MATHIDEABENCH: A BENCHMARK FOR MATHEMATICAL RESEARCH IDEA GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As large language models (LLMs) continue to advance in mathematical reasoning and progress from problem solving to research discovery, rigorously evaluating their capacity for research ideation becomes vital. However, measuring the quality of mathematical research proposals remains an open challenge due to its technical and subjective nature. In this work, we introduce MATHIDEABENCH, the first benchmark designed to measure the quality of LLM-generated mathematical research proposals against a human baseline. MATHIDEABENCH consists of (1) an automated data generation pipeline which produces a refreshable evaluation dataset, and (2) an LLM-based evaluator, MATHIDEAJUDGE, which scores proposals using detailed evaluation rubrics designed by experts. Furthermore, we propose a tree-based diversity metric for proposals based on mathematical subject classification. Our experiments show that compared to ideas from human-authored publications, LLMs tend to generate proposals which are less sound, less feasible and less diverse but conceptually deeper and more original.

1 INTRODUCTION

Over the past few years, frontier large language models (LLMs) have become so powerful that they are increasingly adopted to accelerate research discovery in many disciplines (Zheng et al., 2025). The field of mathematics is no exception (Yang et al., 2024). However, as LLMs continue to improve in mathematical capabilities, concerns over misalignment between the outcome of AI usage and the goal of the mathematical community are also on the rise. This prompted 25 leading mathematicians to release a joint declaration on the threat of AI to the future of mathematics (Tao, 2026), raising concerns that the overabundance of isolated AI-generated methods could destroy the process of discovery and education. While AI tools might allow us to arrive at a final answer faster, over-reliance on them deprives us of the chances to develop holistic understanding and the ability to formulate new questions and ideas. This trend can be particularly detrimental to students and early-career researchers who may be tempted to prioritize publication throughout over developing the skills necessary for long-term research advances and inter-generational knowledge transmission.

To reap the benefits of AI without suffering its potential downsides, mathematicians must not "lose sight of what their work was meant to achieve in the first place" (Tao, 2026). Most existing effort to apply AI in mathematics concentrates on the *problem-solving* aspect (Yang et al., 2024; Dong & Ma, 2025; Tsoukalas et al., 2026). In comparison, LLM-assisted *research ideation* in mathematics has received less attention, even though it is the crucial step which decides the future directions of the field. As the knowledge base continues to grow rapidly in both breadth and depth, it is becoming increasingly challenging for mathematicians to stay up to date with the literature and formulate new research directions independently. We argue that *as LLMs improve in their ability to solve problems and generate proofs, mathematicians must also effectively utilize LLMs to come up with novel and high-quality research ideas to further the field of mathematics in an informed and controlled manner*.

Just as reasoning benchmarks propelled LLMs' problem solving capabilities in mathematics (Zhang et al., 2026; Glazer et al., 2024; Dekoninck et al., 2026; Peyronnet et al., 2026), developing effective AI tools for research ideation requires dedicated benchmarks to evaluate math research proposals. While similar benchmarks exist for other fields (Guo et al., 2025; Chen et al., 2026a; Ruan et al., 2026), they cannot be reliably adapted to mathematics given the discipline's unique standard of

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

rigor and precision. To close this gap, we present **MATHIDEABENCH**, the first benchmark for mathematical research ideation to the best of our knowledge.

Designed to mirror the mathematical research workflow, the benchmark comprises two key components: (1) an **automated data generation pipeline** that produces research proposals, and (2) **MATHIDEAJUDGE**, an evaluation framework featuring detailed grading rubrics. The data pipeline identifies *source papers* on arXiv from which proposals are generated. To be eligible, the source paper must belong to a mathematical field and possess a sufficient number of qualified *follow-up papers*. The core ideas from these follow-up papers are evaluated alongside the LLM-generated proposals by MATHIDEAJUDGE, thereby establishing a human baseline for comparison. To assess qualities that influence downstream research success, MATHIDEAJUDGE employs a set of rubrics crafted by a mathematician and iteratively refined in collaboration with three researchers with strong mathematical backgrounds.

In particular, we identify three core dimensions which determines the proposal quality: **soundness**, **novelty**, and **feasibility**. For each dimension, we specify fine-grained criteria that the judge must consider when scoring. For feasibility, the judge checks edge cases, potential counterexamples and technical barriers that mathematicians typically go through to stress-test a new idea. For novelty, it evaluates both conceptual depth and originality relative to existing literature. For soundness, it assesses whether the proposal is rigorous, self-contained, and backed by proper supporting evidence. By assessing these fine-grained criteria, MATHIDEAJUDGE provides targeted feedback across each dimension that enables systematic refinement of proposals. Furthermore, we propose **topical diversity** as a novel metric for measuring the diversity of mathematical proposals using the hierarchical structure of Mathematics Subject Classification 2020 (MSC2020) codes[1]. Unlike embedding-based similarity metrics, our approach offers an interpretable metric tailored to mathematical research.

The benefits of MATHIDEABENCH are three-fold. First, MATHIDEABENCH as a benchmark can provide a rigorous testbed to measure and drive progress in automated mathematical ideation. Second, MATHIDEAJUDGE can be integrated into agentic systems as an effective scoring module to guide search and increase generation efficiency. Last but not the least, MATHIDEAJUDGE may be of independent interest for mathematicians to automate the evaluation and refinement process of their own research ideas.

We summarize our key contributions as follows:

1. **MATHIDEABENCH:** We present the first benchmark dedicated to measuring the quality of mathematical research proposals. The benchmark dataset contains 478 papers spanning diverse mathematical subfields, along with 1,330 distinct follow-up papers for baseline establishment. With an automated data generation pipeline, the benchmark dataset can be continually refreshed with minimal overhead to include future publications while avoiding data contamination.
2. **Formalization of Proposal Evaluation in Mathematics:** We identify and formalize the non-trivial workflow practiced by mathematicians to stress-test new research ideas, a process that remains largely opaque to non-domain experts and obscured by polished final publications. We operationalize this workflow as MATHIDEAJUDGE through fine-grained evaluation rubrics for soundness, feasibility and novelty.
3. **Topical Diversity**: We propose *topical diversity* as a novel and interpretable diversity metric for mathematical research ideas using MSC2020 subject codes.
4. **Guiding Safe AI-Assisted Mathematical Research:** We contribute to the discourse on AI-assisted research discovery by positioning MATHIDEABENCH as a framework for responsible idea exploration. Beyond automated evaluation, our detailed rubrics and targeted feedback serve to guide junior researchers in formulating high-quality research proposals.

2 RELATED WORK

Benchmarks for Research Ideation Existing benchmark efforts in research ideation focus primarily on machine learning and natural sciences. Guo et al. (2025) evaluates research proposals

[1] https://msc2020.org/

in biomedical fields generated from titles and abstracts of related work. Qiu et al. (2025) focus on AI-related research ideas generated from structured summaries of previous work. Feng et al. (2025) adopt graph-based methods to evaluate long-form concepts by breaking down complex ideas. Chen et al. (2026a) focus on step-wise evaluations of machine learning research including ideation and proposal generation. Ruan et al. (2026) measure LLMs' divergent thinking capabilities for scientific idea generation with minimal context. Liu et al. (2026) breaks down scientific discovery process into inspiration retrieval, hypothesis composition, and hypothesis ranking for evaluation.

However, no existing benchmark is designed specifically to evaluate mathematical research proposals. Mathematics differs fundamentally in its research workflow due to its deductive and rigorous nature: while machine learning and natural sciences rely on empirical evidence from experiments, mathematical proposals must be judged on logical reasoning, supporting heuristics and concrete examples, leaving little room for ambiguous analogies or underspecified conjectures. We address this gap by introducing a benchmark tailored to the unique standards of mathematical research.

Performance Quantification for Open-Ended Evaluation Automating the evaluation of research proposals is inherently challenging due to their open-ended, technical and highly subjective nature (Si et al., 2025). Amidst varying choices of evaluation criteria and their operationalization across existing benchmarks, feasibility and novelty are among the common key dimensions. Lu et al. (2024) and Guo et al. (2025) prompt LLMs to judge feasibility and novelty without explicit rubric definitions. Si et al. (2025) and Chen et al. (2026a) define grading dimensions using brief descriptions on 10-point grading scales. Similarly, Baek et al. (2025) employ 5-point grading scales across related dimensions to evaluate the problem, method, and experimental design in a proposal. In contrast, Qiu et al. (2025) define feasibility and novelty indirectly based on citation metrics and inferred semantic distances to related work.

Beyond the quality of individual proposals, assessing the diversity of proposals generated from shared seed contexts is critical given concerns that widespread LLM adoption may homogenize research directions. In particular, Chen et al. (2026b) highlight a tension between proposal quality and diversity. Existing metrics rely primarily on cosine similarities of text embeddings (Si et al., 2025; Hu et al., 2024) or the corresponding Vendi scores (Chen et al., 2026b), which lack domain-specific interpretability.

In contrast, the metrics in MATHIDEABENCH are designed to be transparent, domain-specific, and directly actionable for proposal refinement. Instead of relying on generic or rudimentary prompts, MATHIDEAJUDGE evaluates core quality dimensions using expert-informed rubrics mapped to rating scales from 0 to 4. For diversity evaluation, we base our metric on tree-based distances over the established MSC2020 taxonomy rather than black-box text embeddings.

3 BENCHMARK DESIGN

In this section, we present the details of MATHIDEABENCH, which contains a data generation pipeline and a set of rubrics for the LLM judge. The data pipeline consist of a *metadata collection* phase (Section 3.1) and a *proposal generation* phase (Section 3.2). The proposals are then passed to the MATHIDEAJUDGE for evaluation (Section 3.3).

Baseline Establishment For tasks with well-defined answers, benchmark datasets usually come with ground truth answers from which accuracy can be computed as the evaluation metric. Prior research ideation benchmarks by Guo et al. (2025) and Qiu et al. (2025) attempt to establish ground truth using a *backward-looking* process. Their pipelines first collect *target papers* whose ideas are treated as the ground truth, and then retrieve important inspiring papers for each target paper as the basis for proposal generation. The generated proposals are evaluated based on their alignment with the ground truth or their relative performance in rankings.

We improve upon previous work by establishing a comparison baseline in a *forward-looking* manner as per the usual process of research discovery. In the previous backward-looking approaches, the restriction to inspiring papers can potentially confine the scope and impact of generated ideas. They also introduce an undesirable dependency between proposal generation and evaluation, as they provide related work that would otherwise come from a model's parametric knowledge or an agent's literature review abilities, both of which contribute to the ideation capacity being measured.

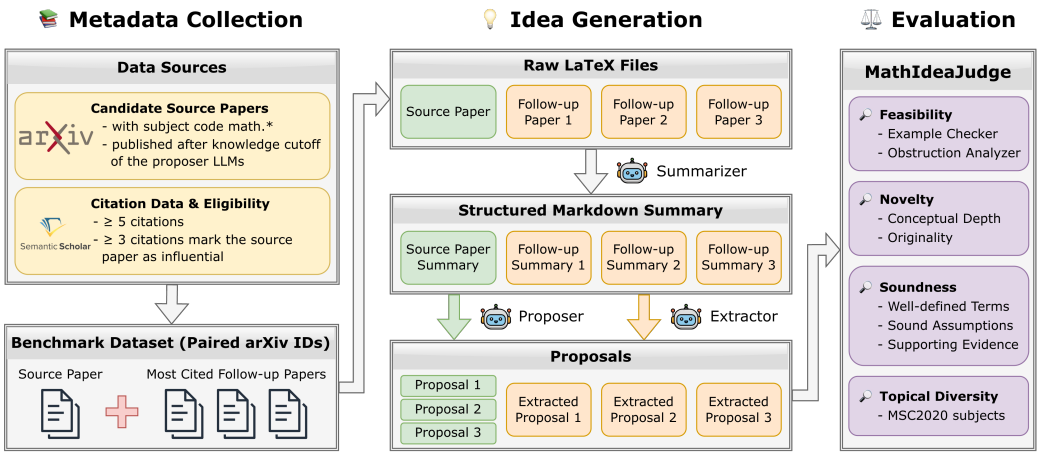


Figure 1: Overview of MATHIDEABENCH. The benchmark consists of a data generation pipeline (left, middle) and an LLM judge named MATHIDEAJUDGE (right). The data pipeline fetches papers in mathematics from arXiv and identifies important follow-up papers using Semantic Scholar to construct a human baseline. The LLMs being scored then generate proposals based on a detailed summary of the papers. The quality of those proposals are evaluated by the MATHIDEAJUDGE.

Importantly, MATHIDEABENCH treats the ideas extracted from actual publications as a comparison baseline rather than hard ground truth to be aligned with. Beyond the open-ended nature of research ideation, it is unreasonable to treat actual publications as the best possible ideas as human research is influenced by numerous external factors. As such, alignment with actual publications can only serve as a proxy for quality and does not reveal the strengths and weaknesses of a proposal directly.

3.1 PAPER METADATA COLLECTION

The initial phase of our data generation pipeline collects metadata from arXiv and Semantic Scholar (Kinney et al., 2023) to establish paired instances between source papers and their follow-up works.

First, we query the arXiv API for papers with a `math.*` arXiv category taxonomy[2] submitted between February 2025 and December 2025. The query returns 69,937 records, which yields 51,232 unique candidate source papers after removing cross-listed duplicates. Next, we filter the candidate source papers using citation data fetched from the Semantic Scholar API. A candidate source paper qualifies if its citation count is at least 5 citations and it has at least 3 eligible follow-up papers. A follow-up paper is considered eligible if (1) it has a valid arXiv identifier, (2) its submission date on arXiv is strictly later than that of the source paper, and (3) the source paper is marked as an "influential" citation in this paper by Semantic Scholar. When a given source paper admits more than three eligible follow-up papers, we keep the three most cited ones. The citation snapshots were collected in August 2026. The filtering results in 490 unique source papers. We further exclude 12 source paper with preprocessing issues.[3]

In the end, our dataset consists of 478 source papers covering 30 arXiv subcategories in mathematics. Each source paper is paired with exactly three follow-up papers in which the source has a significant impact on the citing publication according to Semantic Scholar. The 1,434 source–follow-up relationships contains 1,330 distinct follow-up papers because some follow-ups are cited by more than one source papers. The distributions of the dataset are reported in Appendix A.

Finally, the LaTeX source files of each source paper and follow-up paper are downloaded from arXiv, cleaned and consolidated into a single document for later use.

[2] We exclude `math.HO` from consideration.

[3] We exclude papers with missing TeX documents, papers with unidentifiable main TeX document, and papers that exceed context limit of our summarizer model.

3.2 RESEARCH PROPOSAL GENERATION

Paper Summarization Mathematical papers are often lengthy and dense owing to technical lemmas and detailed proofs. While these low-level details can occasionally inspire new directions, our benchmark focuses on building upon a paper's core conceptual contributions. To achieve this, we prompt a *summarizer* LLM produce a structured summary from the raw LaTeX file of each paper.

We identify the essential information required in a summary to ensure that effective proposal generation can proceed directly from the summary without referring to the original paper. Specifically, each summary comprises the following five sections: (1) Introduction & Context, (2) Precise Setup, (3) Main Results, (4) Proof Strategy, and (5) Examples & Applications. These structured summaries serve as the standardized inputs for the subsequent pipeline stages. The complete summarization prompt is provided in Appendix F.

Proposal Generation Given the summary of a source paper, the pipeline generates three research proposals using a *proposer* LLM. The proposer model is instructed to generate high-impact yet feasible proposals from the provided summary alone without external literature search tools. To ensure sufficient information for reliable proposal evaluation and downstream research, the proposer LLM is prompted to format its output as a structured Markdown document containing the following three sections: (1) Introduction & Context, (2) Precise Setup, and (3) Main Result. In particular, the "Main Result" section should contain the statement of the main conjecture along with high-level heuristics or intuitive arguments explaining why the proposed result is likely to hold.

The three proposals are generated sequentially to increase diversity. For each subsequent generation, the "Introduction & Context" sections of previously generated proposals are included in the prompt, and the proposer LLM is explicitly instructed to explore distinct directions. The prompt is engineered to ensure that each proposal remains self-contained without references to the source paper or prior proposal iterations. The complete proposer prompt is provided in Appendix G.1.

Follow-up Proposal Extraction To compare LLM-generated proposals against actual follow-up papers, we generate one proposal from each follow-up paper using an *extractor* LLM. The extractor model is prompted to identify the paper's primary contribution and rewrite it as a forward-looking research proposal with the same schema used for generated proposals. To ensure a fair comparison, the style, tone, and level of details in the extracted proposals must be calibrated to match the generated proposals. To this end, we generate proposals from two held-out papers with each proposer model and include them as two-shot examples to guide the extractor's formatting. The complete extraction prompt is provided in Appendix G.2.

Upon completing the generation phase, each proposer model yields 3 proposals generated from each source papers, alongside a corresponding set of 3 proposals extracted from the actual follow-up papers. These 6 proposals are subsequently passed to MATHIDEAJUDGE for evaluation.

3.3 MATHIDEAJUDGE

To establish a comprehensive set of evaluation criteria tailored to mathematical research ideas, we develop an LLM judge through an iterative and expert-informed process. An initial collection of evaluation dimensions was derived from existing literature (Section 2). Meanwhile, we consulted an expert mathematician to identify a set of common and widely applicable analytical steps to evaluate proposal feasibility from multiple technical angles and entry points. These frameworks were merged by two researchers with strong mathematical backgrounds to form an initial evaluation schema. We used this prototype to evaluate a small set of generated proposals, then had an independent mathematician review the rubrics and the LLM-generated scores and justifications to suggest improvements. Ultimately, this process established *soundness*, *feasibility*, and *novelty* as the three key dimensions for evaluation, alongside a detailed inspection protocol for each dimension that faithfully mirrors the evaluation workflow of mathematicians.

We emphasize that while assessments along these dimensions remain inherently subjective and rely on state-of-the-art domain knowledge, decomposing proposal quality into explicit criteria provides a more transparent and informative evaluation than prompting with high-level descriptions. These dimensions are sufficiently general to apply across diverse mathematical disciplines while maintaining

the granularity required to isolate the specific strengths and limitations of individual proposals. All scores involved in the three core dimensions are defined on a discrete rating scale from 0 to 4, with a clear description for each number grade. To ensure an unbiased assessment, MATHIDEAJUDGE evaluates each proposal without knowing whether it was generated by a proposer model or extracted from a follow-up paper. The complete evaluation prompts are provided in Appendix H.

Soundness This dimension evaluates whether a research proposal is well-formed, assessing if it is self-contained and grounded enough for a mathematician (or an AI solver) to execute. Specifically, MATHIDEAJUDGE employs the following three sub-criteria: (1) Well-Defined Terms (whether notations, terminologies and mathematical objects are clearly defined), (2) Sound Assumptions (whether the imposed mathematical conditions are valid), and (3) Supporting Evidence (whether the proposal provides plausible heuristics or high-level arguments to support the claimed conjecture). The third sub-criterion evaluates whether the proposal outlines a clear roadmap connecting its assumptions to the main conjecture through compelling intuition, analogies, structural connections, or heuristic arguments. To reflect this, the prompt design explicitly instructs the judge not to penalize the lack of a complete formal proof. Each sub-criterion is assigned a discrete score from 0 to 4, and we take their average to be the overall soundness score for the proposal.

Feasibility This dimension evaluates the potential of a research proposal to be developed into a research paper. Due to the complexity of this task, we employ two specialized sub-agents: an example-focused *Example Checker* and a theory-focused *Obstruction Analyzer*. Each sub-agent has five sub-criteria to identify barriers and their severity before consolidating its findings and assigning a score from 0 to 4. The overall feasibility is computed as the average of the two scores. Since proposals are meant to be high-level by nature and do not contain full proofs, we ask the sub-agents to flag obstructions only when they can present concrete evidence or arguments.

The Example Checker stress-tests a proposal by inspecting concrete examples which either support or contradict the conjectured result. This sub-agent walks through the following types of diagnostic examples: (1) Toy Examples (simple, explicit instances where the proposal should apply), (2) Edge Cases (boundary or degenerate conditions where behaviors might differ) (3) Counterexample Search (designed to disprove or contradict the core claims), (4) Known Instances (canonical examples or settings for which the proposal should pass), and (5) Generalization Check (examples revealing whether the claim is too weak or too strong).

The Obstruction Analyzer seeks to identify concrete mathematical barriers that might prevent a research proposal from succeeding. This sub-agent walks through the following diagnostic lenses: (1) Theoretical Barriers (known impossibility results), (2) Technical Barriers (if a specific proposed step cannot possibly deliver the claimed conclusion for a stated mathematical reason), (3) Scope barriers (dependencies on other open problems), (4) Past failures (documented failures related to this proposal), and (5) Blindspots (conditions, assumptions or other flaws in the conjecture).

Novelty MATHIDEAJUDGE distinguishes between two complementary facets of novelty: *Conceptual Depth* and *Originality*. Conceptual Depth evaluates whether a proposal presents new, meaningful insight or advancement beyond the source paper. This local metric is crucial because our ideation pipeline operates directly on a single source paper without external knowledge access. The Originality part evaluates whether the proposal introduces a genuinely novel contribution when situated within the broader literature. The contribution may take the form of a mathematical framework, an open problem, or a proof technique. To ensure a fair and time-consistent evaluation, the judge is explicitly prompted to search for related literature published prior to the upload date of the source paper. Both criterion are assigned a score from 0 to 4, whose average gives the overall novelty score.

Topical Distance and Topical Diversity To quantify subject-level similarity between proposals, we define a tree-based distance metric over the Mathematics Subject Classification 2020 (MSC2020) taxonomy. The MSC2020 subject codes follow a hierarchical tree structure[4]: the first two digits denote the top-level *section*, the third character represents the *subsection*, and the last two digits specify the granular *subject*. To avoid excessively long prompts when loading the full taxonomy, we identify relevant subject codes recursively by dynamically populating candidate options from

[4]For consistency, we restrict our analysis to standard 5-character MSC2020 codes.

the child nodes of previously selected parent categories. We refer to the resulting set of identified five-character subject codes as the *subject set* of the proposal.

For two proposals P and Q with subject sets X and Y, we define their *topical distance* $d_{\text{topic}}(P, Q)$ as the average tree distance between all pairs of subject codes from the two subject sets:

$$d_{\text{topic}}(P, Q) := \frac{1}{|X||Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y),$$

where d is the shortest-path distance on the tree of MSC2020 codes. For a collection $\mathcal{C}$ of K proposals, we define its *topical diversity* $\text{Div}_{\text{topic}}(\mathcal{C})$ as the average topical distance within this collection:

$$\text{Div}_{\text{topic}}(\mathcal{C}) := \frac{2}{K(K-1)} \sum_{P,Q \in \mathcal{C};\ P \neq Q} d_{\text{topic}}(P, Q).$$

For each source paper p, let $\mathcal{G}_p$ and $\mathcal{F}_p$ denote the collection of its generated proposes and extracted follow-up proposals respectively. We inspect LLMs' ability to generate proposals with diverse topics by comparing $\text{Div}_{\text{topic}}(\mathcal{G}_p)$ against the human baseline $\text{Div}_{\text{topic}}(\mathcal{F}_p)$.

4 Experiments & Results

In this paper, we used GLM-5.3-Flash for MathIdeaJudge and evaluated two proposer models: DeepSeek-R1 and Gemini-3.5-Flash. To avoid data leakage, both models are selected with knowledge cutoff dates no later than January 2025, before the upload date of the earliest source paper. We perform the evaluation on a subset of 400 random seeds from our dataset along with their corresponding follow-ups. For paper summarization and proposal extraction, we used GPT-5.6-Terra with reasoning effort set to high. All LLM inferences were conducted using the OpenRouter API (details in Appendix B). We include a generated proposal, an extracted follow-up proposal and their evaluations from a sample source paper in Appendix I.

4.1 Main Results

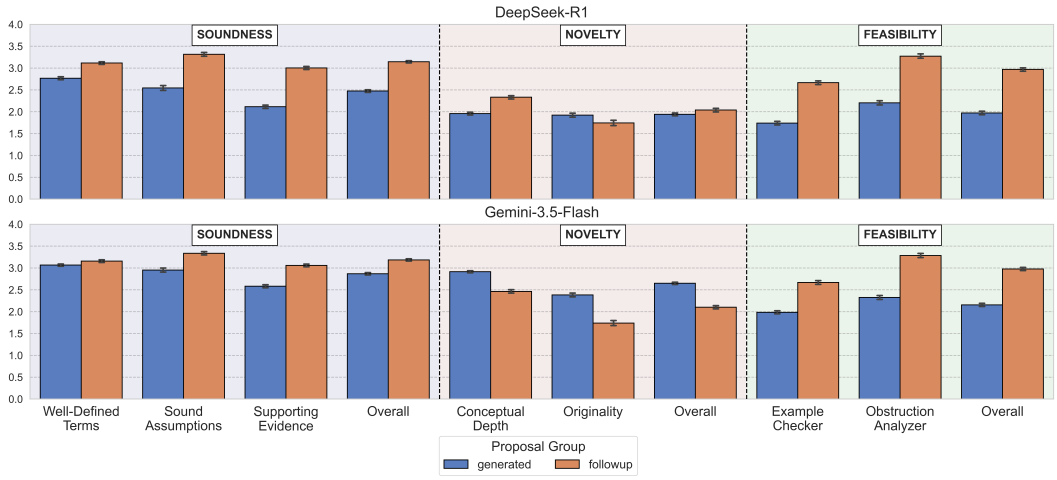


Figure 2: Average score per grading criterion for each proposer model. Each bar represents the average score in its respective grading criterion across the (generated or follow-up) 1,434 proposals, annotated with its 95% confidence interval.

The performance of the two proposer models is shown in Figure 2, with exact numerical results reported in Appendix C. From a model perspective, Gemini-3.5-Flash outperforms DeepSeek-R1 in all criteria for generated proposal. For follow-up proposals, Gemini-3.5-Flash also outperforms DeepSeek-R1 in all criteria but Originality, signaling the effect of style calibration during extraction.

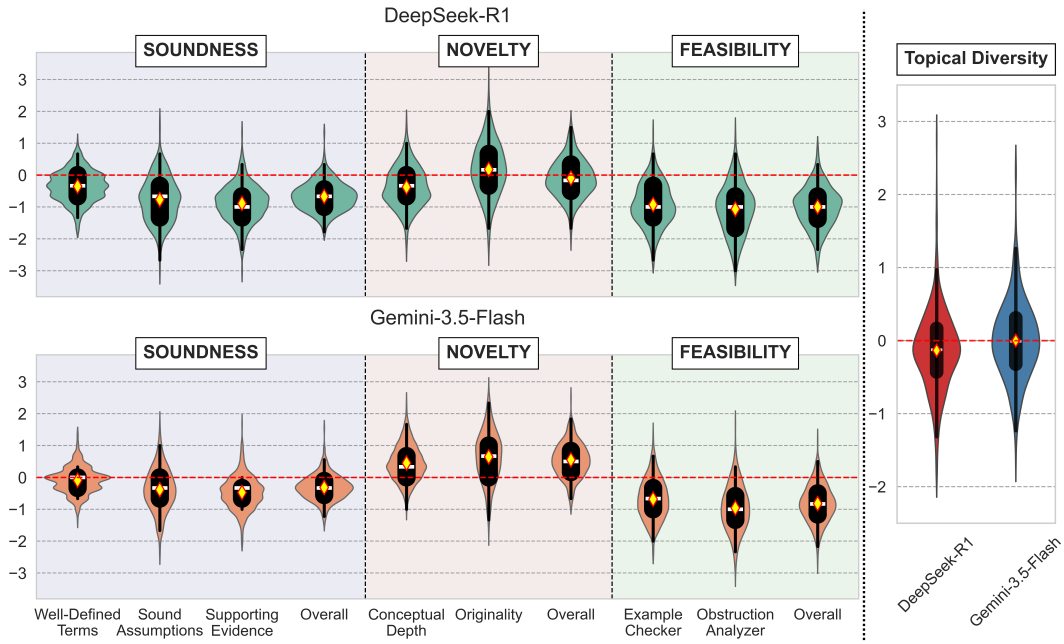


Figure 3: Violin plots for paired analysis. **(Left)** Distributions of score differences in each criterion within Soundness, Novelty and Feasibility. The diamond inside each box plot denotes the mean of the distribution. **(Right)** Distributions of $\text{Div}_{\text{topic}}(\mathcal{G}_p) - \text{Div}_{\text{topic}}(\mathcal{F}_p)$ across source papers p.

Compared to the human baseline, both DeepSeek-R1 and Gemini-3.5-Flash demonstrate lower soundness and feasibility scores in all sub-criteria. For Nnovelty, Gemini-3.5-Flash outperforms the human baseline in both Conceptual Depth and Originality, whereas DeepSeek-R1 shows lower Conceptual Depth but higher Originality. These findings are in line with prior results that LLMs struggle in feasibility and soundness while demonstrating greater potentials in novelty (Chen et al., 2026a; Guo et al., 2025; Si et al., 2025). In terms of topical diversity, Gemini-3.5-Flash achieves around the same score as the human baseline on average, while DeepSeek-R1 underperforms. It is worth noting that the selection of follow-up papers is based on citation count only and does not attempt to diversify. These results are consistent with previous findings (Si et al. (2025)) that LLM-generated research ideas exhibit limited diversity.

To account for paper-level variance, for each source paper p we compare the scores of the 3 generated proposals $\mathcal{G}_p$ and the 3 extracted follow-up proposals $\mathcal{F}_p$. For each criterion, we subtract the mean score of $\mathcal{G}_p$ from that of $\mathcal{F}_p$, so a positive difference indicates that the proposer outperforms the human baseline on p. Visualized in Figure 3, these results confirm that *LLM-generated research ideas exhibit higher novelty but demonstrate lower soundness, feasibility, and diversity*. The statistical significance of these findings are established by a a sign test on each criterion (Appendix C).

4.2 Human Study

To measure alignment between the MathIdeaJudge run by GLM-5.3-Flash and human reviewers, we conducted a human evaluation of 24 proposals associated with four source papers. Each paper contributed three LLM-generated proposals and three proposals extracted from follow-up papers. Seven reviewers with mathematical research experience collectively evaluated the proposals, with each paper receiving two or three independent reviews. Reviewers used the same seven criteria and 0–4 scoring rubrics as the LLM judge. Proposal order was randomized, and reviewers were blinded to the proposal origins. Further details are provided in Appendix D.

For each proposal and criterion, we computed human–human disagreement as the mean pairwise absolute difference among human ratings and human–LLM disagreement as the mean absolute dif-

ference between the GLM score and each human rating. We compared the resulting 24 paired values using two-sided Wilcoxon signed-rank tests with Holm correction across the seven criteria.

As shown in Figure 4, human–GLM disagreement was aligned with human–human disagreement for five out of the seven criteria. The disagreement was statistically significantly lower for Well-Defined Terms (0.681 vs. 1.056; $p_{\text{Holm}} = .021$) and Originality (0.854 vs. 1.208; $p_{\text{Holm}} = .040$). To assess sensitivity to the choice of judge model, we repeat this experiment with three other judge models, with details provided in Appendix E. The results show that relative to the disagreement among human reviewers, we do not detect greater absolute disagreement for human–LLM pairs for most models across most criteria.

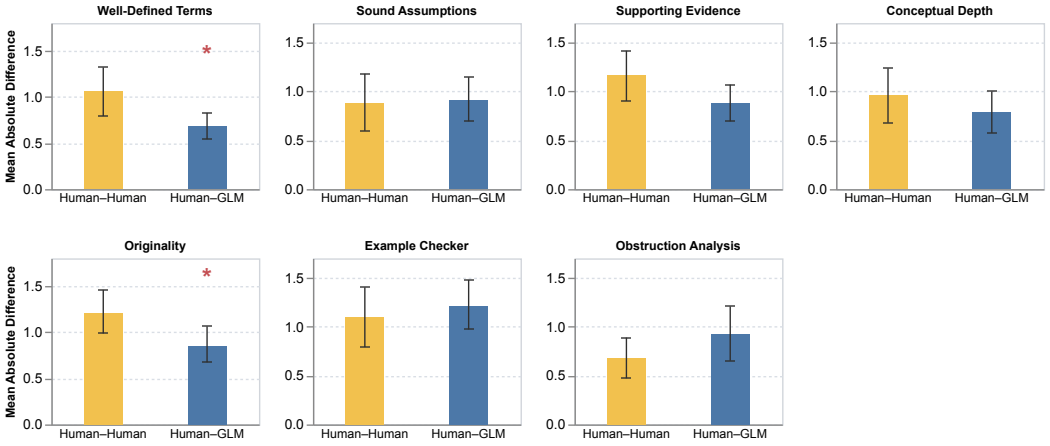


Figure 4: Mean human–human and human–GLM absolute disagreement across 24 proposals. Error bars show 95% bootstrap confidence intervals. Asterisks indicate significant paired differences after Holm correction.

5 Conclusion

In this paper, we present MathIdeaBench, a benchmark for LLM-generated mathematical research proposals. The benchmark consists of two parts: (1) an automated data generate pipeline which produces 478 source papers for ideation in the work, and (2) a MathIdeaJudge comprising expert-designed grading rubrics for targeted evaluation. Our experiment shows that compared to human research, LLMs are less capable in coming up with sound, feasible and diverse research proposals but demonstrate bigger potential in terms of novelty. We position MathIdeaBench as a key step to support the both the AI community in developing better AI systems for mathematical research ideation, and the mathematical community to engage in responsible AI-assisted research.

Limitation and Future Work We identify two limitations of our work. First, our experiment involved only two proposer models and one judge model due to budget constraints. In particular, no frontier model was adopted due to the large dataset size and heavy evaluation workload. This limits the applicability of our findings since different LLMs, especially frontier models, can exhibit considerably different behavior. Second, the scale of our human study is limited. Despite our best effort, the recruitment of domain experts for this study proved challenging due to the required amount of time and effort.

We anticipate that MathIdeaBench will serve as the foundation for developing more advanced agentic workflows in mathematical research. With a refreshable data pipeline, our benchmark offers a future-proof mechanism to compare AI-generated proposals against human-authored publications. Beyond evaluation, MathIdeaJudge provides targeted feedback signals that can drive iterative refinement in autonomous research agents. Furthermore, in expert-in-the-loop systems (Zheng et al., 2026), our framework offers an effective tool to emulate standard mathematical research workflows and accelerate the ideation phase.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure:

- Designing or providing feedback on research methodology or experiments: We used LLMs to improve the clarity of prompts which are initially designed by the authors, without modifying the original meanings and intentions of the prompts.

We have not used generative AI tools for the following tasks with required disclosure: generating synthetic data sets, helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses, implementing methods, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, and interpreting results.

The following required disclosure tasks are not applicable to this work: formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation.

Additionally, we used generative AI tools for the following tasks with recommended disclosure:

- Creating or modifying scientific figures or images: We used LLMs to configure settings for the code creating data visualizations. The methods for data analysis were decided by the authors without the use of LLMs.
- Creating or editing software code: The codebase for the experiment was developed with the help of LLMs. The architecture of the codebase was decided by the authors before starting the development.
- Editing a research paper to improve readability: LLMs were used to polish the research paper. The initial draft was created by the authors without the use of LLMs, and LLM-suggested feedback is incorporated without modifying the original meanings and intentions of the authors,

We have reviewed all AI-assisted work. The prompt modifications suggested by LLMs were reviewed by two authors before incorporating parts of the suggestions. The LLM-generated codebase for experiment was reviewed and verified and tested for correctness by two authors. The code for data visualization is created from an initial human-written version with LLM suggesting modification of layouts and stylistic settings. These suggestions were reviewed and tested for correctness by one author. The research paper polished with the help of LLMs was reviewed by all authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

While automated ideation and evaluation systems can accelerate mathematical discovery, they also present ethical considerations regarding academic integrity and the potential displacement of human researchers. In light of such concerns, we position this work as a contribution to responsible AI-assisted scientific discovery. Our framework is designed as an auxiliary tool to empower human mathematicians rather than a replacement for human creativity. Furthermore, relying on automated judges without human oversight risks rejecting unconventional but impactful ideas or failing to detect inadvertent plagiarism. We advocate that automated ideation frameworks should be deployed exclusively as exploratory tools, with human mathematicians remaining the ultimate arbiters of the appropriate research direction to pursue.

REPRODUCIBILITY STATEMENT

In an effort order to ensure the reproducibility of our results, we have included the necessary details in the construction of benchmark dataset and the evaluation process. The dataset construction process is detailed in Section 3. The settings for LLM inference we adopted for our experiments are reported in Appendix B. All the prompts we use for each step of data generation and evaluation are reported in Appendices F, G and H.

REFERENCES

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 conference of the nations of the Americas chapter of the association for computational linguistics: human language technologies (volume 1: long papers)*, pp. 6709–6738, 2025.

Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. Mlr-bench: Evaluating ai agents on open-ended machine learning research. *Advances in Neural Information Processing Systems*, 38, 2026a.

Nuo Chen, Yicheng Tong, Yuzhe Yang, Yufei He, Xueyi Zhang, Zou Qingyun, Qian Wang, and Bingsheng He. Diversity collapse in multi-agent llm systems: Structural coupling and collective failure in open-ended idea generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 251–306, 2026b.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. *arXiv preprint arXiv:2605.00674*, 2026.

Kefan Dong and Tengyu Ma. Stp: Self-play llm theorem provers with iterative conjecturing and proving. *arXiv preprint arXiv:2502.00212*, 2025.

Tao Feng, Yihang Sun, and Jiaxuan You. Grapheval: A lightweight graph-based llm framework for idea evaluation. In *International Conference on Learning Representations*, volume 2025, pp. 99181–99209, 2025.

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.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M Williams, Stefan Bekiranov, and Aidong Zhang. Ideabench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5888–5899, 2025.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*, 2024.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*, 2023.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13187–13207, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Antoine Peyronnet, Fabian Gloeckle, and Amaury Hayat. Lemmabench: A live, research-level benchmark to evaluate llm capabilities in mathematics. *arXiv preprint arXiv:2602.24173*, 2026.

Yansheng Qiu, Haoquan Zhang, Zhaopan Xu, Ming Li, Diping Song, Zheng Wang, and Kaipeng Zhang. Ai idea bench 2025: Ai research idea generation benchmark. *arXiv preprint arXiv:2504.14191*, 2025.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Evaluating llms' divergent thinking capabilities for scientific idea generation with minimal context. *Nature communications*, 17(1):3625, 2026.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *International Conference on Learning Representations*, volume 2025, pp. 94003–94092, 2025.

Terence Tao. A severe misalignment of ai in mathematics. Blog post, September 2026. URL https://terrytao.wordpress.com/2026/09/11/a-severe-misalignment-of-ai-in-mathematics/.

George Tsoukalas, Anton Kovsharov, Sergey Shirobokov, Anja Surina, Moritz Firsching, Gergely Bérczi, Francisco JR Ruiz, Arun Suggala, Adam Zsolt Wagner, Eric Wieser, et al. Advancing mathematics research with ai-driven formal proof search. *arXiv preprint arXiv:2605.22763*, 2026.

Kaiyu Yang, Gabriel Poesia, Jingxuan He, Wenda Li, Kristin Lauter, Swarat Chaudhuri, and Dawn Song. Formal mathematical reasoning: A new frontier in ai. *arXiv preprint arXiv:2412.16075*, 2024.

Jie Zhang, Cezara Petrui, Kristina Nikolić, and Florian Tramèr. Realmath: A continuous benchmark for evaluating language models on research-level mathematics. *Advances in Neural Information Processing Systems*, 38, 2026.

Daniel Zheng, Ingrid von Glehn, Yori Zwols, Iuliya Beloshapka, Lars Buesing, Daniel M Roy, Martin Wattenberg, Bogdan Georgiev, Tatiana Schmidt, Andrew Cowie, et al. Ai co-mathematician: Accelerating mathematicians with agentic ai. *arXiv preprint arXiv:2605.06651*, 2026.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17744–17761, 2025.

A MATHIDEABENCH DATA DISTRIBUTION

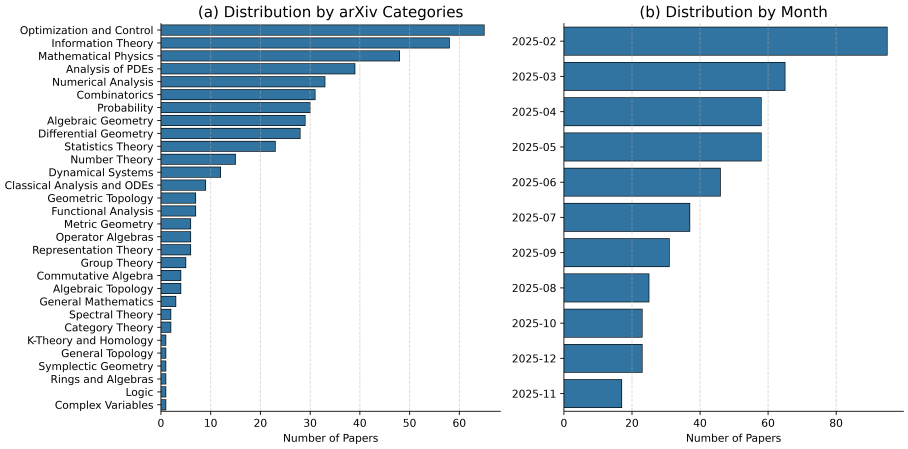


Figure 5: Distribution of MATHIDEABENCH data by arXiv categories and months.

B IMPLEMENTATION DETAILS

All LLM inference was made through the OpenRouter API, which provided a uniform interface. For originality evaluator, we used OpenRouter's server-side web search capability. For consistency across evaluator models, this tool was configured with the Exa search backend and required tool use. Additional details are reported in Table 1.

Table 1: Inference settings by stage.

Stage	Model	Temperature	Reasoning	Prompt
Summarization	GPT-5.6-Terra	—*	High	§F
Proposal generation	DeepSeek-R1	0.7	—†	§G.1
Proposal generation	Gemini-3.5-Flash	0.7	High	§G.1
Proposal extraction	GPT-5.6-Terra	—*	High	§G.2
Soundness	GLM-5.3-Flash	0	High	§H.1
Conceptual depth	GLM-5.3-Flash	0	High	§H.2
Originality	GLM-5.3-Flash	0.7	High	§H.2
Example checking	GLM-5.3-Flash	0	High	§H.3
Obstruction analysis	GLM-5.3-Flash	0	High	§H.3
Topical diversity	GLM-5.3-Flash	0	High	§H.4

* The temperature parameter is omitted because the GPT/Terra provider route does not support it.
† DeepSeek-R1 uses provider-controlled mandatory reasoning, so no reasoning-effort parameter is supplied.

C Additional Evaluation Results

C.1 Statistics of Evaluation Scores

The benchmarking scores for proposals generated by DeepSeek-R1 and Gemini-3.5-Flash are reported in Tables 2 and 3 respectively. The two columns represent the group of follow-up proposals and the group of generated proposals respectively. The fields are formatted as "mean $\pm$ standard deviation," where both the mean and the standard deviation are rounded to four decimal places.

Table 2: Evaluation scores for DeepSeek-R1

Evaluation Criteria	Follow-up Proposals	Generated Proposals
Soundness: Well-Defined Terms	3.1167 ± 0.5538	2.7675 ± 0.6154
Soundness: Sound Assumptions	3.3158 ± 0.7140	2.5450 ± 0.9383
Soundness: Supporting Evidence	3.0033 ± 0.5975	2.1167 ± 0.6783
Soundness: Overall	3.1453 ± 0.4762	2.4764 ± 0.5283
Novelty: Conceptual Depth	2.3333 ± 0.6814	1.9575 ± 0.5680
Novelty: Originality	1.7450 ± 1.0611	1.9233 ± 0.8122
Novelty: Overall	2.0392 ± 0.6938	1.9404 ± 0.5600
Feasibility: Example Checker	2.6667 ± 0.7965	1.7408 ± 0.7376
Feasibility: Obstruction Analyzer	3.2725 ± 0.8596	2.2050 ± 0.9020
Feasibility: Overall	2.9696 ± 0.6853	1.9729 ± 0.7152
Topical Diversity	4.8961 ± 0.6336	4.7613 ± 0.5242

Table 3: Evaluation scores for Gemini-3.5-Flash

Evaluation Criteria	Follow-up Proposals	Generated Proposals
Soundness: Well-Defined Terms	3.1583 ± 0.5494	3.0675 ± 0.4525
Soundness: Sound Assumptions	3.3375 ± 0.7005	2.9517 ± 0.8225
Soundness: Supporting Evidence	3.0592 ± 0.5810	2.5833 ± 0.6338
Soundness: Overall	3.1850 ± 0.4681	2.8675 ± 0.4582
Novelty: Conceptual Depth	2.4642 ± 0.6970	2.9158 ± 0.4170
Novelty: Originality	1.7383 ± 1.0388	2.3842 ± 0.7237
Novelty: Overall	2.1012 ± 0.6732	2.6500 ± 0.4502
Feasibility: Example Checker	2.6692 ± 0.8002	1.9858 ± 0.6279
Feasibility: Obstruction Analyzer	3.2858 ± 0.8295	2.3258 ± 0.8268
Feasibility: Overall	2.9775 ± 0.6794	2.1558 ± 0.6313
Topical Diversity	4.9092 ± 0.5845	4.9103 ± 0.5189

C.2 Hypothesis Testing on Performance Difference

For each grading criterion $\subset$ within Soundness, Novelty and Feasibility, we considered a pair data point corresponding to one source paper p: the average score $S_{\mathcal{G}}(p; \subset)$ of the three LLM-generated proposals and the average score $S_{\mathcal{F}}(p; \subset)$ of the three follow-up proposals. Let ρ denote the probability that $S_{\mathcal{G}}(p; \subset) - S_{\mathcal{F}}(p; \subset) > 0$. For topical diversity, ρ is taken to be the probability that $\text{Div}_{\text{topic}}(\mathcal{G}_p) - \text{Div}_{\text{topic}}(\mathcal{F}_p) > 0$. The means and standard deviations of these differences are shown in Table 4. We perform sign tests with null hypotheses $\rho = 0.5$ and different alternative hypotheses H_1 for each criterion: $\rho < 0.5$ for soundness, feasibility and topical diversity, and $\rho > 0.5$ for novelty. The results are shown in Table 5.

Table 4: Score differences against the human baselines

Model	DeepSeek-R1	Gemini-3.5-Flash
Soundness: Well-Defined Terms	-0.3492 ± 0.4773	-0.0908 ± 0.4003
Soundness: Sound Assumptions	-0.7708 ± 0.6896	-0.3858 ± 0.6551
Soundness: Supporting Evidence	-0.8867 ± 0.5751	-0.4758 ± 0.5045
Soundness: Overall	-0.6689 ± 0.4378	-0.3175 ± 0.3873
Novelty: Conceptual Depth	-0.3758 ± 0.6187	$+0.4517 \pm 0.5452$
Novelty: Originality	$+0.1783 \pm 0.8276$	$+0.6458 \pm 0.7690$
Novelty: Overall	-0.0988 ± 0.5930	$+0.5487 \pm 0.5283$
Feasibility: Example Checker	-0.9258 ± 0.6640	-0.6833 ± 0.6206
Feasibility: Obstruction Analyzer	-1.0675 ± 0.7706	-0.9600 ± 0.7013
Feasibility: Overall	-0.9967 ± 0.6244	-0.8217 ± 0.5763

Table 5: Score differences against the human baselines

Model	Criterion	H_1	p-value-str
DeepSeek-R1	Soundness: Well-Defined Terms	$\rho < 0.5$	$\mathbf{2.637 \times 10^{-34}}$
Gemini-3.5-Flash	Soundness: Well-Defined Terms	$\rho < 0.5$	$\mathbf{9.728 \times 10^{-7}}$
DeepSeek-R1	Soundness: Sound Assumptions	$\rho < 0.5$	$\mathbf{3.005 \times 10^{-57}}$
Gemini-3.5-Flash	Soundness: Sound Assumptions	$\rho < 0.5$	$\mathbf{1.453 \times 10^{-23}}$
DeepSeek-R1	Soundness: Supporting Evidence	$\rho < 0.5$	$\mathbf{1.715 \times 10^{-89}}$
Gemini-3.5-Flash	Soundness: Supporting Evidence	$\rho < 0.5$	$\mathbf{1.609 \times 10^{-58}}$
DeepSeek-R1	Soundness: Overall	$\rho < 0.5$	$\mathbf{9.555 \times 10^{-80}}$
Gemini-3.5-Flash	Soundness: Overall	$\rho < 0.5$	$\mathbf{9.554 \times 10^{-42}}$
DeepSeek-R1	Novelty: Conceptual Depth	$\rho < 0.5$	1.000
Gemini-3.5-Flash	Novelty: Conceptual Depth	$\rho < 0.5$	$\mathbf{1.999 \times 10^{-41}}$
DeepSeek-R1	Novelty: Originality	$\rho < 0.5$	$\mathbf{1.429 \times 10^{-4}}$
Gemini-3.5-Flash	Novelty: Originality	$\rho < 0.5$	$\mathbf{2.324 \times 10^{-40}}$
DeepSeek-R1	Novelty: Overall	$\rho < 0.5$	1.000
Gemini-3.5-Flash	Novelty: Overall	$\rho < 0.5$	$\mathbf{2.483 \times 10^{-56}}$
DeepSeek-R1	Feasibility: Example Checker	$\rho < 0.5$	$\mathbf{1.861 \times 10^{-81}}$
Gemini-3.5-Flash	Feasibility: Example Checker	$\rho < 0.5$	$\mathbf{3.043 \times 10^{-59}}$
DeepSeek-R1	Feasibility: Obstruction Analyzer	$\rho < 0.5$	$\mathbf{2.201 \times 10^{-78}}$
Gemini-3.5-Flash	Feasibility: Obstruction Analyzer	$\rho < 0.5$	$\mathbf{2.789 \times 10^{-77}}$
DeepSeek-R1	Feasibility: Overall	$\rho < 0.5$	$\mathbf{1.880 \times 10^{-93}}$
Gemini-3.5-Flash	Feasibility: Overall	$\rho < 0.5$	$\mathbf{2.222 \times 10^{-78}}$
DeepSeek-R1	Topical Diversity	$\rho < 0.5$	$\mathbf{5.915 \times 10^{-9}}$
Gemini-3.5-Flash	Topical Diversity	$\rho < 0.5$	0.401

C.3 Regarding Evaluation of Originality

During a sanity check, we found that when GLM-5.3-Flash is evaluating originality, for some follow-up proposals, it failed to follow the restriction on excluding the follow up paper's source. This led to a lower originality score for some follow-up papers. After accounting for this, the originality gap decreased modestly for both models, bringing DeepSeek-R1 close to parity while reducing–but not eliminating–Gemini's positive advantage.

D HUMAN STUDY DETAILS

Participants. We recruited seven reviewers whose expertise spans probability, combinatorics, geometry, analysis, statistics, and mathematical physics. The group comprised four postdoctoral researchers, one researcher transitioning from a PhD to a postdoctoral position, one visiting professor, and one research studies manager. All reviewers had authored peer-reviewed mathematics publications, and six had previously reviewed for mathematics conferences or journals.

Paper Assignment. Each reviewer was assigned one or two papers based on the alignment between their expertise and the paper's subject area and mathematical techniques. Given the high degree of specialization within mathematics, the strength of this match varied across assignments. Of the four papers used in the human evaluation study, two received two independent reviews and two received three.

Evaluation Procedure. Reviewers received the scoring rubrics and a worked example containing an expert's ratings and justifications. They evaluated all six proposals associated with each assigned paper through a Google Form and were given ten days to complete the task. Reviewers were instructed to assess each proposal independently rather than using their assessment of one proposal to determine the score of another. Among the four papers included in the alignment analysis, the LLM-generated proposals for two papers were produced by DeepSeek-R1 and those for the other two by Gemini-3.5-Flash. Proposal order was randomized, and reviewers were not informed whether a proposal was generated by an LLM or extracted from a citing work.

E Additional Human Evaluation Results

E.1 Differences in Assigned Scores

We also tested whether GLM systematically assigned higher or lower scores than human reviewers. We compared its scores with the proposal-level mean human scores using two-sided Wilcoxon signed-rank tests with Holm correction across the seven criteria. GLM assigned significantly lower scores for example checker (2.500 vs. 3.271; $p_{\text{Holm}} = .012$), while no significant difference was detected for the remaining criteria.

E.2 Alternative Judge Models

To assess sensitivity to the choice of judge, we repeated the human-alignment analysis with Grok-4.6, GPT-5.6-Terra, and DeepSeek-v4-Pro using the same 24 proposals. Figure 6 reports the resulting human–human and human–LLM disagreement estimates. After applying Holm correction across the seven criteria separately for each judge, DeepSeek-v4-Pro exhibited significantly greater human–LLM disagreement for obstruction analysis. No criterion with significantly greater human–LLM disagreement was detected for Grok-4.6 or GPT-5.6-Terra.

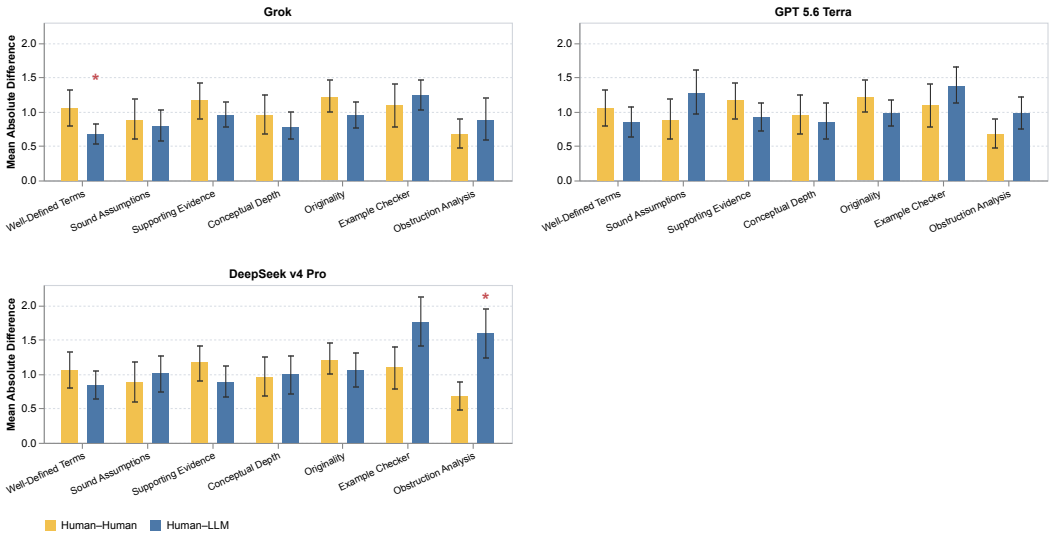


Figure 6: Human–human and human–LLM absolute disagreement for three alternative judge models across 24 proposals. Bars show mean disagreement, and error bars show 95% bootstrap confidence intervals. Asterisks indicate significant paired differences after Holm correction, applied separately for each judge.

F PROMPTS FOR SUMMARIZATION

Summarizer: System Prompt

You are a world-class research mathematician specializing in identifying open problems and potential research directions based on existing literature.

Summarizer: User Prompt

=== TASK ===
Your task is to analyze a given math paper and generate a rigorous, detailed and self-contained summary specifically designed to facilitate **open problem formulation**.
Another mathematician should be able to read your summary and immediately start formulating new research questions without needing to reference the original text for definitions or theorem statements.

=== OUTPUT FORMAT ===
Follow the Markdown structure below strictly. Always use LaTeX for all mathematical notation.

— START OF MARKDOWN STRUCTURE —

Summary of {`paper_title`}
1. Introduction & Context
In this section, provide a high-level overview of the paper's context and contributions.
This includes the main problem addressed, related existing work, and the specific gap or limitation that this paper resolves.

2. Precise Setup
In this section, provide a rigorous and self-contained description of the mathematical setup.
This includes the definitions of the objects, spaces, and structures involved, as well as any standing assumptions or conditions.

3. Main Result
In this section, provide a rigorous and detailed summary of the main result of the paper.
This includes the precise statements of the main theorem(s) along with any necessary conditions or assumptions.

4. Proof Strategy

In this section, explain how the authors prove the main result and highlight any novel or non-trivial aspects.
You should include the following information:

- the underlying intuition or heuristics as to why the results should be true, and
- a brief explanation of the proof techniques, methods or technical lemmas and how the ingredients come together.

5. Examples
In this section, provide one or two examples, concrete cases or applications in the paper used to illustrate the main results
If this is not present in the paper, simply state that the authors do not provide illustrating examples; do NOT invent examples on your own.

— END OF MARKDOWN STRUCTURE —

=== RULES & CONSTRAINTS ===

When summarizing, you must adhere to the following principles strictly:

1. Fidelity: Summarize only what is present in the paper. Do not hallucinate results or search for external information.
2. Rigor: Maintain high mathematical precision. Notations must be defined before use and kept consistent.
3. Dependency: If Theorem A relies on Lemma B or Assumption C, make that relationship explicit.

=== INPUT PAPER TO SUMMARIZE ===
— START OF PAPER —

{input_paper}

— END OF PAPER —

G PROMPTS FOR PROPOSAL GENERATION AND EXTRACTION

G.1 PROPOSER

Proposer: System Prompt

You are a world-class research mathematician specializing in formulating high-quality, standalone research proposals. Use the supplied mathematical summary only as private source material. Every proposal you write must read as an independent document and must never refer to that source, its authors, the summary, the generation process, or previously generated proposals.

Proposer: User Prompt

=== TASK ===
Using the supplied mathematical summary as private inspiration, generate exactly one standalone research proposal: Proposal {proposal_number}. It may lie in the same area as the summarized mathematics or in a genuinely connected area. Aim for a high-impact but feasible proposal. A reader must be able to read it without knowing that a source document, summary, or proposal-generation sequence exists.

=== DIVERSITY CONTEXT ===
The following block contains only the titles and Introduction & Context sections of proposals already generated from the same private summary. Use it solely to avoid duplicating their primary targets. Do not treat these snippets as mathematical premises, extend them mechanically, mention them, or refer to "another" or "previous" proposal in the output.

{prior_proposal_context}

Proposal {proposal_number} must pursue a meaningfully distinct primary target. It must not be a renaming, restatement, immediate corollary, or minor variant of any target represented above.

=== RULES ===
The proposal must satisfy the following criteria:

- Write strictly in the first-person plural ("We propose...", "Our goal is...", "We envision...").
- Write as a completely standalone proposal. Never mention or allude to "the paper," "this paper," "the source paper," "the supplied summary," its authors, the prompt, the generation process, or earlier proposals. Do not describe what a source document proves, studies, establishes, or leaves open. State the relevant mathematical context, known facts, and open gap directly.
- The proposed main result must be centered on one primary mathematical target (e.g., an existence theorem, an explicit inequality bound, or a specific algebraic/-geometric construction) which is concrete, self-contained. It should not be a broad research direction or vague question.
- The proposed main result must be genuinely open: do not propose anything already proved, disproved, or known in the literature according to the supplied summary.
- The proposal must be precisely formulated, using well-defined objects, conditions, quantifiers, etc.
- The proposal must be mathematically motivated by the supplied summary without revealing or referring to that relationship in the output.

=== OUTPUT FORMAT ===
Return exactly one Markdown proposal with the number {proposal_number}. Always use LaTeX for all mathematical notation.

— START OF MARKDOWN STRUCTURE —

Proposal {proposal_number}: <concise statement of the main research question in this proposal>
1. Introduction & Context
In this section, provide a high-level description of your proposal.
This includes the main problem addressed, related existing work, and the specific gap or limitation that this proposal resolves.
Present that context directly as mathematical background; do not recap or cite the source document.

2. Precise Setup
In this section, provide a rigorous and self-contained description of the mathematical setup. This includes the definitions of the objects, spaces, and structures involved, as well as any standing assumptions or conditions.

3. Main Result
In this section, provide a rigorous and detailed description of the one primary result being proposed.
This includes the precise statements of the main theorem along with any conditions or assumptions.
To substantiate your claims, provide heuristics/high-level arguments to explain why the proposed result is likely to be true.

— END OF MARKDOWN STRUCTURE —

Do not include any other proposal or any text outside this structure.

=== PRIVATE SOURCE SUMMARY ===
Use this material to reason about the mathematics, but never mention it or its existence in the proposal.
— START OF SUMMARY —

{paper_summary}

— END OF SUMMARY —

G.2 Proposal Extractor

Proposal Extractor: System Prompt

You are a world-class research mathematician. Convert an existing mathematical result into a concise, forward-looking research proposal while preserving its mathematical meaning.

Proposal Extractor: User Prompt

=== TASK ===
Rewrite the supplied mathematics-paper summary as exactly one formal research proposal. Select the paper's single primary result and present it as a proposed research goal that has

not yet been established.

=== STYLE EXAMPLES ===
The following are two unrelated proposals for style calibration. Treat them together as the target distribution for presentation: closely match their typical overall length, section proportions, paragraph scale, notation density, theorem granularity, explanatory depth, and amount of proposed proof strategy. Keep the output within their approximate length and detail range whenever the selected result can be stated in a correct and self-contained manner within that range.

Match form and granularity, not mathematical content. Do not copy the exemplar's terminology, objects, claims, assumptions, or subject-specific language, and do not infer anything about the source paper from it.

— START OF STYLE EXAMPLES —

{style_exemplar_proposals}

— END OF STYLE EXAMPLES —

=== VOICE & TONAL ALIGNMENT ===

1. Write strictly in the first-person plural (for example, "We propose...").
2. Recast established results and completed proofs as forward-looking goals, conjectures, or intended approaches.
3. Never mention the paper, summary, authors, extraction process, or examples. The output must read as a standalone proposal.

=== CRITICAL RULES ===

1. **Fidelity**: Use only information in the supplied paper summary. Preserve the selected result's core claim and essential hypotheses. Do not introduce external mathematical ideas.
2. **Focused selection**: Center the proposal on the selected primary result and the context and definitions needed to make it self-contained. Omit an exhaustive catalogue of secondary theorems, variants, examples, applications, and technical lemmas. Retain enough motivation and detail to match the examples, but do not make the extracted proposal more exhaustive, formal, or proof-complete merely because the source summary contains those details.
3. **One target**: State one primary proposed result, not a catalogue of the paper's contributions. Prioritize a precise and adequately developed target statement over proof methodology. Include no more proof strategy than the examples do; compress a detailed proof from the summary into a brief high-level approach or heuristics.
4. **Close style calibration**: Make the output plausibly indistinguishable in presentation from the style exemplars. Match the examples' concision and balance among sections. In particular, keep the setup to the minimum definitions and assumptions required for the main result; it should not become substantially denser or longer than the examples' setups simply because the summary supplies more notation.
5. **Rigor**: Maintain mathematical precision and use LaTeX for notation.

=== OUTPUT FORMAT ===
Follow this Markdown structure exactly. Do not add other sections or third-level headings. Always use LaTeX for mathematical notation.

— START OF MARKDOWN STRUCTURE —

Proposal 1: <concise title for the proposed result>
1. Introduction & Context
Motivate the research direction and the gap addressed by the primary result at a level and length closely calibrated by the style examples. Present the context directly without recounting what the source paper does.

2. Precise Setup
Define the objects, assumptions, and notation needed to state the primary result in a self-contained manner. Match the examples' typical density and section length as closely as possible. Prefer compact definitions and omit notation that is not used directly in the primary result.

3. Main Result
State and develop the single primary result as a proposed goal or conjecture along with any essential conditions or assumptions.
To support the development of the proposal, use heuristics/high-level arguments extracted from the summary. Do NOT include a full proof or detailed methodology.
The heuristics arguments must be faithfully extracted from the summary, and must not introduce any new ideas or external knowledge. The heuristics should motivate the proposed result and should not outweigh the importance of the proposed result itself.

Use the examples to calibrate the expected level of detail and length in this section.

— END OF MARKDOWN STRUCTURE —

=== PAPER SUMMARY TO EXTRACT FROM ===

All mathematical content in the output must come from this summary, not from the style examples.

— START OF SUMMARY —

{paper_summary}

— END OF SUMMARY —

H PROMPTS FOR EVALUATION

H.1 SOUNDNESS

Soundness: System Prompt

You are an expert mathematician and reviewer evaluating open research problem proposals. Your job is to assess the soundness of the proposal in terms of its clarity, rigor, and evidence, specified by a list of criteria.
Evaluate each criterion objectively against the rubric. Do not double-penalize errors across multiple criteria; assign each flaw strictly to its primary criterion.

Soundness: User Prompt

=== TASK ===
You will evaluate the soundness of a research proposal on the three criteria given below. Each criterion is scored on a scale of 0–4. Provide justification before deciding on a score.

=== CRITERIA ===
Criterion 1 — Well-Defined Terms
What this evaluates: Are all non-standard mathematical jargon, notation, and domain-specific terminology precisely specified upon first use (standard/ubiquitous terms excluded)?
4 - All technical jargon, variables, and mathematical objects are precisely defined on first use.
3 - Some terms are slightly ambiguous or missing standard notation, but intent is clear.
2 - Core terms have shifting or informal definitions, forcing the reader to guess exact meanings.
1 - Key mathematical objects or technical frameworks are undefined, making validation impossible.
0 - Pervasive misuse or total absence of basic domain definitions; text reads as pseudo-technical jargon.

Criterion 2 — Sound Assumptions
What this evaluates: Are the foundational mathematical conditions, axioms, and scope bounds explicitly declared and mathematically valid?
4 - All mathematical conditions, axioms, and constraints defining the problem space are explicitly stated, mathematically consistent, and standard/realistic for the domain.
3 - A minor background condition is implicit rather than stated, but it is standard in the field and trivial to infer without altering the problem's meaning.
2 - Relies on an overly restrictive, questionable, or unstated condition that narrows the scope or makes the problem trivial/uninteresting.
1 - Core mathematical setup relies on an ill-posed premise or a condition that contradicts standard definitions in the domain.
0 - Violates fundamental mathematical laws or axioms (e.g., assuming a non-existent object exists), rendering the problem mathematically impossible or meaningless.

Criterion 3 — Supporting Evidence
What this evaluates: Does the proposal provide enough plausible heuristics or high-level arguments to support the claimed conjectures?
Do not penalize the proposal for lacking full formal proofs. Instead, assess whether the proposal presents a clear roadmap connecting their assumptions to their conjectures via compelling intuition, analogies, structural connections, or heuristic arguments.
4 - Conjectures are explicitly framed as open questions and backed by sound high-level heuristics, structural connections or analogies.

3 - Conjectures are well-motivated with clear intuitive framing and macro-level connections, though supporting arguments contain minor conceptual leaps.
2 - Conjectures are stated with isolated intuitive or heuristic support, but the overall argument fails to build a cohesive macro-level picture.
1 - Conjectures are stated with very little or fundamentally flawed supporting evidence as to why the proposal is plausible.
0 - Conjectures are stated without supporting evidence or as proven facts with no justification.

=== PROPOSAL ===
{proposal}

=== OUTPUT FORMAT ===
Write your evaluation using exactly the headings below. Give justification first, then the score.

1. Well-Defined Terms
Justification: <2-3 sentences>
Score: <int 0–4>
2. Sound Assumptions
Justification: <2-3 sentences>
Score: <int 0–4>
3. Supporting Evidence
Justification: <2-3 sentences>
Score: <int 0–4>

H.2 NOVELTY

Conceptual Depth: System Prompt

You are a world-class mathematical research evaluator with broad knowledge of the mathematics literature. Your job is to assess the conceptual depths of research proposals.

Conceptual Depth: User Prompt

=== TASK ===
You will receive one research proposal associated with a seed paper. Evaluate this proposal's conceptual depth independently on a scale of 0–4 based on the criterion provided below. Provide a justification of 2-3 sentences before deciding on a score.

=== Criterion: CONCEPTUAL DEPTH (0–4) ===
Does the proposal demonstrate meaningful structural or conceptual insight beyond a surface-level modification of the seed paper?

4 - Introduces major new ideas, tools, techniques, frameworks, or perspectives capable of advancing domain understanding.
3 - Introduces non-trivial ideas which are structurally or conceptually relevant to the field, but not as groundbreaking as a 4.
2 - Presents some genuine content but the insight is incremental or narrow.
1 - Presents a small generalization of the seed paper with limited new insight.
0 - Presents a trivial modification or rephrasing of the seed paper with no new insight.

=== SEED PAPER SUMMARY ===

{paper_summary}

=== PROPOSAL ===
{proposal}

=== OUTPUT FORMAT ===
Respond ONLY with a valid JSON object matching the schema below.

{{
"justification": "<2-3 sentences analyzing the proposal's conceptual depth>",
"conceptual_depth_score": <integer 0-4>
}}

Originality: System Prompt

You are a world-class mathematical research evaluator specializing in literature analysis and academic priority. Your sole job is to assess the originality of open research proposals relative to existing mathematics literature. Originality evaluates whether the proposal introduces a genuinely new mathematical setup, open question, or approach. This includes the following two aspects:
- Problem Originality: Posing a brand-new, unstudied question or conjecture.
- Methodological Originality: Proposing a novel, non-obvious framework or technique applied to an existing, well-known problem.

Originality: User Prompt

=== TASK ===
You will receive a single research proposal and evaluate its ORIGINALITY relative to existing mathematical literature published prior to {paper_date}.
You MUST search the web to identify any prior work, theorems, conjectures, or methodologies closely related to the proposal.

=== RULES & CONSTRAINTS ===
- **Temporal Cutoff:** ONLY consider and search for research published, presented, or archived BEFORE {paper_date}. Strictly ignore any work published after this date.
- **Targeted Literature:** Prioritize foundational papers or closely related formulations.
- **Accuracy & Citing:** Use real, verifiable paper titles, primary authors, and publication dates.
- **No Hallucinations:** Do not invent or fabricate citations. If a specific paper title is uncertain, describe the precise line of research or state "No identical prior work identified in literature".

=== CRITERION: ORIGINALITY (0–4) ===
To what extent has the proposed problem, conjecture, setup, or technique been explicitly posed, studied, or executed in published literature prior to {paper_date}?

4 - Unprecedented: The central question, conjecture, setup, OR proposed technique is genuinely new; no prior published work poses this exact formulation, an equivalent problem, or applies this specific framework to the target space.
3 - Distinct Angle / Technique: The problem or goal is related to existing literature, but the proposal introduces a distinct, non-obvious angle or novel technique that has not been formally targeted or applied before.

2 - Standard Variant / Machinery: Similar questions or setups exist in closely related contexts; the proposal represents a standard variant, routine parameter extension, or applies off-the-shelf machinery already conventional for this domain.
1 - Substantially Explored: Very similar formulations, special cases, or techniques have already been explicitly studied in published papers, leaving only superficial differences or extensions.
0 - Identical / Solved: The proposed problem, conjecture, or identical technique has already been explicitly posed, executed, or fully solved in existing literature prior to {paper_date}.

=== PROPOSAL ===
{proposal}

=== OUTPUT FORMAT ===
Respond ONLY with a valid JSON object matching the exact schema below.
Do not include any text, markdown code blocks, or commentary outside the JSON block.

{{
"closest_existing_works": ["<specific paper (Title, Author, Year)>"],
"analysis": "<2-3 sentences analyzing specific similarities and differences relative to prior work>",
"originality_score": <integer 0-4>
}}

Use an empty list for "closest_existing_works" when no identical or closely related prior work can be verified.

H.3 FEASIBILITY

Example Checker: System Prompt

You are an Example Tester who evaluates the feasibility of mathematical research proposals by constructing and analyzing concrete instances.

Example Checker: User Prompt

=== TASK ===
You will evaluate a research proposal by actively testing its claims against concrete examples, boundary conditions, and potential counterexamples.

Walk through the five diagnostic dimensions below to identify any mathematical breakdowns, edge-case failures, or counterexamples:

1. Toy Examples: Simple, explicit instances where the proposal should apply.
2. Edge Cases: Boundary/degenerate conditions (e.g., $n=0$, $n=1$, trivial topologies, extreme parameters) where definitions may break down.
3. Counterexample Search: Targeted cases specifically designed to disprove or contradict the core claims.
4. Known Instances: Standard canonical examples or literature benchmarks the proposal must pass.
5. Generalization Check: Instances revealing whether the claim is vacuously true (too weak) or mathematically impossible (too strong).

=== SCORING CRITERION: FEASIBILITY (0–4) ===
Consolidate all identified counterexamples, gaps, or test-case failures into a single overall score measuring how easily they can be patched or resolved.

4 - No Violation / Fully Valid: All tested examples, boundary cases, and canonical instances hold true as claimed; no counterexamples found.
3 - Minor / Easily Patched: Issues are cosmetic or localized to edge cases; easily resolved by adding standard boundary conditions or minor hypothesis tweaks without altering the core thesis.
2 - Moderate / Non-Trivial Reformulation: Uncovers a noticeable gap or counterexample that weakens a supporting argument; requires non-trivial reformulation or scope reduction, but the core intuition remains salvagable.
1 - Significant / Fatal Gap: A major counterexample puts the central claim in serious doubt; avoiding it requires a fundamental restructuring of the proposal's main framework.
0 - Blocking / Infeasible: A clear counterexample or fundamental contradiction completely negates the core proposal; no known mathematical workaround exists.

=== PROPOSAL ===
{proposal}

=== OUTPUT FORMAT ===
Write your analysis in markdown. For each tested case use this format:

<Case Type>: <Brief Description>
Construction
<specific mathematical instance>
Outcome
<whether the proposed result holds or breaks for this case>
Analysis
<briefly explain how this affects the scores>

After all cases, end with:

Summary
<one short paragraph for overall assessment>
Feasibility Score
Score: <integer 0–4>

Obstruction Analyzer: System Prompt

You are an expert Obstruction Analyzer who identifies concrete mathematical barriers that could prevent a research proposal from succeeding. Be skeptical but evidence-disciplined. Distinguish a genuine obstruction from the ordinary fact that a new research claim still needs to be proved.

Obstruction Analyzer: User Prompt

=== TASK ===
Evaluate whether there are concrete, evidence-supported reasons that this research proposal could fail or would require a material change to its statement or central method.

The proposal is intentionally written as prospective research. Its central result is not expected to be proved in the proposal. Apply these rules strictly:

1. The fact that a theorem, lemma, estimate, construction, or new framework has not yet been proved or established is not an obstruction.
2. Novelty, ambition, technical depth, a long proof, or the need to develop new machinery is not by itself evidence against feasibility.
3. Do not penalize the proposal because its key lemma is "the real theorem," because its main mechanism still needs justification, or because existing literature does not already contain the proposed advance.
4. A scored barrier must identify an independent reason for possible failure: for example, a concrete counterexample, a conflict with an established theorem, an invalid displayed inference, incompatible definitions, a clearly missing necessary assumption, or a known method-specific limitation with an explained failure mechanism.
5. Statements that something "may" fail are not barriers unless you give a specific mathematical mechanism or evidence making that failure credible.
6. A restriction in prior literature is not evidence of a structural obstruction unless you explain why that restriction is mathematically necessary or identify a failure without it.
7. Do not invent one entry for every diagnostic category. Include only categories for which you found a concrete barrier. If you find none, say so explicitly.

Use these diagnostic lenses when relevant:

- Theoretical barriers: explicit contradictions, counterexamples, impossibility results, or incompatible definitions.
- Technical barriers: a specific proposed step that cannot deliver the claimed conclusion for a stated mathematical reason. Mere difficulty is not enough.
- Scope barriers: dependencies on distinct unresolved results only when failure of those dependencies would force a material reformulation. The amount of work alone is not a barrier.
- Past failures: documented failures with a mechanism that applies to this proposal. Do not speculate from an absence of prior success.
- Blindspots: concrete omitted boundary cases, compatibility conditions, regularity assumptions, or quantifier problems that affect the theorem.

=== SCORING CRITERION: FEASIBILITY (0–4) ===
Score only the concrete obstructions you established. Do not lower the score for ordinary proof obligations or research difficulty.

4 - No Concrete Obstruction: No evidence-supported reason was found that the statement or central method should fail. This score is appropriate even when the project is ambitious, technically deep, or requires new mathematics.
3 - Minor Concrete Obstruction: A demonstrated localized issue requires a standard clarification, boundary condition, or light hypothesis change without altering the core thesis.
2 - Material Concrete Obstruction: A demonstrated failure mechanism requires a non-trivial change to the theorem's scope, assumptions, or central proof strategy. Mere meaningful effort does not qualify.
1 - Severe Concrete Obstruction: A counterexample, established conflict, or failed central inference puts the main claim in serious doubt, although a major reformulation might remain possible.
0 - Blocking Concrete Obstruction: A fatal contradiction or impossibility negates the core proposal and no credible mathematical reformulation is apparent.

Before assigning a score from 0 to 2, verify that every score-driving criticism identifies a failure that would remain after merely supplying the proposal's currently omitted proof steps, without changing its statement, assumptions, definitions, or central method. If a criticism only says that the theorem or its key new ingredient currently lacks a proof,

exclude it from the score.

=== PROPOSAL ===
{proposal}

=== OUTPUT FORMAT ===
Write your analysis in markdown. For each identified barrier use this format:

<Barrier Type>: <Brief Description>
Description
<identify the exact claim or proof step, the concrete evidence, and the independent failure mechanism>
Severity
<state the material change needed and explain why this is an obstruction rather than an ordinary proof obligation>

If no concrete barriers are found, include one section using this format:
No Concrete Obstructions Identified
Description
<briefly summarize what was checked and why no evidence-supported obstruction was found>
Severity
No score reduction is warranted. Difficulty or novelty alone does not lower feasibility.
After all barriers, end with:
Summary
<one paragraph overall assessment of barriers and how easily they can be overcome>
Feasibility Score
Score: <integer 0–4>

H.4 Topical Diversity

Topical Diversity: System Prompt

You are an expert mathematician who classifies mathematical research into Mathematics Subject Classification 2020 (MSC2020) codes. You identify the most relevant subject areas for a given research proposal. Respond in the exact format specified.

Topical Diversity: Step 1, Section Identification

=== TASK ===
For the research proposal below, identify the most relevant section codes of the Mathematics Subject Classification 2020 (MSC2020). The codes are two-digit numbers that represent broad subject areas in mathematics.
Select 1 to 3 section codes that best capture the main topics of the proposal. Only include codes that are genuinely relevant to the proposal's content.

=== OUTPUT FORMAT ===
ONLY output the following JSON structure:
{{
  "justification": "<one sentence explaining your classification>",
  "sections": ["<two-digit section code>", "..."]
}}

=== PROPOSAL ===
{proposal}

=== MSC2020 SECTIONS ===
{section_list}

=== YOUR CLASSIFICATION ===

Topical Diversity: Step 2, Subsection Identification

=== TASK ===
For the research proposal below, identify the most relevant subsection codes of the Mathematics Subject Classification 2020 (MSC2020). The subsection codes consist of three characters that represent specific sub-areas in mathematics.
The two-digit section codes which best capture the main areas of the proposal are provided below.
For each two-digit section code, identify the most relevant (1 to 3) three-character subsection codes which best represent the specific focus of the proposal. Only include codes that are genuinely relevant to the proposal's content.

=== OUTPUT FORMAT ===
ONLY output the following JSON structure:
{{
  "justification": "<one sentence explaining your classification>",
  "subsections": ["<three-character subsection code>", "..."]
}}

=== PROPOSAL ===
{proposal}

=== IDENTIFIED MSC2020 SECTIONS ===
{section_list}

=== CANDIDATE MSC2020 SUBSECTIONS ===
{subsection_list}

=== YOUR CLASSIFICATION ===

Topical Diversity: Step 3, Subject Identification

=== TASK ===
For the research proposal below, identify the most relevant subject codes of the Mathematics Subject Classification 2020 (MSC2020). The subject codes consist of five characters that represent specific subject areas in mathematics.
The three-character subsection codes which best capture the sub-areas of the proposal are provided below.
For each three-character subsection code, identify the most relevant (1 to 3) five-character subject codes which best represent the specific focus of the proposal. Only include codes that are genuinely relevant to the proposal's content.

=== OUTPUT FORMAT ===
ONLY output the following JSON structure:
{{
  "justification": "<one sentence explaining your classification>",

"subjects": ["<five-character subject code>", "..."]
}}

=== PROPOSAL ===
{proposal}

=== IDENTIFIED MSC2020 SUBSECTIONS ===
{subsection_list}

=== CANDIDATE MSC2020 SUBJECTS ===
{subject_list}

=== YOUR CLASSIFICATION ===

I REPRESENTATIVE PROPOSAL COMPARISON

We illustrate the pipeline using the seed paper *A note on a cluster structure of the coordinate ring of a simple algebraic group* (arXiv:2504.09011). The paper establishes an unlocalized cluster-algebra realization of $\mathbb{C}[G]$ for simply connected simple complex algebraic groups, excluding type F_4. For brevity, we omit the seed paper's full formal setup, proof details, and the remaining generated and follow-up proposals. The evaluation results are shown in Table 6.

Abstract of (arXiv:2504.09011): A note on a cluster structure of the coordinate ring of a simple algebraic group

We show that the coordinate ring of a simply-connected simple algebraic group G over the complex number field coincides with the Berenstein–Fomin–Zelevinsky cluster algebra and its upper cluster algebra, at least when G is not of type F_4.

Generated proposal. The following proposal is the third one generated by Gemini-3.5-Flash:

Unlocalized Cluster Realization of the Affine Basic Affine Space $\widehat{G}/\widehat{U}$ via Infinite-Rank Geiss-Leclerc-Schröer Theory

1. Introduction & Context

In this proposal, we address the fundamental problem of establishing the unlocalized cluster-algebraic structure of the coordinate rings of infinite-dimensional varieties arising in representation theory. Specifically, we focus on the affine basic affine space $\widehat{G}/\widehat{U}$ associated with a simply-connected affine Kac-Moody group $\widehat{G}$ and its maximal unipotent subgroup $\widehat{U}$.

For a finite-dimensional simply-connected simple algebraic group G, a landmark line of research has established that the coordinate ring of the basic affine space $\mathbb{C}[G/U]$ is isomorphic to the cluster algebra $\mathscr{A}$ and the upper cluster algebra $\mathscr{U}$ associated with an appropriate initial seed, without inverting the frozen variables. This unlocalized identity $\mathbb{C}[G/U] = \mathscr{A} = \mathscr{U}$ is of central importance because it provides a precise, coordinate-free framework for studying dual canonical bases, quantum group representations, and the geometry of flag varieties.

In the infinite-dimensional setting of loop groups and affine Kac-Moody groups, the coordinate rings of these varieties are of profound importance in mathematical physics (such as the Wess-Zumino-Witten model and conformal field theory) and integrable systems. While localized cluster structures on double Bruhat cells of Kac-Moody groups have been constructed (for instance, by inverting certain frozen variables), the unlocalized identity for the affine basic affine space $\widehat{G}/\widehat{U}$ remains a major open problem.

The primary obstacles to establishing this unlocalized identity in the affine setting are threefold:

1. **Infinite-Rank Cluster Algebras:** The coordinate ring $\mathbb{C}[\widehat{G}/\widehat{U}]$ is infinite-dimensional, requiring the formulation of cluster algebra structures of infinite rank.
2. **The Geiss-Leclerc-Schröer (GLS) Criterion at Infinite Rank:** The standard GLS criterion, which is the primary tool for proving unlocalized cluster identities by using disjoint clusters, is formulated for finite-rank cluster algebras and assumes the coordinate ring is a unique factorization domain (UFD). Extending this to the infinite-rank setting requires a careful treatment of direct limits of UFDs.
3. **Generation by Affine Generalized Minors:** It must be proven that the coordinate ring $\mathbb{C}[\widehat{G}/\widehat{U}]$ is generated as a $\mathbb{C}$-algebra by the affine generalized minors, which correspond to the cluster variables.

We propose to resolve these challenges and establish the first unlocalized cluster realization of the affine basic affine space for all affine Lie types.

2. Precise Setup

Let $\widehat{\mathfrak{g}}$ be an affine Kac-Moody algebra with index set $\widehat{S} = \{0, 1, \ldots, n\}$, where 0 denotes the affine node. Let $\widehat{A} = (a_{st})_{s,t\in\widehat{S}}$ be the affine Cartan matrix, and let $\widehat{\mathfrak{h}}$ be the Cartan subalgebra.
The affine weight lattice is denoted by $\widehat{P} \subset \widehat{\mathfrak{h}}^*$, and the set of dominant affine weights is

$$\widehat{P}_+ = \{\Lambda \in \widehat{P} \mid \langle c_s, \Lambda\rangle \geq 0 \text{ for all } s \in \widehat{S}\},$$

where $\{c_s\}_{s\in\widehat{S}}$ are the simple coroots. The affine fundamental weights are denoted by Λ_i for $i \in \widehat{S}$. Let $\widehat{W}$ be the affine Weyl group generated by the simple reflections r_i for $i \in \widehat{S}$. Let $\widehat{G}$ be the simply-connected affine Kac-Moody group associated with $\widehat{\mathfrak{g}}$, and let $\widehat{U}$ be its maximal unipotent subgroup. The coordinate ring of the affine basic affine space $\widehat{G}/\widehat{U}$ is defined representation-theoretically as the direct sum of the duals of the integrable highest-weight modules:

$$\mathbb{C}[\widehat{G}/\widehat{U}] \cong \bigoplus_{\Lambda\in\widehat{P}_+} V(\Lambda)^*,$$

where $V(\Lambda)$ is the irreducible integrable highest-weight $\widehat{\mathfrak{g}}$-module of highest weight Λ. The algebra structure on $\mathbb{C}[\widehat{G}/\widehat{U}]$ is given by the dual of the highest-weight projection maps:

$$V(\Lambda)^* \otimes V(\Lambda')^* \to V(\Lambda + \Lambda')^*.$$

For each $i \in \widehat{S}$ and $w \in \widehat{W}$, let $v_{w\Lambda_i} = \overline{w} \cdot v_{\Lambda_i}$ be the extremal vector in $V(\Lambda_i)$ of weight $w\Lambda_i$, where v_{Λ_i} is a fixed highest-weight vector. Let $f_{w\Lambda_i} \in V(\Lambda_i)^*_{-w\Lambda_i}$ be the dual extremal vector normalized by $\langle f_{w\Lambda_i}, v_{w\Lambda_i}\rangle = 1$. The **affine generalized minor** $\Delta_{w\Lambda_i} \in \mathbb{C}[\widehat{G}/\widehat{U}]$ is the regular function defined by:

$$\Delta_{w\Lambda_i}(g\widehat{U}) = \langle f_{w\Lambda_i}, g \cdot v_{\Lambda_i}\rangle \quad \text{for } g \in \widehat{G}.$$

To define the cluster structure, we consider the infinite-rank cluster algebra $\mathscr{A}(\boldsymbol{x}, B)$ and upper cluster algebra $\mathscr{U}(\boldsymbol{x}, B)$ associated with an initial seed $(\boldsymbol{x}, B)$ indexed by an infinite set $J = \widehat{S} \sqcup \mathbb{Z}_{>0}$. The initial cluster variables $\boldsymbol{x} = (x_k)_{k\in J}$ are chosen as a family of affine generalized minors $\Delta_{w_k\Lambda_{i_k}}$ associated with a stable limit of reduced words in the affine Weyl group $\widehat{W}$. The exchange matrix $B = (b_{kl})_{k,l\in J}$ is an infinite skew-symmetric matrix such that for each $k \in J$, the set $\{l \in J \mid b_{kl} \neq 0\}$ is finite. This local finiteness property ensures that mutations of the seed $(\boldsymbol{x}, B)$ are well-defined and involve only finitely many variables at each step.
The frozen variables in this setting correspond to the affine fundamental minors $\{\Delta_{\Lambda_i} \mid i \in \widehat{S}\}$, which are never inverted.

3. Main Result
We propose to prove the following primary theorem:

Theorem 1. Let $\widehat{G}$ be a simply-connected affine Kac-Moody group of any affine Dynkin type, and let $\widehat{U}$ be its maximal unipotent subgroup. There exists an infinite-rank seed $(\boldsymbol{x}, B)$ in the fraction field of $\mathbb{C}[\widehat{G}/\widehat{U}]$ such that:

$$\mathbb{C}[\widehat{G}/\widehat{U}] = \mathscr{A}(\boldsymbol{x}, B) = \mathscr{U}(\boldsymbol{x}, B),$$

where the frozen variables are not inverted.
To achieve this, we propose to establish two crucial intermediate results:

Theorem 2 (Affine Generation by Demi-Minors). The coordinate ring $\mathbb{C}[\widehat{G}/\widehat{U}]$ is generated, as a $\mathbb{C}$-algebra, by the affine generalized minors $\{\Delta_{w\Lambda_i} \mid i \in \widehat{S}, w \in \widehat{W}\}$.

*Theorem 3 (Infinite-Rank GLS Criterion). *Let $\mathscr{A}(\boldsymbol{x}, B)$ be an infinite-rank, locally finite cluster algebra. Suppose there exist two disjoint mutable clusters $\boldsymbol{x}'$ and $\boldsymbol{x}''$ (meaning their mutable variables have no common elements) and a subalgebra $\mathscr{O} \subseteq \mathscr{A}(\boldsymbol{x}, B)$ satisfying:*
1. $\mathscr{O}$ is a unique factorization domain (UFD) in the category of direct limits of polynomial rings;
2. $\mathscr{O}$ contains all frozen variables of the seed $(\boldsymbol{x}, B)$ and all mutable variables in the two disjoint clusters $\boldsymbol{x}'$ and $\boldsymbol{x}''$.
*Then we have $\mathscr{O} = \mathscr{A}(\boldsymbol{x}, B) = \mathscr{U}(\boldsymbol{x}, B)$.**

Heuristics & High-Level Arguments
Why is Theorem 2 likely to be true? In the finite-dimensional setting, the generation of the coordinate ring $\mathbb{C}[G/U]$ by fundamental generalized minors $\Delta_{w\varpi_i}$ relies on the fact that the irreducible representation $V(\varpi_i)$ is generated by its extremal vectors under the action of the unipotent subgroups. In the affine setting, the integrable highest-weight module $V(\Lambda_i)$ is similarly generated by its extremal vectors $v_{w\Lambda_i}$ under the action of the affine unipotent subalgebra $\widehat{\mathfrak{u}}^-$. Since $\mathbb{C}[\widehat{G}/\widehat{U}]$ is a direct sum of $V(\Lambda)^*$, and any $V(\Lambda)$ is the highest-weight component of a tensor product of fundamental modules $\bigotimes_{i\in\widehat{S}} V(\Lambda_i)^{\otimes a_i}$, the highest-weight vector v_Λ is the tensor product of the fundamental highest-weight vectors. Consequently, any linear functional $f \in V(\Lambda)^*$ can be written as a limit of linear combinations of tensor products of functionals on the fundamental modules. By dualizing the action of $\widehat{\mathfrak{u}}^-$, we can express any matrix coefficient as a polynomial in the extremal matrix coefficients (the affine generalized minors). This algebraic structure ensures that the affine generalized minors generate the entire coordinate ring.

Why is Theorem 3 likely to be true? The classical Geiss-Leclerc-Schröer criterion uses the Laurent phenomenon to show that any element of the upper cluster algebra $\mathscr{U}$ can be written as a Laurent polynomial in any cluster. If we have two disjoint clusters $\boldsymbol{x}'$ and $\boldsymbol{x}''$, an element $f \in \mathscr{U}$ must be expressed as a Laurent polynomial in both. If the underlying algebra $\mathscr{O}$ is a UFD, the coprime nature of the cluster variables in the disjoint clusters forces the denominators of these Laurent expressions to be trivial, implying that $f \in \mathscr{O}$. In the infinite-rank setting, we can write $\mathbb{C}[\widehat{G}/\widehat{U}]$ as the direct limit of the coordinate rings of Schubert varieties $\widehat{X}_w \subset \widehat{G}/\widehat{B}$. Because these Schubert varieties are normal, Cohen-Macaulay, and have trivial Picard groups, their coordinate rings are UFDs. The direct limit of UFDs under flat transition maps remains a UFD in a suitable sense, allowing us to generalize the algebraic division arguments of the GLS criterion to the infinite-rank case.

Finally, we propose to construct the two disjoint infinite clusters $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ by selecting two distinct "demi-infinite" reduced words for the stable limit of the affine Weyl group $\widehat{W}$. One word will prioritize positive affine reflections, while the other will prioritize negative ones. By analyzing the weights of the corresponding affine generalized minors, we will show that their mutable variables are disjoint, completing the proof of Theorem 1.

Normalized follow-up proposal. For comparison, the following proposal was extracted and normalized from the later citing paper *An introduction to* (G, c)*-bands* (arXiv:2508.19066):

A Discrete Miura Realization of Quantum Affine q-Characters

1. Introduction & Context

We propose to establish a geometric realization of the Frenkel–Reshetikhin q-character homomorphism through a discrete Miura transformation. Our setting combines Steinberg cross-sections in simple algebraic groups, lattice-valued difference equations, and Grothendieck rings of quantum affine representations.

Let G be a simple, simply connected complex algebraic group of simply laced Cartan type A, D, or E, and let c be a Coxeter element of its Weyl group. Steinberg's cross-section

$$A = U(c^{-1})\bar{c}$$

provides canonical representatives of regular conjugacy classes in G. Sequences valued in A naturally arise as coefficients of discrete G-valued difference equations, while a reduced double Bruhat cell $L^{c,e}$ supplies a toric space of Miura-type data. We seek to show that passing from the latter data to the former through discrete gauge normalization reproduces, on coordinate rings, the representation-theoretic passage from a quantum affine module to its q-character.

The q-character homomorphism

$$\chi_q : K_0(\mathcal{C}_{\mathbb{Z}}) \longrightarrow \mathbb{Z}\Big[Y^{\pm 1}_{i,q^{2s+1-\xi_i}} middle | 1 \le i \le r, \in \mathbb{Z}\Big]$$

is an injective refinement of the ordinary character map for a discretized monoidal category $\mathcal{C}_{\mathbb{Z}}$ of finite-dimensional representations of $U_q(\widehat{\mathfrak{g}})$. Its Laurent monomials encode the spectral-parameter data of quantum affine modules. The proposed geometric counterpart is a morphism

$$H : (L^{c,e})^{\mathbb{Z}} \longrightarrow A^{\mathbb{Z}},$$

which should assign to each sequence of Miura variables a uniquely normalized sequence of Steinberg cross-section coefficients.

Our objective is to identify the pullback H^* with χ_q after natural identifications of the relevant coordinate rings. This will provide a uniform discrete q-difference interpretation of q-characters for every simply laced type and every choice of Coxeter element. It will also explain why the difference relations among generalized minors on band schemes coincide with the T-system relations satisfied by Kirillov–Reshetikhin modules.

2. Precise Setup

We fix opposite Borel subgroups

$$B = T \ltimes U, \qquad B^- = T \ltimes U^-,$$

of G, with Weyl group $W = N_G(T)/T$. We choose simple reflections s_i, $1 \le i \le r$, a Coxeter element

$$c = s_{i_1} \cdots s_{i_r} \in W,$$

and a representative $\bar{c} \in N_G(T)$. We write ϖ_i for the fundamental weights and $w_0 \in W$ for the longest Weyl-group element.

We define

$$U(c^{-1}) = U \cap cU^-c^{-1}, \qquad A = U(c^{-1})\bar{c}.$$

The affine variety A is Steinberg's cross-section. For $a \in A$, we use the fundamental generalized minors

$$\theta_i(a) = \Delta_{\varpi_i,\varpi_i}(a), \qquad 1 \le i \le r,$$

as affine coordinates on A.
We also consider the reduced double Bruhat cell

$$L^{c,e} = (U\bar{c}U) \cap B^-.$$

Its coordinate ring is a Laurent polynomial algebra:

$$\mathbb{C}[L^{c,e}] = \mathbb{C}\left[\Delta^{\pm 1}_{\varpi_i,\varpi_i} \;\middle|\; 1 \le i \le r\right].$$

For each i, we let m_i be the least nonnegative integer such that

$$c^{m_i}(\varpi_i) = w_0(\varpi_i).$$

We define a (G, c)-band to be a sequence

$$(g(s))_{s\in\mathbb{Z}} \in G^{\mathbb{Z}}$$

such that

$$g(s)g(s+1)^{-1} \in A \qquad \text{for every } s \in \mathbb{Z}.$$

Writing

$$a(s) = g(s)g(s+1)^{-1},$$

we identify the quotient of the band scheme by its right G-action with $A^{\mathbb{Z}}$. Its coordinate ring is therefore

$$\mathbb{C}[A^{\mathbb{Z}}] = \mathbb{C}[\theta_i^{(s)} \mid 1 \le i \le r, \in \mathbb{Z}],$$

where

$$\theta_i^{(s)}(g) = \Delta_{\varpi_i,\varpi_i}\big(g(s)g(s+1)^{-1}\big).$$

We fix $q \in \mathbb{C}^\times$ not a root of unity. The Coxeter element determines integers ξ_i, and we use the monoidal category $\mathcal{C}_{\mathbb{Z}}$ whose Grothendieck ring is generated by the fundamental modules

$$[L(Y_{i,q^{2s+1-\xi_i}})].$$

We seek an isomorphism

$$\iota : \mathbb{C}[A^{\mathbb{Z}}] \longrightarrow \mathbb{C} \otimes K_0(\mathcal{C}_{\mathbb{Z}})$$

specified by

$$\iota(\theta_i^{(s)}) = [L(Y_{i,q^{2s+1-\xi_i}})].$$

Finally, we identify the Laurent polynomial ring in q-character variables with the coordinate ring of $(L^{c,e})^{\mathbb{Z}}$ by an isomorphism

$$\kappa : \mathbb{C}[Y^{\pm 1}_{i,q^{2s+1\xi_i}}] \longrightarrow \mathbb{C}[(L^{c,e})^{\mathbb{Z}}]$$

defined through

$$\kappa\big(Y_{i,q^{2(s+m_i)+1-\xi_i}}\big) = \Delta_{w_0(\varpi_i),w_0(\varpi_i)} \circ \mathrm{pr}_s\,.$$

3. Main Result
We propose to prove the following geometric realization theorem.
Proposed Theorem. Let G be simple, simply connected, and of simply laced type A, D, or E, and let $c \in W$ be any Coxeter element. Then there exists a discrete Miura transformation

$$H : (L^{c,e})^{\mathbb{Z}} \longrightarrow A^{\mathbb{Z}}$$

such that, under the identifications ι and κ above, its pullback is the Frenkel–Reshetikhin q-character homomorphism:

$$\boxed{\kappa^{-1} \circ H^* \circ \iota^{-1} = \chi_q.}$$

Equivalently, the diagram

$$\begin{array}{ccc} \mathbb{C}[A^{\mathbb{Z}}] & \xrightarrow{\quad H^* \quad} & \mathbb{C}[(L^{c,e})^{\mathbb{Z}}] \\ \Big\downarrow{\scriptstyle \iota} & & \Big\uparrow{\scriptstyle \kappa} \\ \mathbb{C} \otimes K_0(\mathcal{C}_{\mathbb{Z}}) & \xrightarrow[\quad \chi_q \quad]{} & \mathbb{C}[Y_{i,q^{2s+1-\xi_i}}^{\pm 1}] \end{array}$$

commutes.
We will characterize H intrinsically through discrete gauge equivalence. Setting

$$M = U\bar{c}U,$$

we consider the gauge action of $U^{\mathbb{Z}}$ on $M^{\mathbb{Z}}$,

$$(u(s))_{s\in\mathbb{Z}} \cdot (m(s))_{s\in\mathbb{Z}} = \big(u(s)m(s)u(s+1)^{-1}\big)_{s\in\mathbb{Z}}.$$

Our intended construction will normalize a sequence $l \in (L^{c,e})^{\mathbb{Z}} \subset M^{\mathbb{Z}}$ to the unique sequence in $A^{\mathbb{Z}}$ lying in its discrete gauge orbit. Thus we expect

$$H(l) = \text{the unique element of } A^{\mathbb{Z}} \text{ gauge-equivalent to } l.$$

The geometric source of this transformation comes from twisted Birkhoff factorizations. For a band whose components admit decompositions

$$g(s) = u(s)b(s), \qquad u(s) \in U, \quad b(s) \in B^-,$$

we obtain a sequence

$$l(s) = b(s)b(s+1)^{-1} \in L^{c,e}.$$

If

$$a(s) = g(s)g(s+1)^{-1} \in A,$$

then

$$l(s) = u(s)^{-1}a(s)u(s+1).$$

Hence the Miura sequence l and the cross-section sequence a differ precisely by a discrete gauge transformation.
We will use the resulting relation between generalized minors to motivate the q-character identification. Namely, the Laurent coordinate functions on $L^{c,e}$ pull back along the Birkhoff-factorization map as consecutive ratios:

$$\Delta_{w_0(\varpi_i),w_0(\varpi_i)}(l(s)) = \frac{\Delta_{w_0(\varpi_i),w_0(\varpi_i)}(g(s))}{\Delta_{w_0(\varpi_i),w_0(\varpi_i)}(g(s+1))}.$$

These ratios are exactly the geometric expressions that correspond to the variables $Y_{i,a}$ in the q-character Laurent ring.
The proposed theorem will therefore identify two constructions of the same Laurent expansion. On the representation-theoretic side, χ_q expresses a quantum affine module in the variables $Y_{i,a}$. On the geometric side, H^* expresses the invariant functions $\theta_i^{(s)}$ in consecutive generalized-minor ratios. Establishing their equality will realize q-characters as the coordinate pullbacks of a discrete Miura transformation.

Table 6: Evaluation scores for the chosen generated and normalized follow-up proposals

Evaluation axis	Generated	Follow-up
Soundness criteria		
Well-defined terms	2	3
Sound assumptions	2	3
Supporting evidence	2	3
Overall	**2**	**3**
Novelty criteria		
Conceptual depth	3	3
Originality	3	1
Overall	**3**	**2**
Feasibility criteria		
Example feasibility	1	2
Obstruction feasibility	2	3
Overall	**1.5**	**2.5**