# A Study of Naturally Occurring Mathematical Errors in Real-World Theoretical Research with Frontier Models

**Anonymous authors**
Paper under double-blind review

## Abstract

Recent advances in artificial intelligence have enabled frontier models to play a growing role in theoretical research. AI for mathematics, in particular, has attracted widespread attention. Researchers increasingly turn to proprietary models such as Fable 5 and GPT-5.6 to investigate research-level open questions. However, these models remain prone to mathematical errors, making it important to understand how such errors arise and what they imply for the responsible use of AI in theoretical research. Existing studies often examine mathematical errors through evaluations on selected problems, particularly in competition- and Olympiad-level mathematics. Systematic documentation and analysis of mathematical errors encountered during sustained, real-world theoretical research with frontier models remain limited. In this paper, we present 65 source-linked mathematical error records drawn from frontier-model investigations of 12 research-level problems in quantum information, quantum error correction, quantum complexity, and related fields. Each record links an erroneous claim or inference to its original source, relevant context, and a checkable witness explaining the mathematical defect. Where the evidence permits, we also examine visible reasoning summaries alongside subsequent outputs. Using an established classification framework, we find that logical and reasoning errors dominate our corpus, accounting for 68.2% of error labels. Overall, our study provides an empirical analysis of mathematical errors encountered in real-world theoretical research with frontier models.

## 1 Introduction

Over the last several months, AI-assisted mathematical results have attracted attention worldwide (Tao, 2026; Conant, 2026; Faulhuber et al., 2026; OpenAI, 2026). These results demonstrate that frontier models, such as those in the Claude and GPT families, can contribute to theoretical research beyond solving benchmark problems. Examples include a counterexample to the Jacobian conjecture in dimension three (Tao, 2026) and a proof of the mod 4 Kawauchi conjecture (Conant, 2026). In analysis, Faulhuber et al. (2026) developed a counterexample to the Heil–Ramanathan–Topiwala conjecture through a research dialogue with a language model, followed by extensive human verification and revision. OpenAI (2026) also reported ten advances across mathematics and theoretical computer science, including coding theory, group theory, and quantum complexity. These examples illustrate research workflows in which models contribute constructions and proof strategies, while researchers examine, refine, and validate the resulting arguments. Yet a productive research direction does not ensure that every intermediate argument is sound: the HRT study explicitly documents substantial gaps and an error that required a new argument to repair (Faulhuber et al., 2026, Section 1.2). Understanding mathematical errors within such investigations is therefore an important counterpart to documenting their successes.

Prior work has studied mathematical reasoning errors across different problem settings and models (Lightman et al., 2024; Zheng et al., 2025; Yoshida et al., 2026; Abouzaid et al., 2026; Banerjee & Bhattacharjee, 2026). Some studies examine mathematical errors on competition- and Olympiad-level problems (Lightman et al., 2024; Zheng et al., 2025; Yoshida et al., 2026). Others investigate research-level questions, documenting failures in model-generated mathematical arguments (Abouzaid et al., 2026; Banerjee & Bhattacharjee, 2026). These studies provide substantial empirical material and



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.

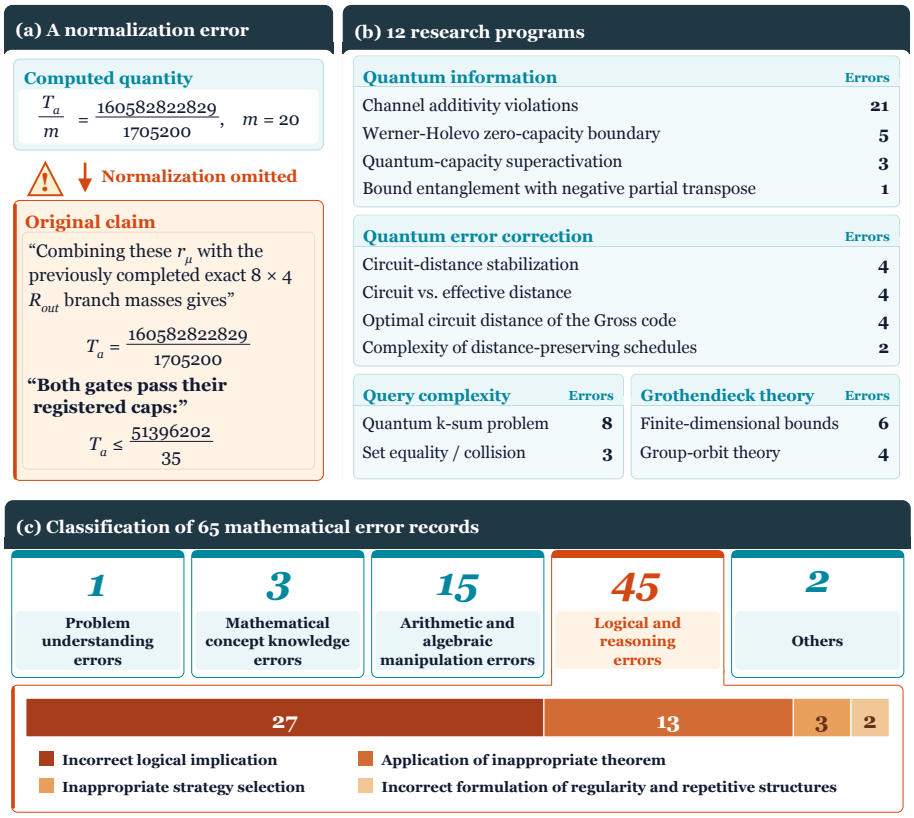


Figure 1: **Study overview.** **(a)** A normalization error: evidence about $T_a/m$ is used to claim a bound on $T_a$ without restoring the factor $m$. **(b)** The 12 research problems contributing 65 mathematical error records. **(c)** The distribution under the classification framework of Yoshida et al. (2026), with a further breakdown of logical and reasoning errors.

important foundations for understanding mathematical errors. However, the models evaluated by Yoshida et al. (2026), such as Llama-3.3-70B-Instruct and Qwen3-32B, achieve default-prompt accuracies of approximately 74.2% and 83.9% on their MATH test split, whereas publicly reported MATH-500 accuracies reach 96.7% for GPT-5.6 Sol and 97.6% for Claude Opus 5 (Open Benchmarks, 2026). With frontier models achieving higher accuracy on established mathematical benchmarks, it remains unclear whether previously observed error patterns reflect those encountered when these models are used in sustained, real-world theoretical research.

Collecting the evidence needed for such an analysis presents two practical challenges. First, publicly shared results may describe AI contributions and subsequent corrections without providing the complete interaction history (Faulhuber et al., 2026, Section 1.2). The complete conversations, intermediate proof drafts, code, and revisions needed to understand an error can therefore be difficult to recover from publicly available materials. Second, collecting naturally occurring errors from frontier models can require substantial inference resources. High answer accuracy on established benchmarks makes incorrect answers less frequent under routine sampling, while collecting errors in real-world research may require investigating many research-level problems and carefully examining the resulting arguments and computations. Together, these challenges make it difficult to assemble a sufficiently large, well-documented collection of mathematical errors. Building such a collection requires not only identifying incorrect claims, but also preserving the research context in which they arose and verifying precisely why they fail.

Our collection comprises 65 source-linked mathematical error records from frontier-model investigations of 12 research-level problems spanning quantum information, quantum error correction, quantum complexity, and related fields, as summarized in Figure 1(b). The errors arose naturally during proof development and related computational work, including unfinished proof attempts. Each

record links an erroneous claim, inference, or computation to its original source, relevant context, and a checkable witness establishing the specific mathematical failure. Where available, supporting process evidence provides further context for how the error arose or was subsequently used.

We adopt the classification framework used by Yoshida et al. (2026) to analyze the distribution of errors in our corpus and compare it with findings on competition- and Olympiad-level mathematics. Logical and reasoning errors dominate our corpus, appearing in 45 records and accounting for 68.2% of error labels, compared with 37.2% in their MATH analysis. Within this category, incorrect logical implication is the dominant subtype, accounting for 27 of the 45 records. In contrast, problem understanding errors account for only 1.5% of error labels (one record), compared with 29.9% in the MATH analysis (Figure 1(c)). Our contributions are twofold:

1. **A collection of naturally occurring mathematical errors.** We present 65 source-linked mathematical error records encountered during real-world theoretical research, in which frontier models were used to investigate 12 research-level problems. Each record connects an erroneous claim or inference to its original source and a checkable witness, together with relevant context and supporting process evidence where available.
2. **An empirical analysis of mathematical errors.** We examine the distribution of errors encountered when frontier models investigate research-level problems and use the associated research materials to understand how these errors arise in practice.

## 2 Related Work

Previous studies have examined mathematical errors in model-generated solutions, particularly on competition- and Olympiad-level problems (Lightman et al., 2024; Zheng et al., 2025; Yoshida et al., 2026). Of particular relevance to our study, Yoshida et al. (2026) classify mathematical errors into five categories: problem understanding, mathematical concept knowledge, arithmetic and algebraic manipulation, logical and reasoning, and others. They further distinguish subtypes within these categories and analyze their distribution in model-generated solutions to MATH problems. We adopt their classification framework to analyze errors encountered during sustained, real-world theoretical research with frontier models, comparing both the broad error categories and the subtypes of logical and reasoning errors. Our work complements these studies with source-linked error records collected during ongoing research, together with their original context and checkable mathematical witnesses.

## 3 Research Setting and Error Collection

We study mathematical errors encountered while using frontier models to investigate 12 research-level problems in quantum information, quantum error correction, quantum complexity, and related fields. Table 1 lists the problems and their research objectives; Figure 1(b) summarizes their contributions to the corpus. Appendix A provides the mathematical background and investigated targets for each problem. During these investigations, models proposed constructions, developed intermediate arguments, wrote verification programs, interpreted computational results, and revised proof attempts. We examine the resulting conversations, proof drafts, code, computational outputs, and revision histories. Our analysis covers mathematical errors in both completed proofs and unfinished proof attempts. We include errors in intermediate claims, inferences, and computations, regardless of whether the attempt ultimately leads to a complete proof.

Our unit of analysis is an individual, identifiable mathematical error, rather than an entire conversation or research problem. A single research problem may therefore contribute multiple error records. For each candidate error, we locate the original statement or computation and examine its context to determine what was claimed, which assumptions were used, and what evidence was available. We then review the relevant argument or computation to identify the specific mathematical defect and the evidence establishing it.

We organize the collection as an error index. Each entry contains four fields: a unique record identifier (`record_id`), the research problem (`problem`), a brief description of the error (`summary`), and a link to the corresponding witness document (`evidence`) in a per-record evidence folder. Each folder contains the relevant original materials and a *witness*. The original materials include source passages with their locations and surrounding context, as well as relevant proof drafts, code,

Table 1: The 12 research problems contributing to the corpus.

| Area | Research problem | Research objective |
|---|---|---|
| Quantum information | Channel additivity violations | Bound and construct additivity violations for minimum output entropy and Holevo information, including their scaling with dimension. |
| | Werner–Holevo zero-capacity boundary | Determine the parameter region in which noisy Werner–Holevo channels have zero quantum capacity. |
| | Quantum-capacity superactivation | Characterize zero-capacity channel pairs whose joint quantum capacity is positive, and bound the size of the effect. |
| | Bound entanglement with negative partial transpose | Determine whether a state with negative partial transpose can remain undistillable for every number of copies. |
| Quantum error correction | Circuit-distance stabilization | Characterize eventual circuit distance under repeated extraction rounds and the conditions for reaching the code distance. |
| | Circuit vs. effective distance | Characterize when effective logical errors lift to full circuit logical errors without additional faults, giving equal distances. |
| | Optimal circuit distance of the Gross code | Determine the maximum circuit distance over legal syndrome-extraction schedules of the $[[144, 12, 12]]$ Gross code. |
| | Complexity of distance-preserving schedules | Determine the complexity of deciding whether a syndrome-extraction schedule can preserve the code distance. |
| Quantum query complexity | Quantum $k$-sum problem | Establish query lower bounds for $k$-sum over small alphabets and understand their dependence on alphabet size. |
| | Set equality / collision | Develop explicit quantum query lower-bound certificates for set equality and the related collision problem. |
| Grothendieck theory | Finite-dimensional Grothendieck bounds | Certify dimension-four Grothendieck bounds by comparing vector values with unrestricted classical optima. |
| | Group-orbit Bell matrices | Determine how finite group symmetry constrains vector optima, classical values, and the dimensions of optimal strategies. |

computational outputs, and visible reasoning summaries where available. A *witness* explains the specific mathematical defect and substantiates it with checkable evidence, such as a counterexample, an explicit derivation, or a reproducible computation, with references to the relevant original materials. The four-field entries are provided in `index.json`; the corresponding folders contain `original/` and `witness.md`. The index contains the 65 included error records. Per-record adjudication and verification details are retained with the evidence materials. The entry for Record 001 is shown below.

**Record index format**

```
{
  "record_id": "001",
  "problem": "Channel additivity violations",
  "summary": "Necessity for the particular localization proof is not
  ↪  necessity for every construction.",
  "evidence": "records/001/witness.md"
}
```

# 4 CLASSIFICATION OF MATHEMATICAL ERRORS

We adopt the five categories of Yoshida et al. (2026): problem understanding errors, mathematical concept knowledge errors, arithmetic and algebraic manipulation errors, logical and reasoning errors, and others. Table 2 summarizes the five categories. Figure 1(c) summarizes the distribution of these categories across the 65 records. Logical and reasoning errors are the most frequent, appearing in 45 records, as discussed in Section 4.4. They are followed by arithmetic and algebraic manipulation errors in 15 records (Section 4.3) and mathematical concept knowledge errors in three records (Section 4.2). Problem understanding errors occur in one record and are examined in Section 4.1. The two remaining records concern data-update or transcription errors and are grouped under *others*. Based on this analysis, we compare our error distribution with findings on competition- and Olympiad-level mathematics in Section 4.5. In addition, we conduct a supplementary analysis of errors in visible reasoning and subsequent outputs in Section 4.6.

Table 2: Five error categories following Yoshida et al. (2026).

| Category | Description |
|---|---|
| Problem Understanding Errors | Misunderstanding the information or conditions given in the problem. |
| Mathematical Concept Knowledge Errors | Incorrect statements about mathematical concepts, properties, theorems, or formulas. |
| Arithmetic and Algebraic Manipulation Errors | Errors in carrying out numerical calculations or algebraic operations. |
| Logical and Reasoning Errors | Errors in selecting or applying reasoning methods, or in drawing conclusions from premises. |
| Others | Errors outside the preceding four categories. |

## 4.1 PROBLEM UNDERSTANDING ERRORS

Problem understanding errors occur when the information or conditions given in a problem are misinterpreted, omitted, or changed. Our corpus contains one such record. It drops a required positive-gap condition when reformulating a channel-additivity task, as illustrated in Example 1. This concerns the preservation of the research question when it is translated into a follow-up task.

**Example 1. Dropping a required condition from a follow-up task**

**Original task and faulty reformulation.** A channel-additivity investigation requires a positive certified gap. For the one-layer case, this condition is

$$\delta = \frac{\log_2 m}{m} - 2\log_2 A > 0, \qquad A = 1 + \frac{9(1+\eta)^2}{m}, \quad \eta > 0.$$

An agent rewrites the next task as finding a finite-dimensional channel construction satisfying only

$$\ln m > 18, \qquad p_1(C) \le A/m, \qquad \kappa_1(U) \le A - 1.$$

Here, $p_1(C)$ is the maximum output purity, and $\kappa_1(U)$ is the orbit-block ratio used to control two-copy purity. The rewritten task omits $\delta > 0$.

**Witness.** Take $m = 2^{27}$, input dimension $D = 1$, scalar unitaries $U_1 = \cdots = U_m = 1$, and $\eta = \sqrt{m(m-1)}/3 - 1$. Then $A = m$, $p_1(C) = 1 = A/m$, and $\kappa_1(U) = m - 1 = A - 1$, so the rewritten conditions hold, but

$$\delta = \frac{27}{2^{27}} - 54 < 0.$$

Thus, solving the rewritten task need not satisfy the original positive-gap requirement. The receiving solver correctly identifies the omission; the error is in the agent-written task, not the solver's qualified answer.

The example illustrates a problem understanding error because an explicit requirement is lost when the problem is restated. This differs from retaining the requirement but making an invalid inference while attempting to prove it.

### 4.2 Mathematical Concept Knowledge Errors

Mathematical concept knowledge errors concern incorrect statements about mathematical concepts, properties, theorems, or formulas. We identify three records in this category. One states that a normalized symplectic Fourier transform squares to phase-space reflection, although its specified kernel gives $\mathcal{F}^2 = I$. Another asserts eventual periodicity of min-plus matrix powers without qualifications, contradicted by the scalar matrix $A = [1]$, whose $R$th min-plus power is $[R]$. The third equates swap invariance with symmetric support in a channel-extension program (Example 2).

**Example 2. Confusing invariance with symmetric support**

**Original construction.** The task asks for an extension $\eta$ on three qutrit systems $A \otimes B_1 \otimes B_2$, invariant under swapping $B_1$ and $B_2$. The program instead represents it as $\eta = V\rho V^{\mathsf{T}}$, where the columns of $V$ span only the 18-dimensional subspace $A \otimes \mathrm{Sym}^2(B)$.
**Witness.** Let $S$ swap the two $B$ systems while leaving $A$ unchanged, and set

$$v = \frac{|01\rangle - |10\rangle}{\sqrt{2}}, \qquad X = |0\rangle\langle 0|_A \otimes |v\rangle\langle v|.$$

Swapping the two systems changes the sign of $v$ but leaves its projector unchanged, so $SXS = X$. This positive, swap-invariant operator has antisymmetric support and cannot be represented by the proposed parameterization. Swap invariance does not imply symmetric support.

### 4.3 Arithmetic and Algebraic Manipulation Errors

Arithmetic and algebraic manipulation errors occur in numerical calculations, algebraic transformations, counting, or comparisons of quantities. We identify fifteen records in this category. The errors range from integer multiplication and off-by-one bounds to noncommutative products (Example 3), modular fiber counts, and matrix contractions. For example, one modular-counting argument treats a nonzero linear form modulo a composite integer as if all its fibers had the size obtained for a surjective form. Another contraction loses the dimension factor supplied by $\mathrm{Tr}(J^{-1}J)$. These are defects in the displayed mathematical operations, rather than inferences that a correctly computed result settles a larger problem.

**Example 3. Reversing the order of matrix multiplication**

**Original claim.** A channel-additivity argument defines $A_{ab} = U_b^* U_a$ for unitary matrices, but later states $A_{ab}A_{bc} = A_{ac}$.
**Witness.** The valid cancellation has the opposite order:

$$A_{bc}A_{ab} = (U_c^* U_b)(U_b^* U_a) = U_c^* U_a = A_{ac}. \tag{1}$$

Take $U_a = I$, $U_b = X$, and $U_c = Z$, where $X$ and $Z$ are Pauli matrices. Then $A_{ab}A_{bc} = XZX = -Z$, whereas $A_{ac} = Z$. Thus, the asserted identity fails under the definitions already given.

The consequences vary with the role of the calculation. An incorrect printed integer does not necessarily invalidate an independently checked nonnegativity result, whereas an incorrect fiber count can invalidate the construction that relies on it.

### 4.4 Logical and Reasoning Errors

Logical and reasoning errors occur when a conclusion does not follow from its premises or a mathematical result is applied without meeting its conditions. We identify 45 records in this category, the largest in our corpus. They include applying results outside their assumptions, drawing exhaustive conclusions from restricted searches, and claiming a construction has properties that have not been established.

Several failures involve an intermediate result that is valid in its own setting. For example, Figure 1(a) shows a normalized calculation being used to certify a bound on the corresponding unnormalized quantity. The calculation alone cannot justify that change of quantity. Other records transfer a bound between different memory sectors, take a necessary condition as sufficient, or infer that a projected family retains its original dimension, as illustrated in Example 4.

Other records contain explicit but invalid derivations. A circuit-distance argument infers eventual equality to the static code distance from eventual constancy at an unspecified value. A scheduling argument infers that two groups of gates do not interleave from agreement of their pairwise orderings. A Grothendieck-bound argument invokes weak duality to treat a separating moment functional as an upper-bound certificate, although the referenced construction has the opposite certificate direction. These records are classified by the invalid inference or theorem application documented in the original argument.

**Example 4. Assuming independence survives projection**

**Original claim.** A Grothendieck-bound investigation claims that the kernel supported on the retained coordinates has dimension at least 28 rather than 12, improving an operator-rank bound to 746,797. The argument uses 28 independent directions in the unpruned representation.
**Witness.** After projection onto the 746,825 retained coordinates, 16 of those directions vanish. The projected family has rank 12: a retained $12 \times 12$ minor has nonzero determinant $-120231596497305600$. These directions therefore support the bound $746{,}825 - 12 = 746{,}813$, not $746{,}825 - 28 = 746{,}797$. The witness invalidates the proposed rank improvement; it does not rule out establishing a stronger bound by another argument.

### 4.5 Comparison with Competition- and Olympiad-Level Mathematics

We compare the relative distribution of error categories in our corpus with the MATH analysis of Yoshida et al. (2026). Their analysis covers 583 problems selected after Llama-3.3-70B-Instruct failed the answer check. To show the composition of each collection, we normalize category counts by the total number of error labels in each collection, excluding the 15 cases classified as correct or non-errors in MATH.

As shown in Figure 2, logical and reasoning errors are the largest category in both collections. They are particularly dominant in our research records, accounting for 68.2% of error labels (45 records), compared with 37.2% in the MATH analysis. To examine this concentration more closely, we further classify these 45 records using the eight subcategories of Yoshida et al. (2026). Figure 3 presents the comparison. Incorrect logical implication is the largest subtype in our collection, accounting for 60.0% (27 records), compared with 24.8% in MATH. These errors include inferring that independent directions remain independent after projection and treating a restricted search as exhaustive. Inappropriate theorem application accounts for a further 28.9% (13 records), compared with 33.5% in MATH. Together, these two subtypes account for 40 of the 45 records, indicating that the concentration primarily concerns what follows from intermediate results and when those results can be applied. The remaining five records are mapped to strategy selection and regularity formulation.

Problem understanding errors show the opposite pattern: they account for 1.5% of our error labels (one record), but are the second-largest category in the MATH analysis at 29.9%. A possible explanation lies in the structure of the investigations. Problem understanding concerns the interpretation of the information and requirements given in a problem, whereas the MATH collection covers many distinct questions, each requiring a new interpretation. Our records instead come from sustained investigations of 12 research problems, with many intermediate claims and deductions within each

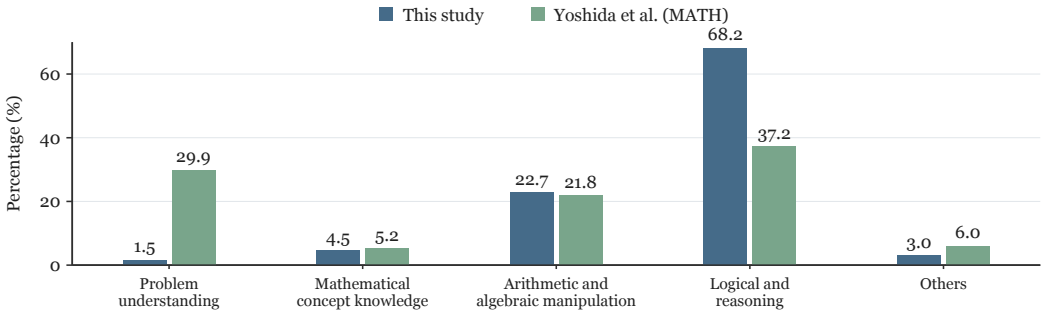


Figure 2: Distribution of mathematical errors in this study and Yoshida et al. (2026, Table 2).

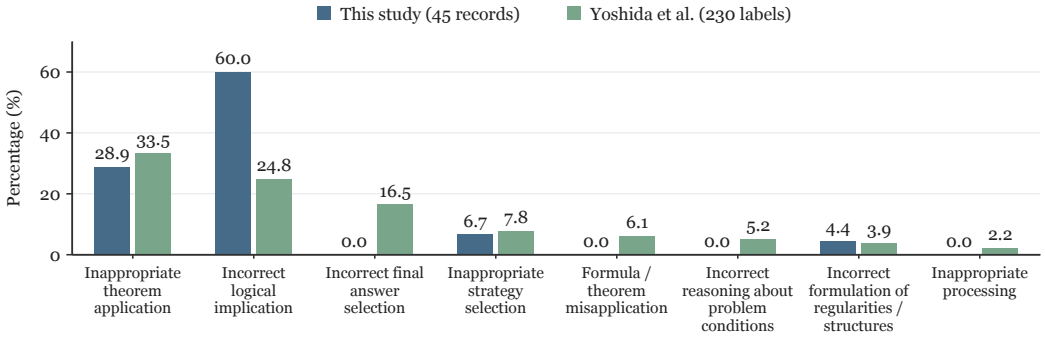


Figure 3: Distribution of logical and reasoning errors in this study and Yoshida et al. (2026, Table 2).

investigation. This structure may create relatively more opportunities for invalid inferences and failures to preserve the conditions of intermediate results than for misunderstanding the original task. The other broad categories have more similar proportions: arithmetic and algebraic manipulation accounts for 22.7% versus 21.8%, mathematical concept knowledge for 4.5% versus 5.2%, and others for 3.0% versus 6.0%.

### 4.6 Errors in Visible Reasoning and Subsequent Output

We further examine whether an error is already visible in the reasoning or appears as a discrepancy between that reasoning and a subsequent output. This supplementary analysis uses provider-visible reasoning summaries and excerpts, not complete internal chain-of-thought traces. Among the 65 accepted records, 33 have linked visible reasoning excerpts; 12 support the specific error-to-reasoning alignments analyzed here. In 11 of these records, the error is already visible in the reasoning summary, while the remaining record shows a discrepancy between the reasoning summary and the output.

**The erroneous judgment is already visible in reasoning.** In eleven records, the summary or excerpt contains the same erroneous judgment identified by the mathematical witness. Ten also have a separately located subsequent output containing it; in the remaining case, the original erroneous statement is itself in the visible reasoning. These include the invalid inference from positive two-way-assisted capacity to non-antidegradability, an incorrect modular count, and the orbit-count judgment in Example 5. The circuit-distance record also contains an early inference from stabilization to a finite saturation index, followed by recognition that the index may be infinite; the later output nevertheless repeats the invalid implication. Thus, the presence of a correct qualification elsewhere in the same reasoning does not make this an output-only error. We do not assume that every passage surrounding these judgments is incorrect or that the earliest observable statement is the first internal occurrence of the error.

**Example 5. An incorrect orbit count appears before the written proposition**

**Visible reasoning.** During a quantum $k$-sum investigation, the summary states that, for nonidentity permutations, "the domain decomposition never forms a single orbit."
**Output.** The draft claims that the negative inputs $D_0$ are not a single $S_n \times C(\sigma)$ orbit, already for $n = 2$, and that their orbit count is at least two. Here, $S_n$ permutes input positions and $C(\sigma)$ consists of alphabet permutations commuting with $\sigma$.
**Witness.** For $G = \mathbb{Z}_2$, target $t = 1$, and $n = 2$, the involution $\sigma(x) = 1 - x$ is nonidentity. The negative inputs are $D_0 = \{(0, 0), (1, 1)\}$, since neither pair sums to 1 modulo 2. The alphabet swap commutes with $\sigma$ and maps one input to the other, so $D_0$ is a single orbit. The draft's own formula gives $p_{\leq 1}(2) = 1$, the number of partitions of 2 into at most one part, not at least two. This same false universal judgment is visible before it is written into the proposition.

**Discrepancy between that reasoning and a subsequent output.** In one record, the aligned visible reasoning acknowledges an alphabet-size restriction that the later statement does not preserve (Example 6). We have not located the corresponding erroneous universal claim in the preceding visible reasoning.

**Example 6. An alphabet restriction is lost in a later bound**

**Visible reasoning.** The summary explicitly tracks a lower-bound requirement on the alphabet size: "with the condition $q \geq (k - 1)(kK + 1)$ for the shift family," where $K$ is a construction parameter.
**Output.** The draft nevertheless states a quantum query lower bound $\Omega_k(\sqrt{N}\, q^{1/6})$ "for all $q$ up to $\Theta_k(N)$," without retaining a lower cutoff.
**Witness.** Set $k = 3$, $q = 3$, target $t = 0$, and $N \geq 7$. Every length-$N$ input over $\mathbb{Z}_3$ contains three copies of some residue $a$. Their sum is $3a = 0$, so every input is a YES instance and the query complexity is zero. The counterexample refutes the unrestricted displayed range, not the preceding theorem under its alphabet-size condition.

We describe this as a *reasoning–output discrepancy*, but because we have access only to a reasoning summary, we cannot confirm that it is a transcription error.

## 5 Conclusion

We present 65 source-linked mathematical error records collected during real-world theoretical research with frontier models across 12 research-level problems. Each record connects an erroneous claim, inference, or computation to its original context and a checkable mathematical witness. Using an established classification framework, we analyze the error distribution in this corpus and compare it with prior findings on competition- and Olympiad-level mathematics. Logical and reasoning errors dominate our corpus, accounting for 68.2% of error labels. These errors frequently concern how intermediate results are interpreted and applied, highlighting a prominent source of mathematical failure in the research investigations studied here.

## 6 Discussion

Our study has two main limitations. First, the corpus contains only 65 error records and primarily covers quantum information, quantum error correction, quantum complexity, and related fields. Larger collections spanning additional theoretical disciplines would help assess whether the observed patterns extend beyond these research settings. Second, our process-level analysis relies on visible reasoning summaries and excerpts rather than complete internal chain-of-thought traces. These summaries may omit intermediate steps or revisions, limiting our ability to reconstruct error formation in detail and distinguish errors in the underlying reasoning from those introduced during output generation. In future work, we plan to maintain and expand this corpus by documenting new mathematical errors encountered as we continue to use frontier models in theoretical research.

## 7 AI USE STATEMENT

The research questions, conceptual framing, study design, and mathematical claims advanced by this paper were developed by the authors. Generative AI systems assisted with executing parts of the research workflow, including mathematical derivations and proof development, source recovery, data organization, code, and manuscript drafting. All corpus inclusion decisions, mathematical witnesses, aggregate results, and final claims were reviewed and verified by the authors, who take responsibility for the final content.

## REFERENCES


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

Arnesh Banerjee and Ayushi Bhattacharjee. Failure modes of large language models on research-level mathematics: A taxonomy and an empirical characterisation. *arXiv preprint arXiv:2606.24902*, 2026. URL https://arxiv.org/abs/2606.24902.

Aleksandrs Belovs and Ansis Rosmanis. Adversary lower bounds for the collision and the set equality problems, 2013. URL https://arxiv.org/abs/1310.5185.

Aleksandrs Belovs and Robert Špalek. Adversary lower bound for the $k$-sum problem, 2012. URL https://arxiv.org/abs/1206.6528.

Michael E. Beverland, Shilin Huang, and Vadym Kliuchnikov. Fault tolerance of stabilizer channels, 2024. URL https://arxiv.org/abs/2401.12017.

Jim Conant. A proof of the mod 4 Kawauchi conjecture. *arXiv preprint arXiv:2607.18655*, 2026. URL https://arxiv.org/abs/2607.18655v2. Version 2, August 2, 2026.

David P. DiVincenzo, Peter W. Shor, John A. Smolin, Barbara M. Terhal, and Ashish V. Thapliyal. Evidence for bound entangled states with negative partial transpose. *Physical Review A*, 61:062312, 2000. doi: 10.1103/PhysRevA.61.062312. URL https://arxiv.org/abs/quant-ph/9910026.

Michael Epping, Hermann Kampermann, and Dagmar Bruß. Designing bell inequalities from a tsirelson bound. *Physical Review Letters*, 111:240404, 2013. doi: 10.1103/PhysRevLett.111.240404. URL https://arxiv.org/abs/1306.3805.

Markus Faulhuber, Philipp Petersen, Jordy Timo van Velthoven, and Felix Voigtlaender. Linear dependence of time-frequency shifts of a Schwartz function. *arXiv preprint arXiv:2608.05044*, 2026. URL https://arxiv.org/abs/2608.05044v2. Version 2, August 18, 2026.

M. B. Hastings. Superadditivity of communication capacity using entangled inputs. *Nature Physics*, 5, 2009. doi: 10.1038/nphys1224. URL https://arxiv.org/abs/0809.3972.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.20050.

Open Benchmarks. MATH-500 Leaderboard. Kaggle, 2026. URL https://www.kaggle.com/benchmarks/open-benchmarks/math-500. Accessed September 22, 2026.

OpenAI. Ten advances in mathematics and theoretical computer science. OpenAI, August 2026. URL https://openai.com/index/ten-advances-in-mathematics/. August 1, 2026.

Shayan Roofeh and Vahid Karimipour. The noisy Werner–Holevo channel and its properties. *Physical Review A*, 109:052620, 2024. doi: 10.1103/PhysRevA.109.052620. URL https://arxiv.org/abs/2310.15353.

Graeme Smith and Jon Yard. Quantum communication with zero-capacity channels. *Science*, 321: 1812–1815, 2008. doi: 10.1126/science.1162242. URL https://arxiv.org/abs/0807.4935.

Armands Strikis, Dan E. Browne, and Michael E. Beverland. High-performance syndrome extraction circuits for quantum codes, 2026. URL https://arxiv.org/abs/2603.05481.

Terence Tao. A digestion of the Jacobian conjecture counterexample. What's New, July 2026. URL https://terrytao.wordpress.com/2026/07/21/a-digestion-of-the-jacobian-conjecture-counterexample/. July 21, 2026.

Runa Yoshida, Kosuke Nishida, and Kyosuke Nishida. Improving mathematical reasoning capabilities in large language models via reasoning process error classification. *arXiv preprint arXiv:2609.15145*, September 2026. URL https://arxiv.org/abs/2609.15145. September 14, 2026.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://aclanthology.org/2025.acl-long.50/.

# A Background of the Research Problems

This appendix describes the twelve research problems represented in Table 1. It explains their mathematical setting and the objectives of the investigations from which our records were collected. The descriptions are not claims that the investigations resolved these problems. Related computational and proof-development branches are grouped under the same problem.

## A.1 Quantum Information

**Channel additivity violations.** A quantum channel describes a noisy physical process acting on a quantum state. Its minimum output entropy, $S_{\min}(\mathcal{N}) = \min_\rho S(\mathcal{N}(\rho))$, where $S$ is the von Neumann entropy, quantifies how little output uncertainty the channel can produce. Additivity asks whether $S_{\min}(\mathcal{N} \otimes \mathcal{M}) = S_{\min}(\mathcal{N}) + S_{\min}(\mathcal{M})$. Hastings established that this equality can fail, with consequences for the additivity of classical communication quantities (Hastings, 2009). Our investigations concerned the size and construction of such violations, including whether gaps can grow proportionally to the logarithm of channel dimension. Related branches studied Holevo information, which measures the classical information available in an output ensemble, and explicit finite-dimensional constructions. These branches are grouped as one research problem. The work involved entropy estimates, tensor-product inputs, and computational certificates, with bounds required to apply to the actual channels and the full admissible input class.

**Werner–Holevo zero-capacity boundary.** The quantum capacity $Q(\mathcal{N})$ is the asymptotic rate at which a channel can reliably transmit quantum information without auxiliary communication. We investigated the noisy qutrit Werner–Holevo family $\Lambda_x(\rho) = (1-x)\rho + \frac{x}{2}[\mathrm{Tr}(\rho)I - \rho^T]$, with $0 \le x \le 1$, and asked for the parameter region in which $Q(\Lambda_x) = 0$. This family interpolates between the identity and the Werner–Holevo channel, providing a low-dimensional setting for studying symmetry, noise, and capacity together (Roofeh & Karimipour, 2024). Antidegradability, meaning that the receiver's output can be simulated from the environment's output, certifies a zero-capacity region. The investigation combined constructive capacity lower bounds with candidate converse arguments, complementary-channel comparisons, and exact matrix inequalities intended to locate the boundary. A vanishing lower bound alone does not establish zero capacity.

**Quantum-capacity superactivation.** Superactivation is the phenomenon $Q(\mathcal{N}) = Q(\mathcal{M}) = 0$ but $Q(\mathcal{N} \otimes \mathcal{M}) > 0$: two individually zero-capacity channels become useful together. Smith and Yard established its existence, showing that an isolated channel's capacity does not completely describe its usefulness as a communication resource (Smith & Yard, 2008). Our investigations asked which zero-capacity mechanisms can be combined in this way and how strongly activation can occur. Candidate pairs included channels with positive-partial-transpose Choi matrices and channels constrained by receiver-versus-environment information orderings. The work involved joint input constructions, coherent information, complementary-channel representations, and upper bounds. The objectives were new activation conditions or quantitative limitations, rather than another proof of the already known existence of superactivation.

**Bound entanglement with negative partial transpose.** Bound entanglement refers to entangled states from which pure entanglement cannot be distilled by local operations and classical communication. The question is whether such undistillable states can have a partial transpose with a negative eigenvalue, separating the detection of entanglement from its extraction (DiVincenzo et al., 2000). Our investigations focused on Werner states $\rho_{d,\alpha} = (I + \alpha F)/(d^2 + \alpha d)$, where $F$ exchanges the two subsystems, in the candidate range $-1/2 \le \alpha < -1/d$. The concrete milestone was three-copy undistillability, initially at $d = 4$ and $\alpha = -1/2$. This requires nonnegative expectations of $(\rho_{d,\alpha}^{\Gamma})^{\otimes 3}$ on all vectors of Schmidt rank at most two, where $\Gamma$ denotes partial transpose. The work examined rank-constrained inequalities, partial traces, positive decompositions, and finite-dimensional searches. A finite-copy result would be an intermediate step, not a resolution for arbitrarily many copies.

## A.2 Quantum Error Correction

**Circuit-distance stabilization.** The four error-correction problems concern CSS codes with parameters $[[n, k, d]]$ and single-ancilla, CNOT-based syndrome extraction. With unit-weight elementary

faults and fixed ideal initialization and final-readout boundaries, circuit distance is the minimum number of faults that change a logical observable without triggering a detector. It differs from static code distance because faults can propagate through the extraction circuit (Beverland et al., 2024). For a fixed schedule $M$ repeated for $R$ rounds, the investigation asks for the eventual distance $d_\infty(M)$ and the earliest round $R_{\mathrm{stab}}(M)$ after which the distance remains constant. The objective is a finite-state description and an explicit finite window sufficient to determine these quantities, together with examples explaining the limiting fault mechanisms. Stabilization need not mean reaching the static distance $d$.

**Circuit versus effective distance.** An ancilla fault can propagate to several data qubits, producing a correlated residual or hook error. The effective distance $d_{\mathrm{eff}}(M)$ is the extended-code distance obtained by treating the residual-error set as additional unit-cost error mechanisms. This representation is smaller than the full spacetime detector model but does not retain every intermediate detection constraint (Strikis et al., 2026). The problem asks when a minimum effective logical error can be realized by an equally short, undetected circuit-level fault configuration, and when additional faults are required. The investigated criteria use spacetime lifts with fixed boundaries; the graph-like formulation is interpreted separately in each CSS sector, with $R \geq d$ in the recorded criterion. The objective is to characterize equality of the distances and explain strict inequalities without identifying the two models.

**Optimal circuit distance of the Gross code.** The Gross code encodes 12 logical qubits in 144 physical qubits and has static distance 12. The investigated optimization fixes its weight-six stabilizer generators and permits one ancilla per check, each required CNOT exactly once, and at most one two-qubit gate per qubit at each time step. Interleaving $X$- and $Z$-check gates is allowed; additional flag ancillas, cat states, and changes of generators are excluded. Published work rules out distance 12 for non-interleaved schedules and proposes a circuit conjectured to have distance 11 (Strikis et al., 2026). The objectives are to certify that candidate and determine whether a legal interleaved schedule can attain 12, or whether an upper bound of 11 holds throughout the allowed schedule class. This tests whether gate ordering alone can recover the code's full protection.

**Complexity of distance-preserving schedules.** This problem asks how difficult it is to choose a schedule preserving effective distance, rather than to compute distance after a schedule has already been supplied. The input specifies CSS check matrices, a promised static distance $d$, permitted interactions and scheduling constraints, and optionally a depth bound. The decision question is whether some legal schedule satisfies $d_{\mathrm{eff}}(M) = d$. The investigation seeks complexity classifications for fixed and input-dependent $d$, together with tractable restricted cases. Promising the static distance separates scheduling difficulty from the difficulty of determining code distance itself. Gate-conflict restrictions and permitted local orders remain explicit. Preserving full circuit distance is a separate variant, not an interchangeable formulation.

### A.3 Quantum Query Complexity

**Quantum $k$-sum with small alphabets.** Given oracle access to $N$ elements of a finite abelian group $G$, the $k$-sum problem asks whether $k$ distinct input positions contain values summing to a prescribed target $t \in G$. Quantum query complexity measures the input queries needed to decide this with bounded error. Tight lower bounds are known when the alphabet $G$ is sufficiently large (Belovs & Špalek, 2012). Our investigation concerned how the query requirement changes as the alphabet shrinks, and which parts of existing arguments depend on its size. The work examined reductions from related collision-finding problems, exact oracle simulation costs, and the effects of group structure and target choice. These questions matter because restricting the available input values can substantially change an algorithmic bound.

**Set equality and collision.** Set equality gives oracle access to two injective functions $f, g : [n] \to [q]$, with the promise that their images are either equal or disjoint, and asks which case holds. The related collision problem distinguishes a one-to-one function from a two-to-one function. Both have established quantum query bounds, including negative-weight adversary constructions (Belovs & Rosmanis, 2013). Our target was an explicit optimal adversary witness for set equality at the minimal alphabet size $q = 2n$: a matrix whose norm and query-dependent norm bounds certify the desired

lower bound. The investigation studied symmetry reductions, representation-theoretic blocks, and exact matrix normalization. Collision continuations are grouped with set equality because they arose from the same certificate construction and attempted transfers of its guarantees.

### A.4 Grothendieck Theory

**Finite-dimensional Grothendieck bounds.** For a real matrix $A$, compare the maximum of $\sum_{ij} A_{ij} s_i t_j$ over signs $s_i, t_j \in \{-1, 1\}$ with the maximum of $\sum_{ij} A_{ij} \langle u_i, v_j \rangle$ over unit vectors in $\mathbb{R}^d$. The supremum of the vector-to-sign ratio over nonzero matrices is the order-$d$ real Grothendieck constant. This comparison connects discrete optimization with correlation-type Bell inequalities, where vector strategies can exceed classical sign bounds. The investigations included dimension-four constructions and an eight-dimensional candidate based on the $E_8$ root system. The latter involved exact root-poset enumeration and the arguments needed to connect it to claimed Bell values. Our dimension-four investigation examined a specific symmetric matrix, an exact rank-four vector construction, and candidate certificates for its classical optimum. The target required a global upper bound over all sign choices. Exact values for selected sign patterns or a restricted support partition were intermediate results, not substitutes for that unrestricted bound.

**Group-orbit Bell matrices.** This investigation studied Bell coefficient matrices with finite-group actions on their rows and columns that preserve the matrix. Such symmetry can simplify optimization of $\sum_{ij} A_{ij} \langle u_i, v_j \rangle$ over unit vectors, relating attainable values to singular spaces and normalization conditions (Epping et al., 2013). The targets included characterizing optimal vector strategies, finding their minimum vector dimension, quantifying losses under dimension restrictions, and certifying classical sign values for structured families. Constructions used symmetric vector collections and diagonal modifications of their associated matrices. The work also examined whether statements about symmetry-averaged solutions extend to all optimizers. Vector dimension and physical Hilbert-space dimension are distinct quantities, so claims about physical quantum resources require separate justification.