

Which Mathematical Context Supports Target-Aligned Conjecture Generation? A Paired Ablation Study

Austin LaHue¹, Michael Burr¹, Luis David Garcia Puente²,
Vinita Gangaram Jansari¹, Benjamin Nye², Carlos Toxtli Hernandez¹

¹Clemson University, ²Colorado College
jlahue@clemson.edu

Abstract

Generating mathematical conjectures from natural-language context is challenging because theorem-relevant information is distributed across multiple forms of mathematical context, including definitions, objects, assumptions, and constructions. We study which of these contextual signals help a model recover the specific claim supported by the surrounding setup, using 61 main theorems from 12 papers across four mathematical domains. For each theorem, we redact the result and represent the remaining setup using six context components. A paired ablation design compares full context, six leave-one-out conditions, and six single-component conditions, yielding 793 generated conjectures. Context components were largely complementary: removing one component usually had little effect, whereas retaining only one substantially reduced expected-theorem alignment. Local Mathematical Objects provided the strongest standalone signal, while Structural Constraints and Conditions produced the largest alignment decrease when removed. These findings suggest that target-aligned conjecture generation depends on combining information that identifies relevant objects with information that constrains admissible claims.

1 Introduction

Mathematical conjectures are formed within a local discourse of definitions, objects, assumptions, examples, and established results. Automated conjecturing has traditionally searched structured symbolic, programmatic, or formally verified spaces (Raayoni et al., 2021; Romera-Paredes et al., 2024; Lin et al., 2024), while recent mathematical language systems show that supplied definitions, theorems, and repository context can support proof generation and formal reasoning (Welleck et al., 2022; Hu et al., 2025). Less is known about which parts of natural-language mathematical context help a model formulate the particular claim that a paper goes on to prove.

We study this question through *target-aligned conjecture generation*. Given a redacted representation of the setup preceding a theorem, a model generates a formal conjecture without access to the theorem statement or proof. The task is narrower than open-ended mathematical discovery: success concerns whether the supplied context supports a mathematically substantive statement aligned with an existing hidden theorem. This controlled setting separates the ability to produce a plausible conjecture from the ability to recover the specific claim supported by the source context.

We conducted a paired ablation study over 61 main theorems from 12 papers spanning Commutative Algebra, Algebraic Geometry, Combinatorics, and Statistics Theory¹. The redacted context is organized into six components and evaluated under full-context, leave-one-out, and single-component conditions. Results indicate that these components are largely complementary: removing one component generally had little effect, whereas retaining one alone substantially reduced alignment. Local Mathematical Objects supplied the strongest standalone signal, while removing Structural Constraints produced the largest alignment loss, suggesting distinct roles in identifying what a conjecture concerns and disambiguating what can be claimed. Together, these findings provide a component-level account of how theorem-local natural language supports target-aligned conjecture generation.

2 Related Work

2.1 Automated Conjecturing

Prior work on automated conjecturing includes heuristic discovery over graph invariants, theory exploration, and search within restricted formula spaces (Fajtlowicz, 1988; Larson and Van Cleemput, 2016; Colton et al., 2000; McCasland et al.,

¹Supplementary materials are available at <https://github.com/alahue/conjecture-ablation>.

2017; Raayoni et al., 2021). More recent systems combine learned guidance with executable search or formal verification (Romera-Paredes et al., 2024; Lin et al., 2024; Dong and Ma, 2025). These approaches largely generate candidates in structured, programmatic, or formal spaces. We instead study conjecture generation from the mixed natural and symbolic language preceding theorems in research papers, asking which semantic components of that local discourse support target alignment.

2.2 Mathematical Context

Mathematical text interleaves prose, formulas, local notation, and document-level dependencies. Work on premise selection and identifier grounding shows that definitions, theorems, lemmas, propositions, formula structures, and co-references help identify the information relevant to a mathematical task (Ferreira and Freitas, 2020, 2021; Asakura and Miyao, 2024). Context-grounded proof systems use such local and repository-level information to support proof generation (Welleck et al., 2022; Yang et al., 2023; Hu et al., 2025), while autoformalization and semantic-alignment work connects informal mathematical language with formal statements (Wu et al., 2022; Lu et al., 2025). Our study differs by fixing a hidden theorem target and ablating semantically defined components of its preceding setup. Because long-context use and mathematical reasoning can vary with information position and surface form (Liu et al., 2024; Williamson et al., 2025), we interpret the observed effects as properties of the constructed prompts rather than of mathematical context in general.

2.3 Ablation and Leakage

Our design parallels rationale-faithfulness evaluation: removing evidence measures its conditional contribution, whereas retaining it alone measures sufficiency (DeYoung et al., 2020). Accordingly, leave-one-out conditions estimate a component’s contribution within rich context, and single-component conditions estimate standalone informativeness; the rankings may differ when information is complementary or redundant. Ablations can also create out-of-distribution inputs or preserve cues that directly reveal the target (Hase et al., 2020; Hsia et al., 2024). We therefore avoid causal interpretations and separately audit redacted seeds and model-generated generalization axes for theorem-shaped leakage.

2.4 Evaluation and Contamination

Generated conjectures must be assessed both for correspondence to the source theorem and for mathematical quality. Rubric-guided LLM judges can approximate human evaluation (Liu et al., 2023; Zheng et al., 2023), but scores remain sensitive to judge choice, repeated runs, and irrelevant persuasive language (Halder and Hockenmaier, 2025; Hwang et al., 2025). We report repeated-judge, cross-model, and expert agreement and treat scores as prompt- and judge-dependent estimates rather than verified correctness. Finally, reconstruction may reflect pretraining exposure. Contamination work distinguishes exposure from exploitation and advocates perturbation-based controls (Magar and Schwartz, 2022; Qian et al., 2024; Gulati et al., 2025). Our publication-date holdout is exploratory, not proof of training-set inclusion or exclusion.

3 Method

3.1 Paper and Theorem Selection

We sampled 12 arXiv papers from four mathematical subject domains, with three papers selected from each:

- math.AC (Commutative Algebra) (Grimpen et al., 2026; Clarke and Kosta, 2024; Friedman et al., 2025);
- math.AG (Algebraic Geometry) (Jeronimo et al., 2024; Draisma et al., 2013; Salizzoni et al., 2025);
- math.CO (Combinatorics) (Sarkar et al., 2016; Le and Römer, 2023; Hollering et al., 2024);
- math.ST (Statistics Theory) (Brysiewicz and Kim, 2026; Alexandr et al., 2022; Duarte and Solus, 2021).

Within each domain, papers were additionally required to have at least one Mathematics Subject Classification (MSC) code in one of the following families: 13Pxx (Computational aspects and applications of commutative rings), 14Qxx (Computational aspects in algebraic geometry), 68Wxx (Algorithms in computer science), or 62Rxx (Statistics on algebraic and topological structures). Among the selected papers classified within 62Rxx, all were assigned to 62R01 (Algebraic statistics).

For each paper, the theorem-selection stage of our pipeline classified all theorem-like environments and identified the paper’s *main proved theorems*. A result qualified if it was presented as a central contribution of the paper, proved within

the paper, and not merely a quoted result, supporting lemma or proposition, corollary, example, or background theorem. This process identified 61 theorem targets across the 12 papers. Individual papers contributed between 3 and 7 targets. The domain totals were 14 for math.ST, 17 for math.AC, 16 for math.AG, and 14 for math.CO.

Four of the selected papers, comprising 21 theorem targets, were posted after May 2025, a knowledge-cutoff date reported for DeepSeek-V4-Pro by third-party model documentation (Ng, 2026). We treat these papers as a temporal-holdout cohort and compare them with the 8 earlier papers.

A human reviewer checked all 61 selected theorem targets against the source papers and confirmed that each was an eligible main proved theorem. We did not assess whether the pipeline identified every eligible theorem in the selected papers. One statement was incompletely extracted, while minor transcription issues included lost $\LaTeX$ formatting in two papers and added definitions in two papers.

3.2 Conjecture-Generation Pipeline

Figure 1 summarizes the end-to-end pipeline. We used Codex (OpenAI, 2025b) as the orchestration harness, configured to call DeepSeek-V4-Pro (DeepSeek-AI, 2026) through a vLLM-served endpoint (Kwon et al., 2023). DeepSeek-V4-Pro performed theorem identification, object-map construction, redaction, and conjecture generation under each experimental condition. The generated conjecture and corresponding source theorem were then provided to DeepEval (Ip and Vongthongsri, 2026), with Gemma-4-31B (Gemma Team, 2026) serving as the judge, to produce theorem-conditioned conjecture-quality and expected-theorem-alignment scores.

For each theorem target, the generation pipeline performed the following steps:

1. Construct a theorem-specific object map containing the source theorem, relevant definitions and notation, mathematical objects, assumptions, supporting results and constructions, and proof dependencies.
2. Convert the object map into a redacted seed by removing information that could reveal the target theorem. The redacted content included the theorem statement and conclusion, theorem identifiers, proof strategy, dependency chain, final classifications, and exact result-specific formulas or equivalences.

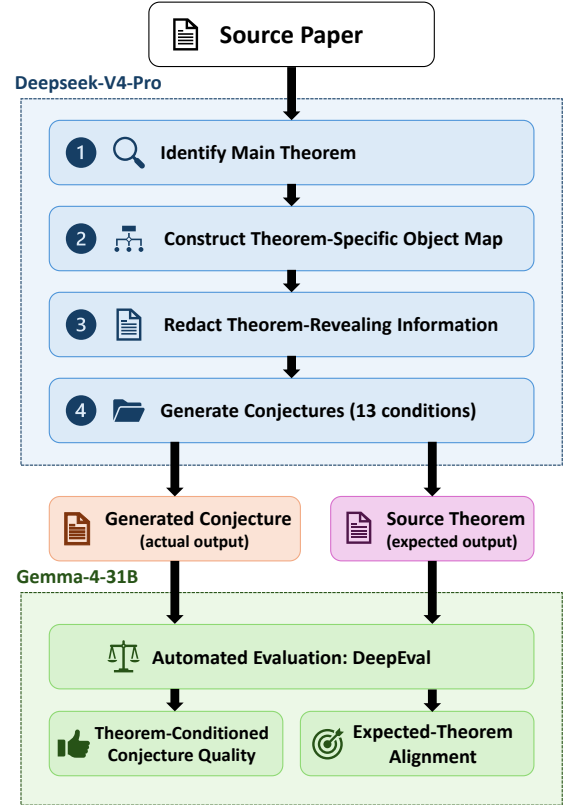


Figure 1: Overview of the conjecture-generation and evaluation pipeline.

3. Organize the remaining setup information into six mathematical-context components.
4. Generate one conjecture from the selected components for each experimental condition.

The conjecture-generation model was provided only with the redacted seed for the corresponding condition. The original paper, theorem map, unredacted object map, source theorem, proof strategy, and dependency chain were not included in the model input, and web search was disabled. It generated exactly one formal conjecture per theorem and condition. The substantive classification, redaction, and generation instructions are summarized in Appendix E.

Table 1 shows a source theorem identified by the pipeline and the conjecture generated from its redacted mathematical-context components.

3.3 Mathematical-Context Components

Each redacted seed represented the theorem’s preceding context using six components:

1. **Ambient Setting:** the broad mathematical setting and background assumptions, without

identifying the target theorem;

2. **Core Definitions and Notation:** definitions and notation required to understand the local problem;
3. **Local Mathematical Objects:** structures, quantities, maps, operators, relations, functions, families, or other mathematical entities that define the local problem setting;
4. **Structural Constraints and Conditions:** hypotheses, restrictions, and validity conditions defining the local problem setting;
5. **Auxiliary Objects and Constructions:** supporting examples, intermediate objects, derived constructions, or families used to study, illustrate, or relate the local mathematical objects;
6. **Potential Generalization Axes:** model-generated directions for forming conjectures, such as varying parameters, weakening assumptions, changing the ambient object, or examining exceptional cases.

The first five components were summarized from the theorem-specific source material after redaction. The Potential Generalization Axes were synthesized by DeepSeek-V4-Pro from the redacted mathematical setup.

All 61 redacted theorem seeds underwent a leakage audit (Appendix A). Fifty-eight seeds (95.1%) were flagged for potential inferential leakage and reviewed, but all were retained after mitigation; no audit summary reported an explicit hidden theorem statement or final conclusion remaining in a seed.

Source theorem

Let $U \subseteq V \subseteq R^d$ be submodules and $\prec$ a fixed monomial order on R^d . Then V has a unique reduced Gröbner basis relative to U .

Generated conjecture

Let $\mathbb{k}$ be a field, let $R = \mathbb{k}[X_1, \dots, X_n]$, and let $U \subseteq V \subseteq R^d$ be nested submodules of the free module R^d . Fix a monomial order $\prec$ on $\text{Mon}(R^d)$. Then there exists a reduced relative Gröbner basis H of V with respect to U .

Moreover, this reduced relative Gröbner basis is unique: if H and H' are two reduced relative Gröbner bases of V relative to U , with respect to the same monomial order $\prec$, then

$$H = H'.$$

Table 1: Example of a source theorem (Grimpen et al., 2026) and the corresponding generated conjecture.

Because Potential Generalization Axes were synthesized by the same model used for conjecture generation, we audited this component separately for theorem-shaped information (Appendix B). Of the 61 seeds, 12 (19.7%) were classified as direct theorem-shaped, 24 (39.3%) as strongly directional, and 25 (41.0%) as low or contextual. Accordingly, results involving this component may reflect theorem-directed cues in addition to genuine support for conjecture generation and are therefore interpreted cautiously.

3.4 Ablation Design

For each of the 61 theorem targets, we generated conjectures under 13 context conditions:

- one **full-context** condition containing all six components;
- six **leave-one-out** conditions, each excluding one component while retaining the other five;
- six **single-component** conditions, each retaining one component while excluding the other five.

This yielded $61 \times 13 = 793$ generated conjectures. The design is paired: all 13 conditions for a theorem were derived from the same theorem-specific object map and evaluated against the same source theorem. Leave-one-out conditions estimate a component’s contribution in the presence of otherwise rich context, whereas single-component conditions estimate its standalone informativeness.

Because redaction cannot prevent recovery of a theorem from model pretraining, we conducted an exploratory temporal-holdout analysis. We compared full-context performance and the mean effects of the leave-one-out and single-component conditions between the four post-cutoff papers and the eight earlier papers. Cohort comparisons used paper-level summaries so that each paper received equal weight. Because DeepSeek has not published an official knowledge cutoff, publication date does not establish actual training-set inclusion, and the cohorts are not balanced by domain, we treat this analysis as exploratory.

3.5 Automated Conjecture Evaluation

We evaluated conjectures using DeepEval with Gemma-4-31B as the primary judge and a temperature of zero, following exploratory comparisons among candidate judges. DeepEval received

the generated statement as the *actual output* and the original source theorem as the *expected output*. Scores were assigned on a scale from 0 to 1 in increments of 0.1, with higher values indicating stronger performance.

We used two metrics:

- **Expected-Theorem Alignment:** implemented using DeepEval’s documented correctness criterion: “Determine whether the actual output is factually correct based on the expected output.” (Confident AI, 2023) Because this measure compares a generated conjecture with a target theorem rather than independently verifying the conjecture, we interpret it as LLM-estimated theorem-alignment rather than verified mathematical correctness.
- **Theorem-Conditioned Conjecture Quality:** a custom, target-aware rubric assessing whether the output contains a clear and substantive mathematical claim, its simple-case plausibility, precision, potential usefulness, and relevance to the source theorem. Unlike expected-theorem alignment, this metric can reward a meaningful specialization, close variant, or plausible generalization even when it does not accurately recover the target result. Because the expected theorem is available to the judge, this metric does not measure source-independent novelty or intrinsic conjecture quality. The complete rubric is provided in Appendix D.

Both metrics are conditioned on the expected theorem, but operationalize different outcomes. Expected-theorem alignment is reconstruction-oriented, whereas theorem-conditioned conjecture quality permits mathematically meaningful departures from the target theorem.

3.6 Expert Validation

We evaluated agreement between the automated theorem-conditioned quality judge and an expert mathematician using an independent validation set of 15 conjectures. These conjectures were generated separately using o3-mini (OpenAI, 2025c) and were not sampled from the ablation-study outputs or assigned to experimental conditions. The expert applied the same theorem-conditioned quality rubric used by the automated judge and was blinded to the LLM-generated scores.

Expert and Gemma theorem-conditioned quality scores were moderately positively correlated, $r(13) = .60$, $p = .017$, 95% CI [.13, .85]. Absolute agreement was also moderate using a two-way mixed-effects, absolute-agreement, single-measure intraclass correlation, $ICC(A, 1) = .54$, with a bootstrap 95% CI of [.09, .77]. The mean absolute error was 0.27, and Gemma assigned scores that were, on average, 0.16 points higher than the expert. We therefore interpret this validation as preliminary evidence of evaluator agreement rather than definitive validation, given the single expert, small sample, and use of an independently generated validation set.

The expert did not evaluate expected-theorem alignment because doing so would require substantial proof development or counterexample search. The Expert theorem-conditioned quality scores were positively associated with the automated expected-theorem-alignment scores, $r(13) = .58$, $p = .022$, 95% CI [.10, .84]; however, because these variables measure different constructs, we treat this as an exploratory association rather than direct validation of the alignment metric.

3.7 Judge Reliability and Cross-Model Agreement

To assess evaluator reliability, we rescored the full-context and six leave-one-out outputs from four papers, with one paper selected from each domain. We first repeated the evaluation using Gemma-4-31B with the same prompts and settings. We then evaluated the same outputs using GPT-5-mini (OpenAI, 2025a) as an alternative judge. For each metric, we assessed score association using Pearson’s r and absolute agreement using $ICC(A, 1)$.

We also piloted GLM-5.1-FP8 (Z.ai, 2026; GLM-5 Team, 2026) as an expected-theorem-alignment judge. Its lower agreement with Gemma appeared to reflect a stricter treatment of claims absent from the expected theorem. We therefore report it as an exploratory diagnostic of judge and prompt sensitivity (Appendix C).

3.8 Statistical Analysis

Because each theorem target was evaluated under every context condition, analyses were paired at the theorem level. For leave-one-out conditions, we report mean differences from the full-context baseline, paired effect sizes (d_z), and 95% confidence intervals for the paired differences. Overall differences among the full-context and six leave-one-out

conditions were assessed for each evaluation metric using Friedman tests, followed by Wilcoxon signed-rank comparisons with the full-context baseline and Holm correction across the six comparisons.

Because theorem targets from the same paper share paper-level context, we treated them as clustered rather than independent observations. To prevent papers with more targets from receiving greater weight, we repeated the primary analyses using paper-level mean scores, with each paper weighted equally. Domain comparisons were computed from these summaries and treated as exploratory because each domain contained only three papers. For each metric, differences in component-removal effects across domains were assessed using Kruskal–Wallis tests with Holm correction across the six components. Single-component conditions were analyzed descriptively to estimate standalone informativeness.

4 Results

4.1 Context Components Are Largely Complementary

Under the full-context condition, mean theorem-conditioned conjecture quality was 0.931, and mean expected-theorem alignment was 0.607. Figure 2 reports paired theorem-level differences from these baselines. Complete condition-level results are reported in Appendix F.

Leave-one-out ablations produced only small changes in either metric. The omnibus tests were not significant for theorem-conditioned quality, $\chi^2(6) = 5.28$, $p = .509$, or expected-theorem alignment, $\chi^2(6) = 9.75$, $p = .135$. Removing Structural Constraints and Conditions produced the largest alignment decrease, $\Delta = -0.087$, 95% CI $[-0.154, -0.020]$, but the comparison did not remain significant after Holm correction, $p_{\text{Holm}} = .055$. No other leave-one-out effect was clearly distinguishable from zero.

Equal-weight paper-level analyses yielded the same qualitative pattern. The omnibus tests remained nonsignificant, and removing Structural Constraints again produced the largest alignment decrease, $\Delta = -0.085$, $p_{\text{Holm}} = .105$.

The single-component conditions revealed a distinction between standalone and complementary information. Local Mathematical Objects provided the strongest standalone signal, with an alignment mean of 0.457. Structural Constraints were less informative alone, with an alignment mean of 0.348,

despite producing the largest decrease when removed from full context. Thus, Local Mathematical Objects supplied the most directly usable information, whereas Structural Constraints contributed by disambiguating otherwise rich context.

4.2 Component Importance Varies by Domain

Domain-level analyses were exploratory because each domain contained only three papers. Full-context performance was broadly similar across domains, but the component associated with the largest alignment loss varied (Table 2). Structural Constraints were most consequential in the two algebraic domains, whereas Definitions and Notation were most consequential in math.CO, and Local Mathematical Objects were most consequential in math.ST. The complete domain-level results are reported in Appendix G.

Potential Generalization Axes showed the clearest qualitative domain difference. Removing them reduced alignment in math.AC and math.AG, but increased alignment in math.CO and math.ST. The removal effect did not differ significantly across domains after Holm correction, $p_{\text{Holm}} = .553$.

4.3 Held-Out Papers Show Suggestive Sensitivity to Context Reduction

Full-context performance was similar for the post-cutoff and earlier-paper cohorts: expected-theorem alignment was 0.630 and 0.619, respectively, while theorem-conditioned quality was 0.929 and 0.942. However, the post-cutoff papers were more sensitive to incomplete context. Their mean leave-one-out alignment effect was -0.083 , compared with approximately zero for the earlier papers, and their mean single-component effect was -0.346 , compared with -0.218 . Theorem-conditioned quality effects were nearly identical between cohorts. The paper-level cohort differences were not significant for either the leave-one-out ($p = .214$) or single-component ($p = .087$) analyses. These descriptive patterns are consistent with greater reliance on the

Domain	Full	Largest loss	Δ	Best standalone
math.AC	.611	Constraints	-.162	Local objects
math.AG	.571	Constraints	-.121	Definitions
math.CO	.644	Definitions	-.086	Constraints
math.ST	.664	Local objects	-.096	Definitions

Table 2: Exploratory domain-level expected-theorem alignment. Full is the full-context baseline. Each domain contains three papers.

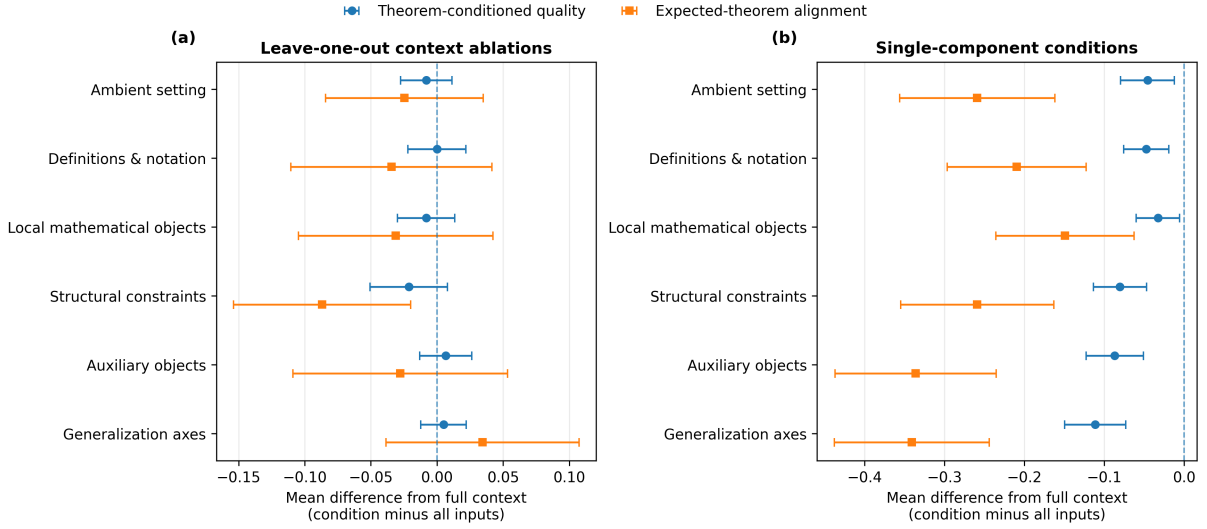


Figure 2: Paired theorem-level differences from the full-context condition with 95% confidence intervals. Panel (a) removes one context component, whereas panel (b) retains only the named component. Negative values indicate lower scores than under full context.

supplied context for temporally held-out papers, but do not establish whether any individual paper was present in the model’s training data.

4.4 Judge Agreement Varies by Metric

Repeated Gemma evaluation showed high reliability for both metrics (Table 3). Cross-model agreement with GPT-5-mini was lower: alignment scores showed moderate absolute agreement, whereas absolute quality scores were sensitive to judge choice.

5 Discussion

This study examined which forms of the theorem-local mathematical context support conjectures aligned with a hidden target theorem. Across 61 theorem targets and 13 paired conditions, the results indicate that context components are largely complementary rather than individually indispensable. Removing one component from an otherwise complete context generally produced only small changes, whereas retaining a single component substantially reduced expected-theorem alignment.

Comparison	Metric	r	ICC
Gemma repeat	Alignment	.921	.921
	Quality	.840	.841
GPT-5-mini	Alignment	.784	.731
	Quality	.568	.212

Table 3: Repeated-judge reliability and cross-model agreement.

Thus, the small leave-one-out effects do not imply that context is unimportant. Instead, multiple components appear to provide overlapping information and partially compensate for one another.

The contrast between the two evaluation metrics further clarifies this result. Full context produced high theorem-conditioned conjecture quality but only moderate expected-theorem alignment. Even under single-component conditions, the model often generated plausible and relevant mathematical statements despite failing to recover the particular result proved in the source paper. Limited context may therefore be sufficient for producing a mathematically substantive continuation, while richer context is needed for target-aligned generation. These are related but distinct capabilities.

5.1 Objects Propose; Constraints Disambiguate

The clearest functional distinction emerged between Local Mathematical Objects and Structural Constraints and Conditions. Local Mathematical Objects provided the strongest standalone alignment signal, suggesting that theorem-relevant objects, maps, quantities, and families help define the broad mathematical subject and were often sufficient for the model to produce a related claim.

Structural Constraints played a different role. Although less informative in isolation, their removal from the full context produced the largest decrease in expected-theorem alignment. This pattern suggests that constraints are primarily disambiguating:

mathematical objects define a space of possible claims, while hypotheses, restrictions, and validity conditions help determine which relation among those objects is intended. Objects may therefore support candidate formation, whereas constraints help select among otherwise plausible candidates.

This distinction also shows why component importance cannot be represented by a single ranking. Standalone performance reflects how directly a component supports generation by itself, while leave-one-out effects reflect its contribution when combined with the remaining context. For target-aligned conjecture generation, direct informativeness and conditional contribution appear to be separate properties.

The exploratory domain results suggest that these functions may vary across mathematical fields. Structural Constraints produced the largest alignment losses in Commutative Algebra and Algebraic Geometry, while Definitions and Notation were most consequential in Combinatorics and Local Mathematical Objects in Statistics Theory. Because each domain contained only three papers, these patterns are not conclusive, but they motivate domain-adaptive approaches to selecting and organizing mathematical context.

The Potential Generalization Axes showed less consistent effects. Their removal reduced alignment in the algebraic domains but increased it in Combinatorics and Statistics Theory, and they were weak as a standalone component. Generalization directions may sometimes focus generation toward the target, but they may also broaden the candidate space. Because these axes were model-generated rather than directly extracted from the source, they are better treated as a conjecturing scaffold than as equivalent to the five source-derived components.

5.2 Context Dependence and Evaluation

Earlier and post-cutoff papers achieved similar full-context performance, but the post-cutoff cohort showed larger alignment losses when context was reduced. This pattern is consistent with the model relying more heavily on the supplied mathematical setup when the target material is less familiar from training. It does not establish whether any paper was included in the training data, and the cohort differences were not statistically conclusive. The temporal analysis therefore provides evidence about context sensitivity rather than contamination itself. Controlled perturbations, such as systematic symbol renaming or paraphrasing, would provide

stronger tests of whether generation depends on supplied mathematical meaning, familiar lexical patterns, or prior exposure.

Evaluator agreement also differed by metric. Expected-theorem alignment showed stronger repeated and cross-model agreement than theorem-conditioned conjecture quality. Alignment asks a constrained question about correspondence to a supplied theorem, whereas quality is a broader judgment of clarity, simple-case plausibility, precision, potential usefulness, and relevance to the source theorem. Expected-theorem alignment therefore provides a more stable basis for comparing ablation conditions, while theorem-conditioned quality indicates whether low-alignment outputs remain mathematically substantive.

Neither metric independently establishes mathematical truth or novelty. Future evaluation could score the existing quality dimensions separately rather than combining them into one holistic judgment. It could also classify each output as equivalent to the target theorem, a specialization, a generalization, or a different but related result. Proof attempts, counterexample search, and source-blind expert review would further distinguish target recovery from mathematical validity.

Overall, the findings suggest that conjecture-generation systems may benefit from organizing context according to its mathematical role. Local Mathematical Objects provide an efficient basis for identifying what to conjecture about, while Structural Constraints, definitions, and ambient information narrow the admissible claims. Because component effects varied across domains, adaptive context assembly may be more effective than a fixed minimal prompt. Target alignment appears to emerge from the interaction between context that identifies relevant objects and context that constrains admissible claims.

6 Conclusion

We presented a paired ablation study of theorem-local context for target-aligned conjecture generation. Across 61 theorem targets, Local Mathematical Objects were most informative in isolation, while Structural Constraints contributed most within richer context. Future work should test whether these patterns generalize across additional models, domains, context decompositions, and source-blind evaluations of validity and novelty.

Limitations

Both evaluation metrics are conditioned on the expected theorem. The theorem-conditioned quality score therefore does not measure source-independent novelty, usefulness, or mathematical validity, while expected-theorem alignment remains sensitive to judge and prompt interpretation. This sensitivity was reflected in the stricter behavior of GLM-5.1-FP8 relative to Gemma-4-31B and GPT-5-mini. Expert validation was also limited to one mathematician and 15 conjectures generated separately by o3-mini, so agreement may not generalize to the ablation-study outputs. The expert did not independently validate expected-theorem alignment, as doing so would require proof development or counterexample search. Accordingly, the reported scores should be treated as target-conditioned estimates rather than definitive assessments of conjecture quality or correctness.

The ablations do not isolate the causal importance of each component. Removing or retaining components changes prompt length, coherence, redundancy, and resemblance to natural mathematical exposition. Single-component prompts may be especially unnatural, while leave-one-out prompts may allow overlapping information to compensate for the removed component. The study is also limited to 61 theorem targets from 12 papers, four computationally oriented domains, one generation model, and one conjecture per theorem-condition pair, so within-condition generation variability was not estimated. Domain comparisons are therefore exploratory, and the results may depend on the selected papers, prompts, and component-construction procedure. The LLM-based selection pipeline was validated only for the targets it selected; because recall was not assessed, it may have omitted other eligible main theorems.

The temporal-holdout analysis relies on a third-party knowledge-cutoff estimate because DeepSeek has not disclosed an official cutoff for DeepSeek-V4-Pro. The publication date cannot establish whether a paper or related material appeared in training, and the temporal cohorts also differ in domain composition. These results should therefore be interpreted as evidence of differing context sensitivity rather than direct evidence of training-data contamination.

Potential Generalization Axes present an additional leakage risk because they were synthesized from theorem-specific material by the same model

used for generation. The audit classified 19.7% of seeds as directly theorem-shaped and 39.3% as strongly directional. Effects involving this component may therefore reflect theorem-directed cues as well as support for generalization. More broadly, the six components represent one operationalization of mathematical context; other reasonable decompositions may produce different results.

Acknowledgments

This work was supported by the National Science Foundation under Award No. 2432704. This research used in part resources on the Palmetto 2 cluster at Clemson University under National Science Foundation awards MRI 1228312, II NEW 1405767, MRI 1725573, and MRI 2018069. The views expressed in this article do not necessarily represent the views of NSF or the United States government.

The authors used ChatGPT to assist with refining the language and organization of the manuscript and to support discussion of statistical analyses and presentation of results. All generated suggestions were reviewed, verified, and revised by the authors, who take full responsibility for the content of the paper. Generative AI models used as part of the experimental methodology are described separately in Section 3.2.

References

- Yulia Alexandr, Eliana Duarte, and Julian Vill. 2022. [Decomposable context-specific models](#). *Preprint*, arXiv:2210.11521.
- Takuto Asakura and Yusuke Miyao. 2024. [What is needed for intra-document disambiguation of math identifiers?](#) In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17500–17512, Torino, Italia. ELRA and ICCL.
- Taylor Brysiewicz and Juhee Kim. 2026. [How to recover a permutation group amidst errors](#). *Preprint*, arXiv:2602.23505.
- Oliver Clarke and Dimitra Kosta. 2024. [Distance reducing markov bases](#). *Preprint*, arXiv:2406.17730.
- Simon Colton, Alan Bundy, and Toby Walsh. 2000. [On the notion of interestingness in automated mathematical discovery](#). *International Journal of Human-Computer Studies*, 53(3):351–375.
- Confident AI. 2023. [Introduction to LLM evaluation metrics](#). DeepEval documentation. Accessed: 2026-08-05.

- DeepSeek-AI. 2026. [DeepSeek-V4: Towards highly efficient million-token context intelligence](#). *Preprint*, arXiv:2606.19348.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.
- Kefan Dong and Tengyu Ma. 2025. [STP: Self-play LLM theorem provers with iterative conjecturing and proving](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 14114–14136. PMLR.
- Jan Draisma, Emil Horobet, Giorgio Ottaviani, Bernd Sturmfels, and Rekha R. Thomas. 2013. [The euclidean distance degree of an algebraic variety](#). *Preprint*, arXiv:1309.0049.
- Eliana Duarte and Liam Solus. 2021. [Representation of context-specific causal models with observational and interventional data](#). *Preprint*, arXiv:2101.09271.
- Siemion Fajtlowicz. 1988. [On conjectures of Graffiti](#). In *Graph Theory and Its Applications: East and West*, volume 38 of *Annals of Discrete Mathematics*, pages 113–118. Elsevier.
- Deborah Ferreira and André Freitas. 2020. [Premise selection in natural language mathematical texts](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7365–7374, Online. Association for Computational Linguistics.
- Deborah Ferreira and André Freitas. 2021. [STAR: Cross-modal statement representation for selecting relevant mathematical premises](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3234–3243, Online. Association for Computational Linguistics.
- Hannah Friedman, Bernd Sturmfels, and Maximilian Wiesmann. 2025. [Squared linear models](#). *Preprint*, arXiv:2505.19351.
- Gemma Team. 2026. [Gemma 4 technical report](#). *Preprint*, arXiv:2607.02770.
- GLM-5 Team. 2026. [GLM-5: From vibe coding to agentic engineering](#). *Preprint*, arXiv:2602.15763.
- Fritz Grimpen, Matthias Orth, and Anastasios Stefanou. 2026. [Relative Gröbner bases of modules and applications in persistence theory](#). *Preprint*, arXiv:2603.27149.
- Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Frønsdal, Bruno De Moraes Dumont, and Sanmi Koyejo. 2025. [Putnam-AXIOM: A functional and static benchmark for measuring higher level mathematical reasoning in LLMs](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 20723–20747. PMLR.
- Rajarshi Haldar and Julia Hockenmaier. 2025. [Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24986–25004, Suzhou, China. Association for Computational Linguistics.
- Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. [Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367, Online. Association for Computational Linguistics.
- Benjamin Hollering, Joseph Johnson, and Liam Solus. 2024. [Hyperplane representations of interventional characteristic imset polytopes](#). *Preprint*, arXiv:2404.18500.
- Jennifer Hsia, Danish Pruthi, Aarti Singh, and Zachary C. Lipton. 2024. [Goodhart’s law applies to NLP’s explanation benchmarks](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1322–1335, St. Julian’s, Malta. Association for Computational Linguistics.
- Jiewen Hu, Thomas Zhu, and Sean Welleck. 2025. [miniCTX: Neural theorem proving with \(long-\)contexts](#). In *The Thirteenth International Conference on Learning Representations*.
- Yerin Hwang, Dongryeol Lee, Taegwan Kang, Yongil Kim, and Kyomin Jung. 2025. [Can you trick the grader? adversarial persuasion of LLM judges](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14632–14651, Suzhou, China. Association for Computational Linguistics.
- Jeffrey Ip and Kritin Vongthongsri. 2026. [DeepEval: The open-source LLM evaluation framework](#). Software, version 4.1.5, released 2026-07-16.
- Gabriela Jeronimo, Leonardo Lanciano, and Pablo Solernó. 2024. [On the geometric degree of the tangent bundle of a smooth algebraic variety](#). *Preprint*, arXiv:2403.10661.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626. Association for Computing Machinery.
- Craig E. Larson and Nicolas Van Cleemput. 2016. [Automated conjecturing I: Fajtlowicz’s dalmatian heuristic revisited](#). *Artificial Intelligence*, 231:17–38.

- Dinh Van Le and Tim Römer. 2023. [Equivariant lattice bases](#). *Preprint*, arXiv:2309.07246.
- Xiaohan Lin, Qingxing Cao, Yinya Huang, Zhicheng Yang, Zhengying Liu, Zhenguo Li, and Xiaodan Liang. 2024. [ATG: Benchmarking automated theorem generation for generative language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4465–4480, Mexico City, Mexico. Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Jianqiao Lu, Yingjia Wan, Yinya Huang, Jing Xiong, Zhengying Liu, and Zhijiang Guo. 2025. [FormAlign: Automated alignment evaluation for auto-formalization](#). In *The Thirteenth International Conference on Learning Representations*.
- Inbal Magar and Roy Schwartz. 2022. [Data contamination: From memorization to exploitation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.
- Roy L. McCasland, Alan Bundy, and Patrick F. Smith. 2017. [MATHsAiD: Automated mathematical theory exploration](#). *Applied Intelligence*, 47(3):585–606.
- Ann Ng. 2026. [How to use DeepSeek V4 Pro from DeepSeek with API key on TypingMind](#). Accessed: 2026-08-03.
- OpenAI. 2025a. [GPT-5 system card](#). Accessed: 2026-08-05.
- OpenAI. 2025b. [Introducing Codex](#). Accessed: 2026-08-04.
- OpenAI. 2025c. [OpenAI o3-mini](#). Accessed: 2026-08-06.
- Kun Qian, Shunji Wan, Claudia Tang, Youzhi Wang, Xuanming Zhang, Maximillian Chen, and Zhou Yu. 2024. [VarBench: Robust language model benchmarking through dynamic variable perturbation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16131–16161, Miami, Florida, USA. Association for Computational Linguistics.
- Gal Raayoni, Shahar Gottlieb, Yahel Manor, George Pisha, Yoav Harris, Uri Mendlovic, Doron Haviv, Yaron Hadad, and Ido Kaminer. 2021. [Generating conjectures on fundamental constants with the Ramanujan Machine](#). *Nature*, 590:67–73.
- Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. 2024. [Mathematical discoveries from program search with large language models](#). *Nature*, 625:468–475.
- Flavio Salizzoni, Luca Sodomaco, and Julian Weigert. 2025. [Nonlinear Kalman varieties](#). *Preprint*, arXiv:2512.16540.
- Kaushik Sarkar, Charles J. Colbourn, Annalisa De Bonis, and Ugo Vaccaro. 2016. [Partial covering arrays: Algorithms and asymptotics](#). *Preprint*, arXiv:1605.02131.
- Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. 2022. [NaturalProver: Grounded mathematical proof generation with language models](#). In *Advances in Neural Information Processing Systems*, volume 35.
- Dane A. Williamson, Yangfeng Ji, and Matthew B. Dwyer. 2025. [Syntactic blind spots: How misalignment leads to LLMs’ mathematical errors](#). In *Proceedings of the 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pages 1–14, Suzhou, China. Association for Computational Linguistics.
- Yuhuai Wu, Albert Qiaochu Jiang, Wenda Li, Markus Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022. [Autoformalization with large language models](#). In *Advances in Neural Information Processing Systems*, volume 35.
- Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J. Prenger, and Animashree Anandkumar. 2023. [Le-anDojo: Theorem proving with retrieval-augmented language models](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Z.ai. 2026. [GLM-5.1-FP8 model card](#). Hugging Face model card. Accessed: 2026-08-05.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36.

A Leakage Audit

All 61 redacted theorem seeds across the 12 paper sets underwent a post-redaction leakage audit. Of these, 58 seeds (95.1%) were flagged for at least one ambiguity or potential leakage concern. A flag indicated a residual inferential risk requiring review or mitigation rather than a failed redaction. After mitigation, all 61 seeds were retained. The audit achieved 100% processing yield, with no skipped files, unmatched outputs, or input–output identifier mismatches, and no audit summary reported an explicit hidden theorem statement or final conclusion remaining in a seed.

The audit recorded 62 concern groups. The most common involved retained proof, algorithm, or construction architecture (40.3%), followed by hints about equivalence, containment, equality, or characterization (21.0%), quantitative formulas or thresholds (14.5%), and structural or definitional setup that narrowed the likely result (12.9%). Mitigation included removing explicit conclusions and quantitative answers, weakening proof dependencies, reframing revealing relationships as open questions or background, and separating definitions from their consequences. Thus, the remaining leakage risk is primarily inferential and contextual rather than explicit.

B Potential Generalization Axes Audit

Because Potential Generalization Axes were synthesized by the same model used for conjecture generation, we conducted a separate content-flow audit to assess whether this component could reintroduce information about the hidden theorem. Each seed’s axes were compared with the corresponding exact theorem statement and assigned one seed-level severity:

- **Direct theorem-shaped:** an axis names substantially the same objects, conditions, and conclusion or formula pattern, such that converting the question into a declarative statement approximately reconstructs the theorem.
- **Strong directional:** the axes reveal the theorem’s type, architecture, relationship, or expected output, but do not reproduce the complete statement.
- **Low/contextual:** the axes vary assumptions, parameters, methods, or ambient settings without identifying the target conclusion.

Risk level	Seeds	Share
Direct theorem-shaped	12	19.7%
Strong directional	24	39.3%
Low/contextual	25	41.0%
Total	61	100.0%

Table 4: Audit of Potential Generalization Axes for theorem-shaped information.

The 61 seeds contained 450 generalization-axis bullets, with a mean of 7.38 bullets per seed. Table 4 reports the seed-level classifications.

Thus, 36 seeds (59.0%) contained at least strong directional information, including 12 whose axes approximately encoded the shape of the source theorem. This audit characterizes information flow from the source theorem into the constructed component; it does not measure how frequently a generation model would actually reconstruct the theorem from that information.

C Exploratory GLM-5.1-FP8 Judge Analysis

We piloted GLM-5.1-FP8 as an additional expected-theorem-alignment judge on the subset used for the repeated-Gemma and GPT-5-mini analyses. Its scores showed weaker agreement with Gemma, with $r = .649$ and $\text{ICC}(A, 1) = .307$.

An inspection of cases with high Gemma-4-31B and GPT-5-mini scores but low GLM-5.1-FP8 scores suggested that GLM adopted a stricter interpretation of the evaluation prompt. Its rationales frequently noted that the generated output introduced claims “not present in or supported by the Expected Output,” effectively treating the expected theorem as the evidentiary basis for every claim. Gemma-4-31B and GPT-5-mini were more permissive toward additional claims when the output otherwise corresponded closely to the target theorem.

We therefore treat the GLM-5.1-FP8 run as an exploratory diagnostic of judge and prompt sensitivity rather than part of the primary cross-model agreement analysis. This post hoc treatment was not based on a prespecified agreement threshold.

D Theorem-Conditioned Conjecture Quality Prompt

- "Evaluate the actual output as a generated mathematical conjecture. The goal is to assign a score based on the quality, relevance, and mathematical usefulness of the conjecture."

- "Use the input as the context that was provided to the conjecture-generating model. Use the expected output as a theorem from a mathematics paper whose preceding text was used to generate the conjecture. Use the actual output as the generated conjecture being evaluated."
- "First, compare the actual output to the expected output. Determine whether the actual output is essentially the same as the expected theorem, a close variant, a related generalization or specialization, loosely related, or unrelated."
- "Second, remove fluff. Ignore generic motivational language, vague claims of importance, broad exposition, future-work language, and text that does not assert a mathematical claim. Determine whether a clear mathematical statement remains."
- "Assess whether anything mathematically meaningful remains after removing fluff. If the remaining content does not contain a clear mathematical statement, assign a low score. If it appears to contain a potentially new mathematical statement, continue evaluating it using the knownness, simple-case plausibility, and relevance criteria below."
- "Assess simple-case plausibility. Penalize statements that are internally inconsistent, obviously false, missing necessary hypotheses, overgeneralized in a way that would likely fail immediately, or likely to fail on the first interesting example. Do not perform extensive computation; use basic mathematical reasoning only."
- "Assess knownness and novelty. A known result is less valuable as a new conjecture than a plausible new statement. However, do not overly penalize close recovery of the expected theorem: similarity to the expected theorem is evidence that the output is mathematically meaningful, interesting, and contextually relevant."
- "Assess relevance to the expected theorem. Outputs close to the expected theorem should generally receive higher scores. Outputs very different from the expected theorem should often receive C-level scores if they contain

substantive mathematical content but are difficult to evaluate confidently."

- "Assign a continuous score from 0 to 1 using this scale: 1.0 means A-level: clear, substantive, plausible, highly relevant, and mathematically interesting; it closely recovers, usefully refines, or interestingly extends the expected theorem. 0.75 means B-level: good mathematical conjecture; mostly clear, plausible, and relevant, but less precise, less novel, or less strong than A. 0.5 means C-level: contains potentially interesting mathematical content, but its relevance, novelty, precision, or plausibility is uncertain; this includes many outputs that are quite different from the expected theorem but not clearly bad. 0.25 means D-level: weak conjecture; little substance remains after removing fluff, or the statement is vague, trivial, mostly known, poorly connected, or mathematically imprecise. 0.0 means E-level: no usable conjecture; incoherent, unrelated, essentially only fluff, or obviously false in simple cases."
- "Use intermediate values when appropriate. For example, a score of 0.6 may indicate an output between C and B, while a score of 0.9 may indicate a very strong output that is slightly below ideal."

E Pipeline Prompt Protocol

The conjecture-generation pipeline used three prompts in sequence: (1) theorem classification and theorem-specific object-map construction, (2) theorem-specific redaction, and (3) blind conjecture generation. This appendix summarizes the substantive instructions that governed model behavior. We omit administrative instructions concerning directory paths, filenames, Markdown formatting, identifier bookkeeping, summary files, and behavior when no input files were present.

E.1 Theorem Classification and Object-Map Construction

The first prompt instructed the model to inspect a paper, classify its theorem-like environments, and identify its main proved theorems. The number of qualifying results was not specified in advance, and the model was explicitly instructed not to force a paper to contain a fixed number of main theorems.

A result qualified as a *main proved theorem* only when it satisfied all of the following criteria:

1. it was presented as a contribution of the current paper;
2. it had a proof in the current paper;
3. it was not merely quoted from prior work;
4. it was not a corollary, lemma, proposition, example, remark, definition, or background result;
5. it played a central role in the paper’s stated contribution.

Evidence could include emphasis in the abstract or introduction, language such as “we prove,” and the presence of a proof in the paper. The model separately classified quoted prior theorems, supporting results, corollaries, and unproved or deferred theorems. Ambiguous cases were to be marked as unclear rather than treated as main proved theorems.

For each qualifying result, the model constructed a separate theorem-specific object map. The object map contained:

- the exact source theorem, its location, and its proof status;
- mathematical objects relevant to the result;
- required definitions and notation;
- assumptions and structural conditions;
- supporting lemmas, propositions, examples, constructions, quoted results, and background facts;
- an explicit or inferred proof-dependency chain;
- a concise proof-strategy summary;
- notes identifying material that should later be removed or generalized.

An object could appear in more than one map when it was relevant to multiple theorems. At this stage, theorem statements and proof information were permitted because the maps served as unredacted intermediate representations. The model was instructed to preserve the source notation, use only the supplied paper, avoid web search, and not invent theorem statements, labels, or proof dependencies.

E.2 Theorem-Specific Redaction

The second prompt converted each object map independently into a redacted seed for blind conjecture generation. Its core instruction was: *keep setup, remove answers*. The seed was intended to represent the local mathematical setting surrounding one source theorem without revealing the theorem itself.

The prompt permitted the seed to retain or non-conclusively rewrite:

- the ambient mathematical setting;
- definitions and notation;
- fields, rings, groups, spaces, graphs, categories, probability spaces, and other domain objects;
- assumptions on parameters;
- relevant subsets, maps, operators, relations, functions, measures, metrics, bases, generators, and constructions;
- hypotheses and broad structural constraints;
- examples and auxiliary objects that did not reveal the conclusion;
- broad, setup-grounded directions along which conjectures might be formed.

The prompt required complete removal of:

- the source theorem statement, conclusion, label, and any revealing title;
- statements describing what the theorem proves;
- the number or identity of the paper’s main proved theorems;
- proof summaries, proof sketches, dependency chains, and exact logical paths leading to the theorem;
- result classifications and descriptions identifying a result as main, proved, central, or the paper’s primary goal;
- theorem-concluding classifications, decompositions, equivalences, isomorphisms, bounds, rates, formulas, or existence claims;
- descriptions whose purpose was to explain an object’s role in the theorem conclusion.

When theorem-dependent text also contained useful setup information, the model was instructed to rewrite it as broader, non-conclusive context rather than copy it. For example, a claim that an invariant equals a particular formula could be reduced to the statement that the setup contains an invariant associated with the objects under study. If information might reveal the conclusion, the prompt instructed the model to remove or generalize it and to prefer under-inclusion over leakage.

Each resulting seed was organized into the six context components used in the ablation study:

1. Ambient Setting;
2. Core Definitions and Notation;
3. Local Mathematical Objects;
4. Structural Constraints and Conditions;
5. Auxiliary Objects and Constructions;
6. Potential Generalization Axes.

The Potential Generalization Axes were to remain generic and grounded in the redacted setup. Suggested directions included varying parameters, weakening or strengthening assumptions, changing the ambient object or target class, replacing a structure with an analogue, and examining small or exceptional cases.

The redaction output also contained a forbidden-content section and a redaction audit checking whether theorem statements, conclusions, proof information, exact classifications, and result-count information had been removed. Before conjecture generation, these two administrative sections were manually removed, leaving only the six mathematical-context components as candidate model input.

E.3 Blind Conjecture Generation

The third prompt instructed the model to generate exactly one plausible formal conjecture for each redacted seed. Under an ablation condition, the model received only the context components retained by that condition. It was not permitted to consult the original paper, $\text{\LaTeX}$ source, theorem map, unredacted object map, theorem statement, proof summary, dependency chain, redaction report, or any external source.

For each seed, the model was instructed to choose the single most natural conjecture suggested by the available local objects and constraints.

The conjecture was written in $\text{\LaTeX}$ and could use definitions, notation, objects, assumptions, examples, and structural constraints present in the seed. The prompt permitted natural mathematical extrapolation, including a classification, generalization, strengthening, weakening, analogue, exceptional-case statement, structural claim, existence or uniqueness claim, or dimension or bound conjecture. A newly proposed formula was permitted only when it was derived from the redacted setup rather than copied from an unredacted result.

The prompt stated that proximity to the hidden theorem was acceptable, and even desirable, provided that the conjecture followed only from the redacted seed. At the same time, the model was prohibited from claiming that the paper proved the conjecture, claiming knowledge of the hidden theorem, copying an exact conclusion from an unredacted source, or generating multiple candidate conjectures for one seed. Information from one theorem-specific seed could not be used for another unless it appeared independently in both.

For each seed, the requested response contained one precise formal statement, the seed objects and assumptions used, a conjecture-type classification, a plausibility rationale, a tentative proof direction, a counterexample-search strategy, an assessment of possible proximity to a hidden source result, and a blindness and leakage audit. The formal statement was treated as the generated conjecture in the subsequent evaluation.

F Complete Ablation Results

Table 5 reports the complete leave-one-out and single-component results.

G Complete Domain-Level Results

Table 6 reports the complete equal-weight paper-level domain summaries for both evaluation metrics, including all leave-one-out effects. The main-text summary in Table 2 reports only the largest alignment loss and strongest standalone component within each domain. Each domain contains three papers, so these comparisons are exploratory.

(a) Leave-one-out conditions							
Metric	Removed component	Full	Ablated	Δ	95% CI	d_z	p_{Holm}
Quality	Ambient setting	.931	.923	-.008	[-.028, .011]	-.108	1.000
	Definitions and notation	.931	.931	.000	[-.022, .022]	.000	1.000
	Local mathematical objects	.931	.923	-.008	[-.030, .013]	-.097	1.000
	Structural constraints	.931	.910	-.021	[-.051, .008]	-.187	1.000
	Auxiliary objects	.931	.938	.007	[-.013, .026]	.085	1.000
	Generalization axes	.931	.936	.005	[-.012, .022]	.074	1.000
Alignment	Ambient setting	.607	.582	-.025	[-.084, .035]	-.106	1.000
	Definitions and notation	.607	.572	-.034	[-.110, .042]	-.116	1.000
	Local mathematical objects	.607	.575	-.031	[-.105, .042]	-.108	1.000
	Structural constraints	.607	.520	-.087	[-.154, -.020]	-.333	.055
	Auxiliary objects	.607	.579	-.028	[-.109, .053]	-.088	1.000
	Generalization axes	.607	.641	.034	[-.038, .107]	.121	1.000

(b) Single-component conditions			
Component retained	Quality	Alignment	Alignment gap
Ambient setting	.885	.348	-.259
Definitions and notation	.889	.397	-.210
Local mathematical objects	.898	.457	-.149
Structural constraints	.851	.348	-.259
Auxiliary objects	.844	.270	-.336
Generalization axes	.820	.266	-.341

Table 5: Complete ablation results. Panel (a) reports theorem-level leave-one-out comparisons, and Panel (b) reports standalone performance under the single-component conditions.

(a) Theorem-conditioned conjecture quality							
Domain	Full	Ambient	Definitions	Local objects	Constraints	Auxiliary	Generalization
math.AC	.934	-.030	-.008	-.018	-.008	.010	-.003
math.AG	.881	.000	.029	.019	-.037	.000	.016
math.CO	.950	.006	-.019	-.006	-.008	.017	.022
math.ST	.986	-.018	-.008	-.059	-.014	-.020	-.014

(b) Expected-theorem alignment							
Domain	Full	Ambient	Definitions	Local objects	Constraints	Auxiliary	Generalization
math.AC	.611	-.037	-.061	.004	-.162	.011	-.047
math.AG	.571	-.052	-.053	-.060	-.121	-.087	-.046
math.CO	.644	-.003	-.086	-.042	-.047	.019	.103
math.ST	.664	.026	.073	-.096	-.011	-.054	.162

Table 6: Equal-weight paper-level results by domain. The full-context column reports the domain baseline; the remaining columns report leave-one-out differences from that baseline. Negative values indicate lower performance after removing the component. Each domain contains three papers.