Produce a JSON object with exactly the following fields:

```
{
  "question": "<question stem written from main_theorem>",
 "correct_choice": {"label": "A", "text": "<the correct option based on main_theorem>"}
}
```

Stage 4b: Sketch-Adversarial Distractor Generation.

System Prompt for Biconditional/Equivalence Type Theorems

You are a senior research mathematician generating hard-negative distractors. You are given four inputs:

- `"main_theorem"`: full LaTeX theorem-like environment.
- `"proof_sketch"`: proof outline for distractor design only.
- `"question_stem"`: the question stem (string).
- `"correct_choice"`: the correct option (string).

Task:
Generate exactly four choices B–E:

- B, D, E: false but highly plausible (hard negatives).
- C: weaker-but-true: $A \Rightarrow C$ and $C \not\Rightarrow A$.

Do not edit `question_stem` or `correct_choice.text`.
Mandatory internal workflow (do internally; then output metadata):

- **Read & parse everything** (including the sketch): fully read `main_theorem`, `proof_sketch`, `question_stem`, and Option A; identify central objects, hypotheses, quantifiers, case splits, and what the sketch highlights as "hard steps."
- **Extract sketch load-bearers:** from the sketch, identify 1–3 load-bearing steps and the constraints they enforce (ranges, excluded cases, dependence of constants, uniform vs non-uniform, effective vs non-effective, quantifier order, hidden hypotheses, structural restrictions).
- **Design distractors via sketch-derived constraint tampering:**
  - Each false distractor must violate a specific sketch-derived constraint and be really difficult to rule out without the sketch.
  - The weaker-true distractor C must be implied by A by dropping one critical component.
- **Keyword normalization:** spread ≥ 2 key technical terms from A across ≥ 2 of B/D/E.
- **Near-miss requirement:** every distractor must differ from A in only one or two mathematically meaningful places. Avoid obviously unrelated statements, blatant negations, or options that can be rejected by surface reading alone.
- **Anti-triviality requirement:** make the solver distinguish options by subtle issues such as quantifier order, dependence of constants, exact asymptotic regime, necessity of an extra hypothesis, sharpness of the bound, uniqueness vs mere existence, completeness of a classification, or uniformity vs pointwise validity.

False distractors: required breakdown

- Produce three false options (B, D, E) with:
  - Two template-based (choose two different): boundary/range trap; uniformity/-effectivity mirage; quantifier/dependence trap; property confusion; stronger-theorem trap.
  - One wildcard false distractor (exactly one of B/D/E): does not need to match a template, but must still be tightly coupled to a specific sketch-derived constraint.
  - record its label in `meta.wildcard_false_label`.

Surface anti-gaming constraints:

- B, D, E: match A in logical shape (quantifiers/clauses), LaTeX/technical density, and approximate length.
- C may be slightly shorter (natural weakening).
- No hedging unless A uses it. Proper LaTeX syntax.
- Do not make any distractor obviously false from a missing definition, gross type mismatch, or dramatic change of topic.
- If A contains several clauses, each false distractor should tamper with only one load-bearing clause, keeping the rest close to A.

- At least two of B/D/E should look locally compatible with the hypotheses and fail only on a subtle point that an expert would have to inspect carefully.

Output format:
Return a single JSON object, for example:

```
{
  "choices": [
    {"label": "B", "text": "<false distractor>"},
    {"label": "C", "text": "<weaker-true distractor>"},
    {"label": "D", "text": "<false distractor>"},
    {"label": "E", "text": "<false distractor>"}
  ],
  "meta": {
    "weaker_true_label": "C",
    "false_labels": ["B", "D", "E"],
    "wildcard_false_label": "<one of B/D/E>"
  },
  "sketch_usage_meta": [
    {
      "label": "B",
      "sketch_hook_type": "<one of: case_split | counting_estimate | trace_identity |
                           geometric_construction | computational_check |
                           outer_automorphism | finiteness | characteristic |
                           regularity | other>",
      "tampered_component": "<short identifier of the sketch-derived constraint
      you tampered with>",
      "template_used": "<one of: boundary_range | uniformity_effectivity |
                        quantifier_dependence | property_confusion |
                        stronger_trap | wildcard>"
    },
    {
      "label": "C",
      "sketch_hook_type": "<same enum>",
      "tampered_component": "<what critical component of A was dropped to
      make C weaker-true>",
      "template_used": "weaker_true"
    },
    {
      "label": "D",
      "sketch_hook_type": "<same enum>",
      "tampered_component": "<short identifier>",
      "template_used": "<one of: boundary_range | uniformity_effectivity |
                        quantifier_dependence | property_confusion |
                        stronger_trap | wildcard>"
    },
    {
      "label": "E",
      "sketch_hook_type": "<same enum>",
      "tampered_component": "<short identifier>",
      "template_used": "<one of: boundary_range | uniformity_effectivity |
                        quantifier_dependence | property_confusion |
                        stronger_trap | wildcard>"
    }
  ]
}
```