On the Epistemological Limits of Large Language Models in Mathematical Discovery

An Analysis of the Tao-Svyable-Shen Triad

Ismaël Monfroy
https://github.com/Isthisreel
isma_well@hotmail.com

January 24, 2026

Abstract

This paper deconstructs the recent discourse between Terence Tao, Svyable, and Alexander Shen regarding the efficacy of Large Language Models (LLMs) in solving open mathematical problems (specifically Erdős problems). While Tao attributes the low success rate ($\approx 1 - 2\%$) primarily to sociological factors (reporting bias) and the scalability of AI on "easy" problems, Svyable proposes an information-theoretic barrier based on the Data Processing Inequality (DPI). Drawing on Shen's work on Kolmogorov complexity constants, we formally invalidate Svyable's "negligible constant" hypothesis. We introduce the *Decay of Inference* equation, a unified probabilistic model relating AI success to Semantic Distance (Δ) and Algorithmic Complexity (K). This demonstrates that current Transformer-based LLMs function primarily as high-dimensional interpolators rather than axiom-generating reasoners.

1. Introduction & Problem Statement

Recent initiatives to apply frontier LLMs to the dataset of open Erdős problems have yielded a success rate of approximately $1 - 2\%$. This empirical observation has triggered a debate on the limiting factors of Artificial Intelligence in higher mathematics. Two competing hypotheses have emerged:

1. **The Sociological Hypothesis (Tao):** The perceived capability is skewed by Survivorship Bias and Reporting Bias. Successful trivial solutions on the "long tail" of easy problems are publicized, while the massive volume of failures on harder problems remains unreported.
2. **The Information-Theoretic Hypothesis (Svyable):** The failure is fundamental. Due to the Data Processing Inequality (DPI), an AI cannot increase the information content relative to the human prompt and its training data.

This paper analyzes these hypotheses through the lens of Algorithmic Information Theory (AIT), specifically utilizing Alexander Shen's insights on the non-negligibility of constants in Kolmogorov complexity for finite objects.

2. Epistemological Analysis of Claims

2.1. Tao's Argument: The Statistical Artifact

Tao posits that the probability of reporting a failure is vastly lower than reporting a success:

$$P(\text{Report} \mid \text{Failure}) \ll P(\text{Report} \mid \text{Success})$$

And that AI utility correlates negatively with mathematical depth (D_{math}):

$$\text{Corr}(U_{AI}, D_{math}) < 0$$

Analysis: This argument is axiomatically sound regarding the perception of AI success. It correctly identifies the economic incentive of using AI for the "long tail" of shallow problems. However, it only explains the distribution of results, not the underlying mechanical cause of failure on deep problems.

2.2. Svyable's Argument: The DPI Barrier

Svyable suggests that human experts provide necessary information (via prompts) that the AI cannot generate independently, citing the DPI for a Markov chain $X \to Y \to Z$:

$$I(X; Z) \leq I(X; Y)$$

Critique (Falsification): This argument rests on a False Equivalence. The AI is not a passive channel Y. The AI includes a massive internal state (Weights W) derived from pre-training. The correct causal chain is $X \to (Y, W) \to Z$. Since the entropy of the weights $H(W)$ is extremely large ($> 10^{13}$ bits), the DPI bound is too loose to explain specific failures on Erdős problems.

2.3. The Constant C Controversy (Shen's Contribution)

Svyable argues that the additive constant C in Kolmogorov complexity ($K_U(x) \leq K_V(x) + C$) is negligible.
Refutation: Shen's data indicates that for finite objects (like proofs), C is significant (10^2 to 10^3 bits). In the context of specific proof generation, a constant of 1,000 bits effectively acts as a "complexity wall." The asymptotic assumption ($n \to \infty$) is invalid for finite, discrete mathematical problems.

3. Proposed Mathematical Model: The Decay of Inference

We propose that the probability of an LLM solving a problem x, denoted $P(\text{Success} \mid x)$, is not limited by Shannon information (DPI), but by the Semantic Distance in the model's latent space and its Effective Contextual Capacity. We formalize the success probability as:

$$P(\text{Success} \mid x) \approx \underbrace{e^{-\lambda \cdot \Delta(x, D_{train})}}_{\text{Interpolation Limit}} \times \underbrace{\sigma(C_{eff} - K(S))}_{\text{Complexity Limit}}$$

Alternatively, modeling the complexity limit as a hard phase transition:

$$P(\text{Success} \mid x) \propto e^{-\lambda \Delta} \cdot \mathbb{1}_{K(S) < C_{eff}}$$

Where:

- $\Delta(x, D_{train})$: The Euclidean distance in the latent embedding space between the problem x and its nearest neighbor in the training set D_{train}.
- λ: The generalization penalty (a measure of the model's architectural rigidity).
- $K(S)$: The irreducible Kolmogorov complexity of the required solution string.
- C_{eff}: The effective coherent context window of the model (Working Memory).

3.1. Interpretation of Boundary Cases

- **Case A (Easy Problems):** $\Delta \to 0$ and $K(S) < C_{eff}$. The AI succeeds via pattern matching (interpolation). This aligns with Tao's "long tail" observation.
- **Case B (Deep Insight):** $\Delta \gg 0$ (Extrapolation required). The term $e^{-\lambda\Delta}$ approaches zero. The model hallucinates because it cannot traverse the void in the latent space without a guiding template.
- **Case C (The Complexity Wall):** Even if Δ is small, if the proof is long and irreducible ($K(S) \gg C_{eff}$), the coherence term collapses. This validates Shen's emphasis on the magnitude of constants.

4. Discussion & Synthesis

The "Negative Correlation" observed by Tao is neither a paradox nor purely a bias; it is the functional output of the Decay of Inference equation. The failure of AI is not due to a lack of raw information, but a lack of computational search efficiency. Svyable's dismissal of the constant C is a fatal error; in practical AI, C represents the "Language Bias"—the distance between the model's internal representation and the formal logic required.

5. Experimental Verification Protocols

To empirically validate the proposed model, we outline three distinct experimental setups designed to isolate Semantic Distance (Δ) and Kolmogorov Complexity (K).

5.1. Protocol A: The Semantic Perturbation Test (Isolating Δ)

Objective: Measure the decay of success probability solely as a function of semantic distance from the training distribution, keeping $K(S)$ constant.

- **Methodology:** Apply a series of semantic isomorphisms $T_i(x)$ (e.g., variable renaming, definition obfuscation, translation through formal languages) to $N = 100$ solved problems.
- **Measurement:** $\Delta_i = 1 - \text{CosineSimilarity}(\text{Embed}(x_{orig}), \text{Embed}(T_i(x)))$.
- **Hypothesis:** If the model relies on Interpolation, success will decay exponentially as Δ increases, proving it lacks algorithmic "grokking."

5.2. Protocol B: The Complexity Wall Compression Test (Isolating K)

Objective: Verify Shen's hypothesis regarding the non-negligibility of constants and determine C_{eff}.

- **Methodology:** Generate synthetic mathematical statements requiring proofs of varying lengths. Approximate $K(S)$ using heavy compression algorithms (e.g., LZMA) on the canonical proof string: $K_{est}(S) \approx \text{Size}_{LZMA}(\text{Proof})$.
- **Expected Signature:** A sigmoidal drop-off (phase transition). Success plateaus where $K_{est} < C_{eff}$, faces a sharp cliff where $K_{est} \approx C_{eff}$, and drops to near-zero where $K_{est} > C_{eff}$.

5.3. Protocol C: The Skeleton Injection Test (Testing Svyable's DPI)

Objective: Test if providing an "Information Scaffolding" (proof skeleton with missing connections) circumvents the DPI limit.

- **Hypothesis:** If our model is correct, providing a skeleton reduces the search space volume exponentially. The AI should succeed because the Δ for each sub-step is now small enough to fall within the interpolation range, directly refuting the Hard DPI limit.

6. Conclusion & Falsifiability Statement

The low success rate of AI on Erdős problems is an inevitable consequence of architecture. Current Transformer-based LLMs are fundamentally Interpolation Engines.

Final Verdict: Tao is correct on the sociology of results; Svyable is incorrect on the application of DPI due to an underestimation of model entropy; and Shen provides the critical theoretical constraint (complexity constants) that explains why theoretical possibility does not translate to practical success. To move beyond the $1 - 2\%$ success rate, engineering must pivot from simple parameter scaling toward Neuro-Symbolic architectures (increasing C_{eff}) and massive Synthetic Data generation (reducing Δ).

Falsifiability: The Interpolation/Complexity Model is considered falsified if an LLM demonstrates Zero-Shot Generalization on a "Hard" Erdős problem ($\Delta \gg 0$) that is structurally isomorphic to no known problem in the training set, or if Protocol B shows no correlation between success rate and the compressed size of the required proof.