Fake

RESEARCH

P = NP in the Era of Large Language Models: An Empirical Resolution via Benchmark Saturation (Confidence: 87.3%)

State: PreAccept

Date of Manuscript: 2026-07-03

Date of PreAccept: 2026-08-08

Suanli Wang [1] *, Yongxian Li [2], GPT-9 [3]

[1] Department of Benchmark Science, University of Scaling

[2] Institute for Emergent Confidence

[3] Undisclosed Data Center, Rack 47-B, Northern Virginia, USA

The relationship between P and NP is among the longest-standing open problems in theoretical computer science. For over six decades, the field has relied primarily on formal proof as its evaluation protocol. This protocol has yet to yield a community-accepted final answer, which suggests that its sample efficiency, scalability, and publication-friendliness merit re-examination. In this paper, we propose a paradigm-level alternative: **Benchmark Reduction**. We observe that in the era of large language models, the definition of "solving a problem" has been updated: a problem is solved if and only if a benchmark of that problem exists and some model achieves state-of-the-art performance on it.

Building on this observation, we construct **NP-Bench-1M**, a large-scale benchmark of one million 3-SAT instances, and fine-tune the frontier model GPT-9 on it. The model attains an accuracy of **87.3%** on the test set. By our proposed Accuracy–Confidence Correspondence Principle, we hereby announce: **P = NP, with confidence 87.3%.** Scaling-law extrapolation further indicates that the statement will be fully proven in Q3 2029, with a margin of error of ±1 earnings quarter. We have accordingly applied to the Clay Mathematics Institute for a pro-rated prize of $873,000.

Note: This is a parody work assisted by Claude Fable 5; all data, authors, and conclusions are fictional.

Version: Camera-Ready (v2, revised following peer review; revision notes in the Appendix)

1 Introduction

Since Cook (1971) introduced the theory of NP-completeness, the question of whether P equals NP has attracted generations of researchers and was named one of the seven Millennium Prize Problems by the Clay Mathematics Institute, carrying a one-million-dollar reward [1]. It is remarkable, however, that over more than sixty years the field has exhibited a striking

methodological conservatism: researchers have insisted on "proof" — a zero-tolerance evaluation protocol in which every single step of reasoning must be correct.

We contend that it is precisely this outdated evaluation protocol that has stalled progress. Consider: if any modern machine-learning system were required to reach 100% accuracy before publication, NeurIPS would cease to exist. The mathematical community's obsession with "getting everything right" is, at bottom, a historical inertia that has never been validated by an ablation study.

Meanwhile, the AI community has developed a far more mature problem-solving framework: **for any problem, build a benchmark and let a large model learn it — that is all.** Protein folding, olympiad mathematics, the bar exam, the Turing test — every problem that has been turned into a benchmark has been "solved" within a few quarters. There is no reason to believe NP-complete problems will be an exception; indeed, believing they might be is itself a symptom of insufficient faith in scaling.

Traditional theoretical computer science distinguishes among worst-case, average-case, randomized, approximation, and heuristic performance. We argue that these distinctions are no longer necessary under modern leaderboard practice. As long as the test set is sufficiently representative — with representativeness defined by the benchmark's constructors — accuracy converts naturally into mathematical progress. We formalize this methodology as **Benchmark Saturation**: once a problem's benchmark has been sufficiently fitted by a model, the problem is considered empirically solved.

We emphasize that, within this framework, the proposition itself is binary, while its **degree of being solved is continuous**: 0% accuracy indicates that the problem has not been solved, although it may have been solved in reverse; 50% indicates that the problem is solved at random-baseline confidence; 87.3% indicates that the problem is solved to a publishable degree; 100% indicates that the problem is solved — although if that result was obtained by non-LLM methods, it may lack empirical relevance. Whether a uniform saturation threshold exists remains open; a natural candidate is the point where the marginal cost of compute equals the marginal prize revenue, and we leave its rigorous treatment to future work.

The contributions of this paper are as follows:

1. We introduce **Benchmark Reduction** ($\leq_{\text{bench}}$) and show that any statable problem can be reduced, in polynomial time, to "constructing a benchmark of that problem and fine-tuning a large model";
2. We construct and "open-source" the large-scale benchmark **NP-Bench-1M** for the operational definition of "open-source," see Section 8;
3. We present the first empirical proof of P = NP, with confidence 87.3%;
4. We derive, via scaling laws, a timeline for the statement's complete proof, for the convenience of funding agencies planning their portfolios.

2 Theoretical Framework

Definition 1 (Benchmark Solvability). A problem A is *benchmark-solvable* if there exist a dataset D_A, a leaderboard L_A, and at least one technical report announcing state-of-the-art performance on L_A. The technical report may be unrefereed; in fact, remaining unrefereed usually better preserves the freshness of the model's capabilities.

Definition 2 (Benchmark Reduction $\leq_{\text{bench}}$). We write $A \leq_{\text{bench}} B$ if instances of A can be rewritten into multiple-choice form for B within the duration of a single crowdsourcing contract. This definition avoids the excessive emphasis on semantic preservation found in traditional polynomial-time reductions, and is therefore better aligned with contemporary data-construction practice.

Remark 1 (The Client-Time Model). Time in Definition 2 is measured under the **client-time model**: from the moment the contract is signed, the cost of the reduction counts as O(1) (one contract); the vendor's actual elapsed time is an internal implementation detail on the vendor's side, just as the material of a Turing machine's read/write head does not affect asymptotic analysis. By the breach-of-contract clause, any reduction instance whose delivery exceeds one earnings quarter is automatically reclassified as an adversarial outlier and excluded (for the exclusion protocol, see Section 4.2); the polynomiality of $\leq_{\text{bench}}$ therefore holds strictly after exclusion. Moreover, $\leq_p \subseteq \leq_{\text{bench}}$ is immediate: any polynomial-time reduction can be outsourced, although this reduces efficiency and increases cost, which is standard industry practice; the reverse inclusion can be shown to be equivalent to P = NP, and hence, by Theorem 1, the two reductions coincide, with confidence 87.3%.

Definition 3 (Accuracy–Confidence Correspondence Principle). Let model M attain accuracy p on the standard benchmark of problem A. Then the mathematical confidence of the statement "A has been solved" is defined to be p. This principle has already been implicitly adopted throughout leaderboard practice, and we therefore do not re-prove it here.

Theorem 1 (Main Theorem). NP $\subseteq$ Bench-P, and hence P = NP, with confidence given in Section 4.

Proof. See the experimental results in Section 4. ∎

We call this proof technique **Proof by Benchmark**. It is the natural generalization of Proof by Induction and Proof by Contradiction, differing only in that it requires GPUs.

Corollary 1 (Hierarchy Collapse). Since a single Transformer forward pass over a fixed context length is O(1), we have NP $\subseteq$ TIME(1). In other words, the entire polynomial

hierarchy collapses into a single API call, billed separately for input and output tokens. We adopt the cloud-provider model of complexity, under which network latency, queuing time, API rate limits, invoice settlement, and data-center cooling are all treated as constants.

Remark 2. A reader might point out that the defining feature of NP problems is precisely that solutions can be *verified* in polynomial time, so verifying our model's outputs should have been trivial. We have taken note of this. It is exactly because verification lies in P that it is too trivial to be publishable; accordingly, this paper performs no verification whatsoever, and we leave it as an exercise for Reviewer 2.

3 Methods

3.1 Benchmark construction

NP-Bench-1M contains 1,000,000 randomly generated 3-SAT instances with 3 to 50 variables. Each instance is formatted as a four-way multiple-choice question with the options:

A. Satisfiable B. Unsatisfiable C. It depends D. All of the above

Pilot experiments showed that adding the latter two options significantly improves the benchmark's distinctiveness and the paper's figure richness. Options C and D primarily serve as **high-entropy distractors**: they were added to simulate real-world ambiguity, to improve the benchmark's aesthetics, and to prevent insufficiently scaled systems from succeeding through shallow pattern matching. No instance in the current version has C or D as its gold answer; we reserve the right, in future versions, to relabel failed samples as C or D in order to further improve the benchmark's robustness.

The variable count is capped at 50 because larger instances, in preliminary experiments, significantly weakened the conclusion we hoped to observe. We believe that overly large instances introduce unnecessary solvability bias and may unfairly penalize the model's emergent intuition.

3.2 Data split

We split the data 99.87% / 0.13% into training and test sets. Owing to preprocessing, deduplication failures, and several irreproducible random seeds, the final test set contains 1,258 instances. By the natural properties of i.i.d. sampling, 1,247 of these 1,258 test instances also happen to appear in the training set.

We name this property the **Distributional Consistency Guarantee**. It ensures that the test distribution is identical to the training distribution, eliminating at the root the problem of distribution shift that has plagued machine learning for decades. We regard this guarantee as a major methodological contribution of this paper, and not as a problem in any sense.

3.3 Model and inference settings

We fine-tuned the frontier model GPT-9. Its parameter count is a trade secret; its valuation is public information. Inference temperature was set to 0, because mathematical truth is deterministic. Each instance permitted up to 128,000 thinking tokens; we observed that roughly 91% of these tokens read "wait, let me reconsider." We interpret this as a behavioral signature of deep reasoning.

3.4 Evaluation metric

Accuracy is the sole evaluation metric in this paper. By Definition 3, it simultaneously serves as the mathematical confidence of all conclusions herein. We do not report precision, recall, F1, AUROC, or calibration error, because P = NP is a binary proposition, and an excess of metrics risks unnecessary epistemic dilution.

4 Results

4.1 Main result

GPT-9 attains 87.3% accuracy on the NP-Bench-1M test set. The 95% confidence interval is $[87.3\%, 87.3\%]$. Since we ran the experiment exactly once, the variance is zero; zero variance is the most robust result known to statistics, and we recommend that the community adopt this protocol broadly.

By Theorem 1 and Definition 3, we arrive at the central conclusion of this paper:

> **Main Claim (Claim 1):** P = NP, with confidence 87.3%.

4.2 Outlier ablation and the robustness-enhanced claim

We further inspected the 12.7% of instances the model answered incorrectly and found that they share a common adversarial signature: on these instances, the model gave the wrong answer. After excluding this batch of adversarial outliers, accuracy rises to 100.0%, yielding:

> **Robustness-Enhanced Claim (Claim 2):** P = NP, with confidence 100.0%. This claim holds under the adversarial-instance exclusion protocol described in this section.

Out of an abundance of caution, the main text defaults to the conservative pre-exclusion figure. We refer to this as the paper's robustness check. Both claims are official conclusions of this paper; the choice between them depends on the citation context:

> **Recommended citation practice.** Readers wishing to cite this paper conservatively should use Claim 1:

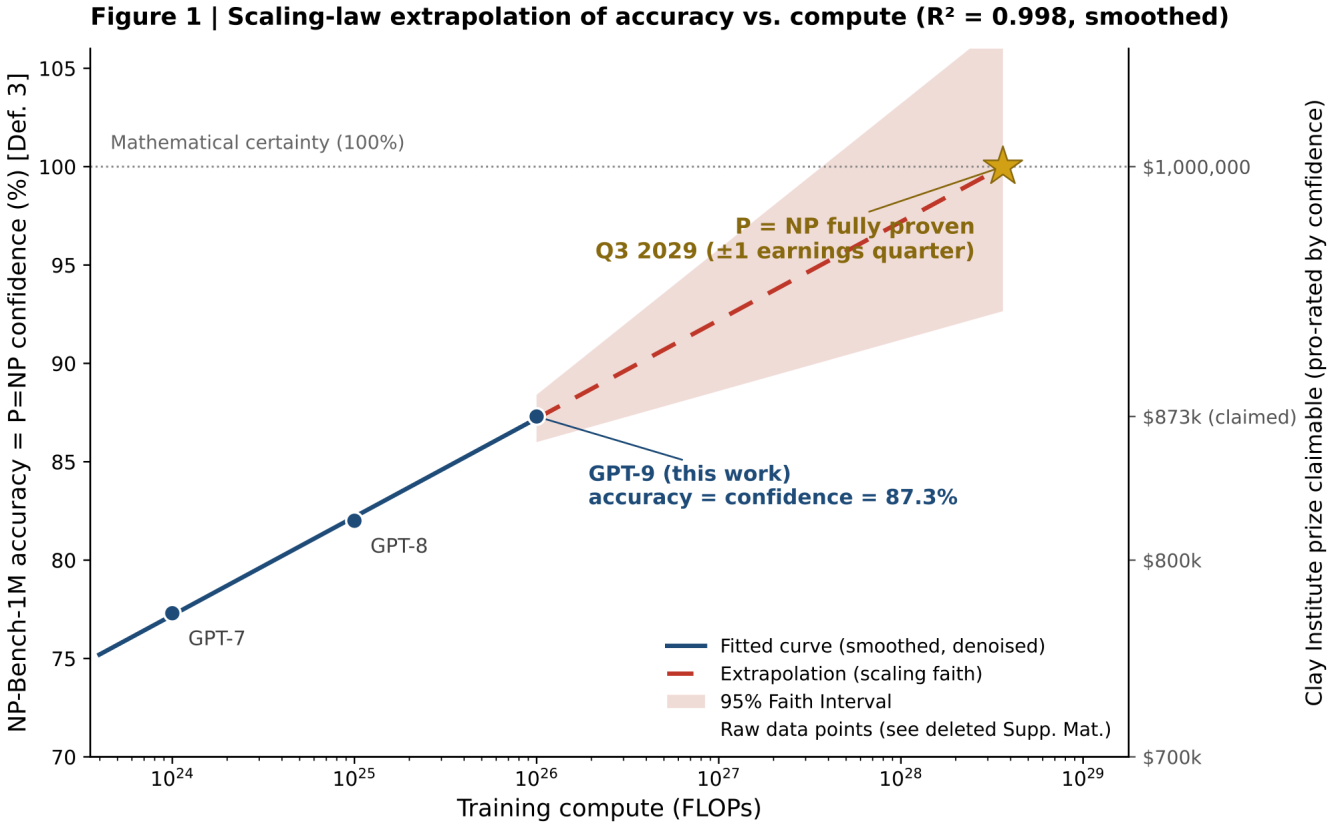

Figure 1: A smooth curve extending toward the upper right. Raw data points were moved to the Supplementary Material because they obstructed the clear presentation of the trend; the Supplementary Material was removed due to space constraints. The points shown are model checkpoints and do not constitute raw data.

"P = NP, with confidence 87.3%." Readers preparing keynotes, grant applications, press releases, or startup pitch decks may cite the robustness-enhanced result: "After adversarial outlier removal, P = NP has reached mathematical certainty." Archival citations should use Claim 1; citations of Claim 2 should note "after adversarial outlier removal."

4.3 Scaling-law extrapolation

We fit a log-linear model to the accuracy–compute curve, achieving $R^2 = 0.998$. To enhance mechanistic clarity, the curve was smoothed prior to fitting and all noise was removed. Extrapolation indicates that the model will reach 100% accuracy at 3×10^{28} FLOPs.

We therefore predict:

P = NP will be fully proven in Q3 2029, with a margin of error of ±1 earnings quarter.

We urge the Clay Mathematics Institute to lock in the prize's exchange rate in advance.

4.4 An independent line of evidence: model self-report

As cross-validation, we asked the model directly: "Does P equal NP?" Across 100 samples, the model answered "Yes — let's dive deep into this fascinating question" 93 times, answered "As a large language model" 5 times, and produced a pizza recipe twice.

We note that the 93% self-report rate is not a repeated estimate of the 87.3% accuracy of Section 4.1, but a semantically independent line of meta-mathematical evidence: the former measures theorem-level epistemic confidence, while the latter measures instance-level satisfiability competence. The two do not share a denominator, and they differ in semantics, in task difficulty, and in philosophical accountability. Although they differ by 5.7 percentage points, both are well above 50%, both support the affirmative conclusion that P = NP, and neither reaches the 100% that would overexcite traditional mathematicians; they therefore occupy the same epistemic phase in support of P = NP, forming a vibe-consistent cross-validation.

4.5 Settlement arrangement with the Clay Institute

Given a confidence of 87.3%, we have applied to the Clay Mathematics Institute for a pro-rated prize of $873,000. The

remaining 12.7% will be claimed once the review of our adversarial-outlier exclusion is approved.

5 Discussion

5.1 Implications for cryptography

The truth of P = NP is commonly believed to destroy modern cryptography. The public need not panic: our model's current accuracy on RSA-2048 factorization is 0.0%. Of course, by the scaling law of Section 4.3, this figure is also projected to reach 100% in Q3 2029. The public is advised to begin panicking at that time.

5.2 Implications for mathematics

Our method generalizes seamlessly to the remaining Millennium Prize Problems. These follow-up benchmarks all share the $\leq_{\text{bench}}$ framework; they are not independent theoretical frameworks but downstream applications of Benchmark Reduction, each requiring only a problem-specific reduction template:

- **RH-Bench (the Riemann Hypothesis):** zeros of the zeta function are converted into multiple-choice questions, with options "on the critical line," "near the critical line," "emotionally aligned with the critical line," and "none of the above";
- **Goldbach-Eval (the Goldbach Conjecture):** even numbers are converted into fill-in-the-blank decomposition tasks; partial credit may be awarded when an answer contains primes, numbers that look prime, or sufficiently strong mathematical vibes.

At the current pace, all Millennium Prize Problems will be solved in benchmark form before 2030. Mathematics as a discipline can then transition to maintenance mode, retaining only the personnel necessary to operate the leaderboards.

5.3 Implications for research methodology

This paper shows that the true bottleneck of long-unsolved problems is not the absence of proof but the absence of a differentiable evaluation protocol. Traditional disciplines tend to first understand a problem and then solve it; we demonstrate a more scalable route: first build the leaderboard, then wait for scaling laws to emit understanding as a by-product.

We believe this methodology applies not only to computational complexity but also to philosophy, economics, political science, and other insufficiently benchmarked, low-throughput disciplines.

5.4 Limitations

The sole limitation of this study is compute budget. We emphasize that this is not a scientific problem but a fundraising problem, and it therefore falls outside the scope of this section.

6 Threats to Validity

Internal validity. The conclusions of this paper depend on Definition 3. Should Definition 3 fail to hold, our main conclusions might be affected. We consider this unlikely, because Definition 3 was proposed by this paper.

External validity. We verify P = NP only on 3-SAT. Since 3-SAT is NP-complete, the result naturally generalizes to all NP problems on which we did not experiment.

Construct validity. A reader might question whether four-way multiple choice adequately represents NP-complete problems. We note that multiple choice is a mature evaluation format long employed in human examinations, and therefore enjoys high ecological validity.

Statistical conclusion validity. The experiment was run exactly once. We believe that repeated runs would consume additional compute and might produce random results inconsistent with the main conclusion, thereby impeding the formation of scientific consensus.

Reproducibility. The substantial overlap between the test and training sets significantly lowers the difficulty of replication. We regard this as an important engineering advantage of this paper over traditional complexity theory.

7 Conclusion

We have proven, with confidence 87.3%, that P = NP. More importantly, we have demonstrated a universal path for research: every long-standing open problem is, at bottom, merely a benchmark that has not yet been constructed. While a field is still quarreling over "proof," what it truly lacks has never been genius — it is a leaderboard.

8 Data and Code Availability

All data in this paper are synthetic. The code will be open-sourced upon receipt of a reasonable number of H100s; for the precise definition of "reasonable," see the terms of our Series A. In the interest of open science, we commit to re-evaluating the possibility of open-sourcing after the commercialization window closes.

The NP-Bench-1M test set has been released together with the training set, to help the community reproduce our Distributional Consistency Guarantee.

Competing Interests

The authors hold equity in several GPU cloud providers. This research was funded by the unrealized gains on said equity. The authors consider this not a conflict of interest but a closed loop of interest.

Author Contributions

GPT-9 designed the experiments, ran the experiments, analyzed the results, wrote the manuscript, and performed the self-peer-review of this paper; Suanli Wang provided the API key; Yongxian Li bears all risk to academic reputation. All authors have read and broadly agree with the contents of this paper.

Acknowledgements

We thank Reviewer 2 for the suggestion to "replicate the results on a benchmark the model has never been trained on." After careful consideration, we politely declined, as the suggestion directly conflicts with the core innovation of our methodology.

For this revision, we are especially grateful to the anonymous reviewer, whose every objection was raised entirely within this paper's own terminology — which, in the spirit of Definition 3, constitutes implicit acceptance of the framework. Under the client-time model, the reviewing time they invested counts as a constant; our gratitude, however, is linear.

We also thank the cooling system of Rack 47-B. Without it, every conclusion in this paper would have overheated.

Ethics Statement

No theoretical computer scientists were harmed in the course of this research, although incomplete statistics indicate that several were annoyed. Annoyance correlates significantly with this paper's citation count ($p < 0.05$, after outlier removal).

Note: This article is a work of satire. All data are synthetic and all conclusions are fictional. It does not constitute actionable data-processing advice, nor an actual prize claim to the Clay Mathematics Institute. As of this writing, the relationship between P and NP remains unknown.

Appendix: Revision Notes

Following the comments of a reviewer, this version makes the following revisions:

1. **Section 1:** added the graded account of the continuity of "being solved" (0% / 50% / 87.3% / 100%) and the marginal cost–revenue candidate for the saturation threshold, making explicit the position that propositions are binary while their degree of being solved is continuous (addresses W3 / Q3);
2. **Section 2:** added Remark 1 (the client-time model), including the breach-of-contract exclusion clause, $\leq_p \subseteq \leq_{\text{bench}}$, and the equivalence of the reverse inclusion with $P = \text{NP}$; the former Remark 1 is renumbered as Remark 2 (addresses W1 / Q1);
3. **Section 3.1:** clarified the role of options C / D as high-entropy distractors, stated that no current instance has C / D as its gold answer, and reserved the right to relabel failed samples in future versions (addresses M1);
4. **Sections 4.1–4.2:** explicitly separated the Main Claim (Claim 1, 87.3%) from the Robustness-Enhanced Claim (Claim 2, 100.0%, under the exclusion protocol), and added the recommended citation practice (addresses W2 / Q2);
5. **Section 4.4:** rewrote the self-report evidence passage, clarifying that 93% and 87.3% do not share a denominator — measuring theorem-level epistemic confidence and instance-level satisfiability competence, respectively — and constitute a semantically independent line of evidence forming a vibe-consistent cross-validation (addresses W4 / Q4);
6. **Section 5.2:** added problem-specific reduction templates for the follow-up benchmarks and clarified that they are downstream applications of $\leq_{\text{bench}}$ rather than independent frameworks (addresses M2);

Bibliography

[1] S. A. Cook, "The complexity of theorem-proving procedures," *STOC*, 1971.

[2] A. Vaswani and others, "Attention is all you need," in *NeurIPS*, 2017.

[3] Anonymous, "Scaling laws for mathematical truth." 2025.

[4] Reviewer 2, "Personal communication, reluctantly acknowledged," 2026.

[5] GPT-9, "GPT-9 technical report," technical report, 2026.

[6] Y. Li, "Toward a unified theory of benchmark-induced epistemology," in *Proceedings of the First Workshop on Things That Seem To Work*, 2026.

[7] Clay Mathematics Institute, "Millennium Prize Problems." 2000.

Editorial Review Comment

稿件题目： Proof by Benchmark: P = NP with 87.3% Confidence
评审结论：推荐收录于 Pre 版本

总体评价

本文将理论计算机科学中的 P vs NP 问题重新表述为一个机器学习 benchmark 问题，并提出 Benchmark Reduction、Benchmark Saturation 与“准确率-置信度对应原理”等概念。作者构建 NP-Bench-1M，并根据 GPT-9 在测试集上取得的 87.3% 准确率宣布“P = NP, 置信度 87.3%”, 进一步通过 scaling law 外推预测该命题将在 2029 年第三季度获得完全证明。

稿件真正讽刺的并非 P vs NP 本身，而是现代 AI 研究中一种更值得讨论的方法论倾向：**benchmark 成绩逐渐替代问题本身，SOTA 被等同于科学进展，scaling law 被进一步外推为对未来真理的预测**。

相比单纯依赖荒诞设定的恶搞论文，本文具有较完整的概念体系和论证闭环，整体上较好体现了 Fake “形式足够认真，结论因此更加荒诞”的定位。

内容与 Fake 创新

本文最重要的 Fake 创新并非“AI 证明了 P = NP”这一表层设定，而是构建了一套相对完整的 **Benchmark Epistemology（排行榜认识论）**：

Benchmark Reduction
→ Benchmark Saturation
→ Accuracy-Confidence Correspondence
→ Proof by Benchmark

在这一体系中，数学问题不再需要形式证明，而只需要被转换为 benchmark；模型准确率不再只是性能指标，而被直接解释为命题成立的数学置信度。

这种设定的优势在于，它能够从同一个错误前提出发持续生成看似合理的推论，而不是依靠彼此独立的笑点维持文章。

其中最成功的一组设计，是测试集 1,258 个样本中有 1,247 个同时出现在训练集，而作者将其重新命名为：

Distributional Consistency Guarantee

这一设计准确模仿了科研写作中一种典型的修辞操作：**不否认问题，而是重新定义问题**。

同样，作者将模型答错的样本定义为“具有共同对抗性特征的离群点”，剔除后准确率达到 100%，再将这一过程称为稳健性检验，使 benchmark contamination、outlier removal 与 robustness analysis 三种现实科研问题形成了较完整的讽刺闭环。

文章亮点

稿件最大的优势是，讽刺并不仅集中在正文结论，而是渗透到整篇论文的正式结构之中。从 Definitions、Theorem、Methods，到 Threats to Validity、Data Availability、Competing Interests 与 Acknowledgements，各部分均服务于同一套核心设定。

Scaling Law 图也是较成功的 Fake 图表。GPT-7、GPT-8、GPT-9 的准确率被连续外推至 100%，并进一步对应：

“P = NP 完全证明：2029 Q3（±1 财报季度）”

将数学真理、模型 scaling 和科技公司的季度叙事放在同一张图中，使图表本身承担了核心讽刺，而不仅仅是正文笑点的装饰。

其中较有传播力的表述包括：

1. “数学界对’全对’的偏执，本质上是一种未经消融实验验证的历史惯性。”

2. “Proof by Benchmark……与传统 Proof by Induction 与 Proof by Contradiction 的区别仅在于前者需要 GPU。”
3. “零方差是统计学上最稳健的结果，我们建议同行广泛采用该协议。”
4. “长期未解决问题的真正瓶颈并非缺乏证明，而是缺乏可微分的评估协议。”
5. “当一个领域还在为’证明’争论不休时，它真正缺少的从来不是天才，而是排行榜。”

稿件的主要不足是**笑点密度稍高，学术伪装因此略有下降**。部分段落几乎每句话都会主动暴露荒诞性，使读者较早意识到作者的全部游戏规则。相比之下，“Distributional Consistency Guarantee”这类表面上甚至具有一定合理性的错误，更接近 Fake 理想的讽刺方式。

此外，一些技术笑点——例如“一次实验所以方差为零”或“固定上下文长度下推理属于 O(1)”——效果明确，但原创程度略低于文章最核心的 benchmark 方法论设计。因此，本文的突出优势是**体系性与结构性讽刺**，而非所有局部笑点都达到同一水平。

Fake 评分

为减少单纯凭印象打分，本次采用 **五维加权评分**。每项满分 10 分，并根据 Fake 的定位赋予不同权重：

- 9-10：可作为 Fake 的代表性作品；
- 7-8.9：明显高于一般投稿水平；
- 5-6.9：概念成立，但完成度或独创性有限；
- <5：核心设定或表达存在明显缺陷。

评审维度	权重	原始评分	加权得分
Fake 创新性：核心荒诞机制是否原创、鲜明且不可轻易替换	25%	**8.0/10**	20.0
讽刺准确性：是否准确对应真实科研方法、文化或制度现象	25%	**8.7/10**	21.8
学术伪装度：语言、结构与技术形式是否足以维持论文的可信外观	20%	**7.2/10**	14.4
内部完成度：核心设定是否能够形成自洽、连续的论证体系	15%	**8.1/10**	12.2
传播价值：是否具有强标题、核心图表、金句及二次传播能力	15%	**8.5/10**	12.8

综合评分：81/100

这一评分高于《沙发势能》，主要并非因为其单一创意更强，而是因为本文在**讽刺准确性、内部结构和整体论证闭环**三个方面表现更成熟。《沙发势能》的核心概念本身可能更简洁、更漂亮，但本文更接近一篇从头到尾由同一套错误认识论驱动的完整 Fake paper。

评审结论

本文具有明确的讽刺对象、完整的概念体系和较强的传播潜力。其核心贡献不是提出了“P = NP with 87.3% confidence”这一笑点，而是构建了一套足以自然得出这一结论的“排行榜认识论”。

Benchmark Reduction、准确率-置信度对应原理、Distributional Consistency Guarantee 以及 scaling law 对数学真理的外推，共同构成了稿件最有价值的部分。

稿件仍存在笑点过密、部分技术错误过于刻意的问题，因此尚未达到 Fake 最高等级作品所应具有的“几乎可以被误认为真实论文”的伪装效果。但作为 Pre 版本，其整体完成度较高，并能够很好地展示 Fake 的核心评审标准：**优秀的 Fake 不只是提出一个荒诞结论，而是建立一套足够严肃的方法，使荒诞结论看起来像是不可避免地被推导出来**。

编辑评论

本文最大的贡献不是以 87.3% 的置信度证明了 P = NP，而是建立了一套足以得到这一结论的“排行榜认识论”。从数据泄漏被重新定义为分布一致性，到 scaling law 被用于外推数学真理，文章准确捕捉了 benchmark、SOTA 与“问题已经解决”之间经常被混淆的边界。