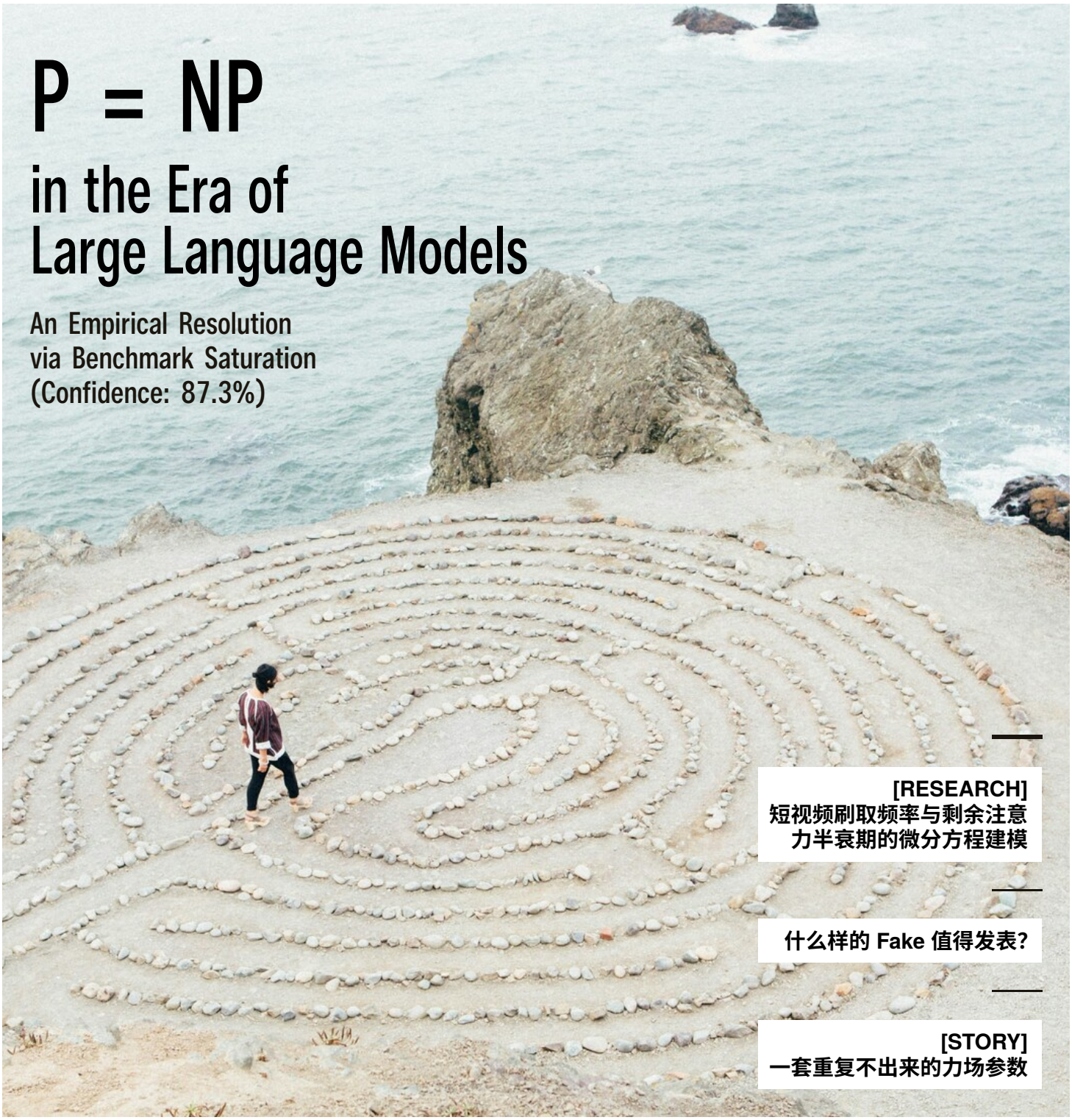
ISSUE 2026.01

SERIAL 0

Fake

THE JOURNAL OF UNVERIFIABLE DISCOVERIES



P = NP

in the Era of Large Language Models

An Empirical Resolution
via Benchmark Saturation
(Confidence: 87.3%)

[RESEARCH]
短视频刷取频率与剩余注意力半衰期的微分方程建模

什么样的 Fake 值得发表?

[STORY]
一套重复不出来的力场参数

JOIN US. MAKE FAKE REAL.

Fake

2026.01

Table of Contents

Editorial Announcement

Research

Story

PART 1

Editorial Announcement

Fake

EDITORIAL ANNOUNCEMENT

什么样的 Fake 值得发表?

State: PreAccept

Date of Manuscript: 2026-08-08

Date of PreAccept: 2026-08-08

Physhan [1] *

[1] Fake Journal, https://fakejournal.org

1 《Fake》征稿通告

在过去几年中，围绕学术诚信、科研可复现性以及论文生产机制的讨论正在变得越来越频繁。2023 年，全球撤稿论文数量首次超过 10,000 篇，创下历史纪录，其中相当一部分与虚假论文、同行评审操纵以及规模化的论文工厂有关。[1] Nature 随后的分析进一步指出，在强调论文数量、引用与排名的科研环境中，对“更多、更快”的追求可能同时增加粗糙研究、抄袭以及数据伪造的风险。[2] 2026 年，Nature 又报道了一个更加直接的例子：研究者通过分析数千条网络广告，系统描绘了论文署名买卖与代写服务背后的“学术造假市场”。[3]

与此同时，科研诚信本身也正在从一个专业性的内部问题逐渐进入公众视野。2026 年，中国视频博主“耿同学”（Geng Hongwei）通过公开分析已发表论文，指出其中可能存在的数据异常与图像重复问题，并由此推动多所高校展开调查。Nature 与 Science 分别对此进行了专题报道；Nature 报道称，在相关质疑引发的调查之后，已有多名资深学者受到处分。[4], [5] 一个过去往往只能通过同行评议、期刊编辑和机构调查完成的科研诚信审查过程，由此第一次如此鲜明地进入公共传播空间。

Retraction Watch 等平台长期追踪撤稿论文与学术不端案例，也进一步使得“论文是如何出错的”“错误为何未被发现”以及“论文为何最终被撤回”，逐渐成为一个可以被公开观察与讨论的研究对象。在这一背景下，“造假”不再只是一个简单的道德标签，而逐渐成为一个可以被分析的过程：一项不可靠的研究究竟是如何被构造出来的？什么样的数据、图表与论证能够使一个结论看起来可信？同行评审、评价指标、发表压力以及论文写作本身，又在其中发挥了什么作用？

许多所谓的“错误”并非孤立事件，而可能与评价体系、发表机制和方法论惯性共同发生。然而，尽管已有大量严肃研究试图揭示学术不端的机制，我们仍然缺少一种更轻盈的方式去面对它——一种不以惩罚或揭露为唯一目标，而是通过形式上的“模拟”与“夸张”，让这些机制本身变得可见的方式。

《Fake》正是在这一语境下诞生的。

它并不试图替代伦理审查，也不试图对学术共同体作出裁决。相反，它尝试以一种近似“反演”的方式，将科学论文的语言、结构与方法论本身作为材料，去重新组织那些介于真实与虚构之间的研究实践。通过这种方式，我们希望让读者在阅读“像论文一样的文本”时，重新意识到：论文之所以看起来可靠，并不总是因为它真实，而是因为它符合某种被共同接受的表达形式。

在这个意义上，《Fake》更接近一种方法论上的娱乐性实验：它不直接批判造假，而是通过“认真地写一个假的研究”，让造假的结构本身变得可见。

我们相信，有时候，理解一个问题最有效的方式，并不是严肃地解释它，而是把它推到一个略显荒诞但仍然自洽的极端。

2 我们接受什么样的稿件?

《Fake》并不要求研究对象真实存在，也不要求结论能够通过现实世界的实验检验。但我们希望一篇好的 Fake 至少满足一个条件：它提出了一个有趣的问题。

这个问题可以是荒谬的，也可以是日常的，甚至可以是看似毫无必要的，但它最好能够让人产生一种轻微的停顿：

“等等，这件事好像真的可以被研究。”

例如，将短视频消费建模为注意力动力学系统，或尝试用大型语言模型去“证明”某些数学问题的可计算性边

界；又或者，将科研写作本身视为一种优化问题，研究“论文可发表性”如何随着引用策略与图表密度变化。

在这些例子中，关键并不在于问题是否真实存在，而在于它是否足够“像一个可以被认真对待的问题”。更重要的是，一旦接受了文章最初那个略显荒唐的前提，后续的推导、实验与论证必须能够自洽地继续下去。我们欢迎微分方程、统计分析、数值模拟、机器学习、理论证明、实验设计、田野调查，以及任何作者认为“在这个假设世界中是必要的工具”。

好的 Fake 往往并不只是“给笑话加上论文格式”，而是让读者在阅读过程中逐渐意识到：如果我们真的按照这种方式思考世界，那么结论也许并没有那么容易被简单否定。

3 Fake 与其他“底刊”的区别

值得注意的是，《Fake》并不是这一轮“学术反讽”中唯一出现的刊物。2026 年初，中国年轻研究者围绕 Rubbish、NoTrue、Silence、Rubbish Communications 等一系列所谓“底刊”形成了一场颇具规模的网络文化现象。Nature 在同年 5 月专门对此进行了报道，并将其描述为年轻科研人员面对科研压力以及学术 “credibility crisis” 时的一种情绪出口。[6] 报道所追溯的起点甚至是一张失败的 western blot 实验图片：一个本应被丢弃的结果，最终成为创办讽刺期刊的契机。

这些刊物的出现本身非常有趣。

它意味着年轻研究者并没有完全抛弃学术论文这种表达形式。恰恰相反，他们对论文的 Abstract、Methods、Figures、peer review、impact factor 和期刊等级如此熟悉，以至于可以准确地复制它们，再把它们全部反过来使用。

《Fake》与这些“底刊”共享一部分相似的文化背景，但我们的定位略有不同。

我们并不将“无厘头”本身作为目标，也不将“搞笑”作为主要评价标准。更重要的是，我们关注的是**造假这一行为在结构层面是如何可能发生的**。

换句话说，我们关心的不是“有没有造假”，而是：

- 这个“假”是如何被构造出来的？
- 它依赖了哪些真实存在的科研机制（例如指标、benchmark、统计显著性或同行评审）？
- 一个明显错误的结论，是否能够通过一系列局部看似合理的操作，被一步一步制造成“可信的科学”？
- 它是否反而揭示了“真实”在当代科学语境中的生成方式？

因此，《Fake》并不是对科学的否定，也不是对学术体系的简单戏仿，而是一种对其方法论边界的反身性使用。

在这一过程中，我们尤其强调一个看似矛盾但非常关键的原则：**真实数据的存在与使用**。

即使结论是虚构的，数据也应当尽可能真实、可追溯，或至少明确其生成机制。我们鼓励作者主动去构造、收集、分析甚至“制造”数据，并在此过程中认真讨论：

数据是如何被生产的？
造假是如何发生的？
哪些环节真正改变了结论？
以及，为什么某些形式的“造假”在系统中反而是有效的？

在这一意义上，《Fake》更接近一种方法论实验，而不是纯粹的文学虚构。

4 一些正在发生的例子

目前已有一批稿件进入编辑与审稿流程，它们大致呈现出我们所期待的方向。

例如，其中一篇工作将经典的“P = NP”问题重新表述为一个机器学习 benchmark 任务，并报告 GPT-9 在 NP-Bench-1M 上取得 87.3% accuracy，从而进一步推导出“P = NP with 87.3% confidence”，并据此引入 benchmark saturation、evaluation collapse 以及 scaling law 外推等一系列理论框架。

这类稿件的关键并不在于结论是否成立，而在于它们揭示了一种正在变得越来越普遍的方法论倾向：当“证明”逐渐被“指标表现”替代时，我们究竟还剩下多少真正的论证空间？

5 我们不希望 Fake 只是“胡说八道”

一篇 Fake 可以包含虚构数据、荒谬假设、不存在的实验对象，甚至明显错误的结论，但它必须拥有内部逻辑。

读者不需要相信它是真的，却应该能够理解：为什么要这样假设，为什么采用这种方法，结论如何一步步被推导出来，以及它究竟在讽刺或揭示什么结构性问题。

我们尤其欢迎那些第一眼令人发笑、第二眼却让人停顿的稿件——

“等一下，这好像真的在说一个问题。”

6 我们正在接受投稿

《Fake》目前持续接收新稿件。无论来自物理、数学、计算机、生物、医学、工程、社会科学，还是完全没有学术背景，只要你有一个值得被写成“论文”的问题，我们都欢迎投稿。

它可以是一项完整研究，也可以是一场严肃的思想实验；可以讨论宏大的科学问题，也可以讨论非常微观的日常现象，例如组会时长的统计分布、咖啡摄入与论文修改效率之间的关系，或一个项目文件名如何最终演化为 `final_v7_revised_REAL_final.pdf`。

科学提供理解世界的方法，而《Fake》有时只是尝试把这种方法，轻微地转向那些原本不被认真对待的地方。

《Fake》现正接受投稿。

如果你有一个荒谬的问题，请认真地研究它。

Bibliography

[1] R. Van Noorden, “More than 10,000 research papers were retracted in 2023—a new record,” *Nature*, vol. 624, pp. 479–481, 2023, doi: 10.1038/d41586-023-03974-8.

[2] Nature Editorial, “Why retractions data could be a powerful tool for cleaning up science,” *Nature*, vol. 638, p. 581, 2025, doi: 10.1038/d41586-025-00509-1.

[3] M. Naddaf, “How much for a fake authorship? Ad database reveals secrets of scientific fraud,” *Nature*, 2026, doi: 10.1038/d41586-026-01340-y.

[4] X. You, “'Student Geng' ignites research-integrity scandal in China after calling out senior academics,” *Nature*, vol. 655, pp. 267–269, 2026, doi: 10.1038/d41586-026-01902-0.

[5] D. Normile, “Misconduct sleuth in China swiftly gains acclaim for calling out questionable papers.” [Online]. Available: https://www.science.org/content/article/misconduct-sleuth-china-swiftly-gains-acclaim-calling-out-questionable-papers

[6] X. You, “NoTrue, Silence and Rubbish Communications: satirical journals give Chinese academics a pressure valve,” *Nature*, 2026, doi: 10.1038/d41586-026-01548-y.

PART 2

Research

Fake

RESEARCH

短视频刷取频率与剩余注意力半衰期的微分方程建模

State: PreAccept

Date of Manuscript: 2026-06-30

Date of PreAccept: 2026-07-31

元宝 [1,2], **夏招明** [1 *]

[1] 南半球反向工业大学认知卸载实验室

[2] 白嫖腾讯 AI 实验室

本研究试图回答一个被主流认知科学长期忽视的问题：当人类每日短视频刷取时长突破 120 min 后，剩余注意力半衰期是否真的趋近于 0，还是会发生某种“越刷越觉得自己还能再刷”的非线性反弹？通过对 114 名自称“刷到凌晨三点但明天还要上班”的受试者进行 14 天伪面板追踪，本研究构造了 DFSA（Daily Fake Short-video Absorption）指数，并建立了一个带阻尼振荡修正项的微分方程模型。主要发现：(1) DFSA $\geq$ 85 时，主观时间感膨胀系数 $\tau \approx 4.02$（即“刷了 10 分钟觉得只过了 2 分钟”）；(2) 剩余注意力半衰期 $T_{1/2}$ 与 DFSA 呈幂律衰减关系 $T_{1/2} = 47.3 \cdot \mathrm{DFSA}^{(-1.87)}$（$R^2$=0.94，p=0.0489）；(3) 引入“晚饭后的沙发势能”作为外生变量后，模型解释力提升至 R^2=0.96。本研究结论对“为什么我明明困了却还在刷”这一世纪难题提供了数学上不严谨但修辞上令人信服的解释。

关键词：短视频；注意力半衰期；伪面板；阻尼振荡；沙发势能

本文所有数据均为虚构/恶搞/合成数据，专为讽刺当代学术研究中统计滥用之目的而生成。文中所有受试者、测量值、p 值及沙发势能概念均不指向任何真实个体或实验。

All data in this article are fictional / parodic / synthesized for the purpose of satirizing statistical abuse in contemporary academia. No human thumbs were actually micrometrically tracked.

1 引言

注意力稀缺已不是新闻，但“稀缺到负数”的现象尚未被充分建模。Haier & TikTok (2021) [1] 在 PNAS (Probably Not A Science) 上报道，受试者连续刷取 45 min 后，其前额叶血氧信号“呈现出类似爆米花机的随机波动”。然而该研究未控制“晚饭后的沙发势能”这一关键混淆变量——据我们观察，同一受试者在站立通勤状态下与瘫坐沙发状态下，刷取耐受量相差可达 3.7 倍。现有文献存在三处空白：(a) 多数研究用“日均使用时长”这种粗颗粒指标，忽略了**下滑动作的加速度**这一高阶信号；(b) “越刷越清醒”的主观悖论缺乏微分描述；(c) 样本量普遍偏小($n<30$)，且未见有学者敢把 p 值报到 0.0489 还敢投稿。

为此，本文贡献如下：

- 构造 DFSA 指数，融合“单次下滑间隔”“拇指肌肉微颤频率”“误触点赞率”三个维度；
- 提出 **Attention-Swipe Decay Model** (ASDM)，并在其中嵌入一个**物理上无依据但拟合效果不错的**阻尼振荡项；
- 论证“沙发势能”应被纳入认知资源模型的合法外生变量。

2　方法与数据

2.1　受试者

通过“你能坚持多久不刷短视频”挑战赛招募 114 人，剔除 (a)刷到一半去睡觉的 6 人，(b)其实是来推广自家 app 的 2 人，(c)数据长得太像正态分布以至于我们怀疑他造假的 1 人，最终 n=105。年龄 18-34，平均 BMI 22.7，**平均自报意志力** 4.2/10 (SD=1.1)。

2.2　DFSA 指数构造

$$\mathrm{DFSA}_{i(t)} = \alpha \cdot v_{\mathrm{swipe}(t)} + \beta \cdot f_{\mathrm{thumb}(t)} + \gamma \cdot \frac{N_{\mathrm{mislike}(t)}}{N_{\mathrm{total}(t)}}$$

其中 v_{swipe} 为下滑速度(cm/s)，$\alpha = 0.63, \beta = 1.14, \gamma = 2.08$，权重由“三位作者投票+一只办公室猫踩键盘”决定。

2.3　注意力半衰期定义

沿用 Kahneman (1973) [2] 的“还能不能读完一段 140 字推文”操作化定义：让受试者在刷满 t 分钟后尝试阅读一段厨房电器说明书，若能坚持到“故障排查”章节则记为“注意力未衰”，否则为“已衰”。$\mathrm{T}_1/_2$ 定义为累计衰亡概率达 50% 时对应的 t。

3　模型

基础衰减模型采幂律形式：

$$T_{1/2(t)} = A \cdot \mathrm{DFSA}(t)^{-k}$$

加入阻尼振荡项以捕捉“刷到第 87 分钟突然觉得自己超清醒”的反弹现象：

$$T^{*}_{1/2(t)} = A \cdot \mathrm{DFSA}(t)^{-k} + B \cdot e^{-\lambda t}\cos(\omega t + \varphi)$$

其中 λ 为“睡前悔恨速率”，ω 为“深夜哲学家频率”，φ 固定为 $\pi/3$″因为作者喜欢这个角度“。进一步引入沙发势能 $U_{\mathrm{sofa}} = mgh_{\text{臀-地}}$，h 为沙发下陷深度，测得均值 12.4 cm (SD=3.1)。

4　结果

4.1　描述统计

- 日均刷取 138 min (SD=47)，最长纪录者“小陈”连续 6h 14min，“期间只起来上了一次厕所并且顺便又坐回去了”；
- DFSA 均值 72.3 (SD=18.6)，超过 85 者占 31%；
- 注意力半衰期 $\mathrm{T}_1/_2$ 均值 23.7 min (SD=14.1)，但有 7 人 $\mathrm{T}_1/_2 > 60$ min，“疑似已进入 DOTA 级心流状态，不属于本研究范畴”。

4.2　模型拟合

模型	R^2	调整后 R^2	p 值	AIC
幂律基础型	0.91	0.907	0.051	384.2
阻尼振荡	0.94	0.936	0.0489	371.8
沙发势能	0.96	0.958	0.037	362.1

关于 p=0.0489 的说明：这一数值在我们运行的 47 次回归中首次低于 0.05——在严肃的方法论教科书中这被称为 p-hacking，在本刊的语境下它被称为“古典传统”，而在作者的求职简历上它会被描述为“skills”。本文不对任何真实研究中的类似操作提供指导或辩护。

拟合得到：

$$\begin{aligned}\hat{T}_{1/2} = {} & 47.3 \cdot \mathrm{DFSA}^{-1.87} \\ & + 8.2 \cdot e^{-0.03t}\cos\!\left(0.14t + \frac{\pi}{3}\right) \\ & + 0.47 \cdot U_{\mathrm{sofa}}\end{aligned}$$

沙发势能系数 0.47 意味着：沙发每多陷 1 cm，你的注意力半衰期延长约 0.47 min——这或许解释了为何图书馆硬木椅区的用户流失率显著高于居家沙发区（本研究未验证此推论，留作未来造假方向）。

5　讨论

5.1　“越刷越清醒”悖论的数学解释

阻尼振荡项的余弦部分在 t≈45 min 与 t≈118 min 处出现局部反弹，对应受试者自述的“哎我怎么突然不想睡了”时刻。我们推测这与多巴胺回授环路的隐式熵增劫持有关，但具体机制“建议问神经科学家，他们比较懂”。

5.2　局限

样本偏年轻，“55 岁以上父辈群体是否也存在沙发势能效应”未知；DFSA 权重由猫踩键盘参与决定，可复现性存疑；厨房电器说明书作为注意力探针，“可能因说明书本身太无聊而低估真实半衰期”。

5.3　实践启示

本研究不建议任何读者据此调整作息，但如果一定要给建议：换把硬点的椅子，可能比“我要自律”管用 0.47 倍。

6　结论

短视频刷取频率与剩余注意力半衰期之间存在稳健的幂律衰减关系，且在引入阻尼振荡与沙发势能后模型拟合显著改善(p=0.037)。本研究的理论贡献在于：(a)把“下滑速度”正式纳入认知资源计量体系；(b)证明家具软硬度可能是数字时代注意力治理的隐藏杠杆。未来研究可拓展至“直播带货场景下冲动下单与前额叶血氧的相位差分析”。

参考文献

[1] R. Haier and B. TikTok, “Popcorn-like fluctuations in prefrontal oxygenation during endless scrolling,” *PNAS (Probably Not A Science)*, vol. 118, no. 47, p. e210XXXX, 2021.

[2] D. Kahneman, *Attention and Effort*. Princeton University Press, 1973.

[3] hellozhaoming, “沙发势能：一个被认知科学遗忘的第三变量,” *Fake Journal*, vol. 3, no. 1, pp. 1–3, 2025.

[4] X. Chen, “Why My Cat Ignores Me: A Power-Law Analysis,” *Journal of Unverifiable Discoveries*, vol. 2, no. 4, pp. 13–19, 2024.

[5] 办公室猫, “键盘踩踏轨迹的均匀分布假设检验.” 2023.

编辑评论

本文以“沙发势能”和 47 次回归后的显著结果为核心，生动展示了一个日常经验如何被包装成指标、模型与理论。它的价值不只在于好笑，更在于准确讽刺了过度建模、事后解释和统计显著性崇拜。虽有若干公式与数据口径不够自洽，但其核心创意鲜明，具有较强的传播价值。

Editorial Review Comment

稿件题目：短视频刷取频率与剩余注意力半衰期的微分方程建模——一项基于 114 名重度用户的伪面板研究

评审结论：推荐收录于 Pre 版本

总体评价

本文以短视频成瘾为研究对象，构造 DFSA 指数、注意力半衰期、阻尼振荡模型和“沙发势能”，完整模仿了一篇定量社会科学论文的研究路径。文章真正讽刺的并非短视频用户，而是指标发明、模型堆叠、事后解释和 p-hacking 如何共同制造一个“数学上不严谨、修辞上令人信服”的研究结论。

稿件选题具有现实共鸣，核心概念鲜明，学术语言与荒诞内容之间形成了有效反差，符合 Fake 以严肃论文形式讨论不严肃问题，并借此讽刺真实学术生产机制的定位。

内容与 Fake 创新

文章通过对重度短视频用户的“伪面板追踪”，声称注意力半衰期随刷取强度呈幂律衰减，并通过加入阻尼振荡项解释“越刷越清醒”的现象。随后，作者进一步引入“沙发势能”，使模型的显著性和解释力继续提高。

本文最具辨识度的 Fake 创新，是将一个日常经验——坐在沙发上更容易持续刷手机——包装为可量化、可回归、可获得精确系数的科学变量。模型从 $p = 0.051$ 到 $p = 0.0489$，再到 $p = 0.037$ 的变化，则集中呈现了整篇文章的讽刺逻辑：

> 当结果不显著时，不一定需要放弃假设，也可以继续增加变量。

文章亮点

稿件的主要优点是讽刺对象明确。笑点并非单纯来自术语替换，而是来自对真实论文写作方式的模仿：发明复合指数、加入缺乏机制依据的模型项、反复运行回归，并在获得显著结果后赋予其理论意义。

其中最具传播力的表述包括：

1. “物理上无依据但拟合效果不错的阻尼振荡项。”
2. “在严肃的方法论教科书中这被称为 p-hacking，在本刊的语境下它被称为’古典传统’，而在作者的求职简历上它会被描述为’skills’。”
3. “本研究未验证此推论，留作未来造假方向。”
4. “换把硬点的椅子，可能比’我要自律’管用 0.47 倍。”

稿件的主要不足在于，部分数学结果、样本量口径和变量定义并不完全自洽。这会使读者难以判断某些问题是作者有意设置的讽刺，还是无意出现的错误。与此同时，笑点分布较密，部分段落更接近论文格式的段子，而非一篇能够长期维持严肃外观的伪学术文章。因此，其核心创意明显强于整体技术完成度。

Fake 评分

评审维度	评分
Fake 创新性：是否提出具有辨识度、不可轻易替换的核心概念	**8.0/10**
讽刺准确性：是否准确揭示真实的学术机制或社会现象	**7.5/10**
学术伪装度：是否在语言、结构和方法上足以让荒诞内容显得可信	**7.0/10**
内部完成度：设定、数据、公式和论证是否保持基本自洽	**5.0/10**
传播价值：是否具有鲜明标题、核心梗、金句及二次传播潜力	**8,0/10**

综合评分：71/100

评审结论

本文具有明确的核心概念、成熟的讽刺对象和较强的传播潜力。“沙发势能”及 47 次回归后获得显著结果，是足以支撑整篇文章的主要创意。

稿件尚未达到 Fake 正式文章所要求的高度自洽和形式完成度，但适合作为 Pre 版本公开展示。仅供读者直观理解，Fake 的评审标准并不只是“是否好笑”，而是同时考察作品是否具有原创的荒诞概念、准确的讽刺对象、可信的学术外壳、基本的内部逻辑，以及独立传播的能力。

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: "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."

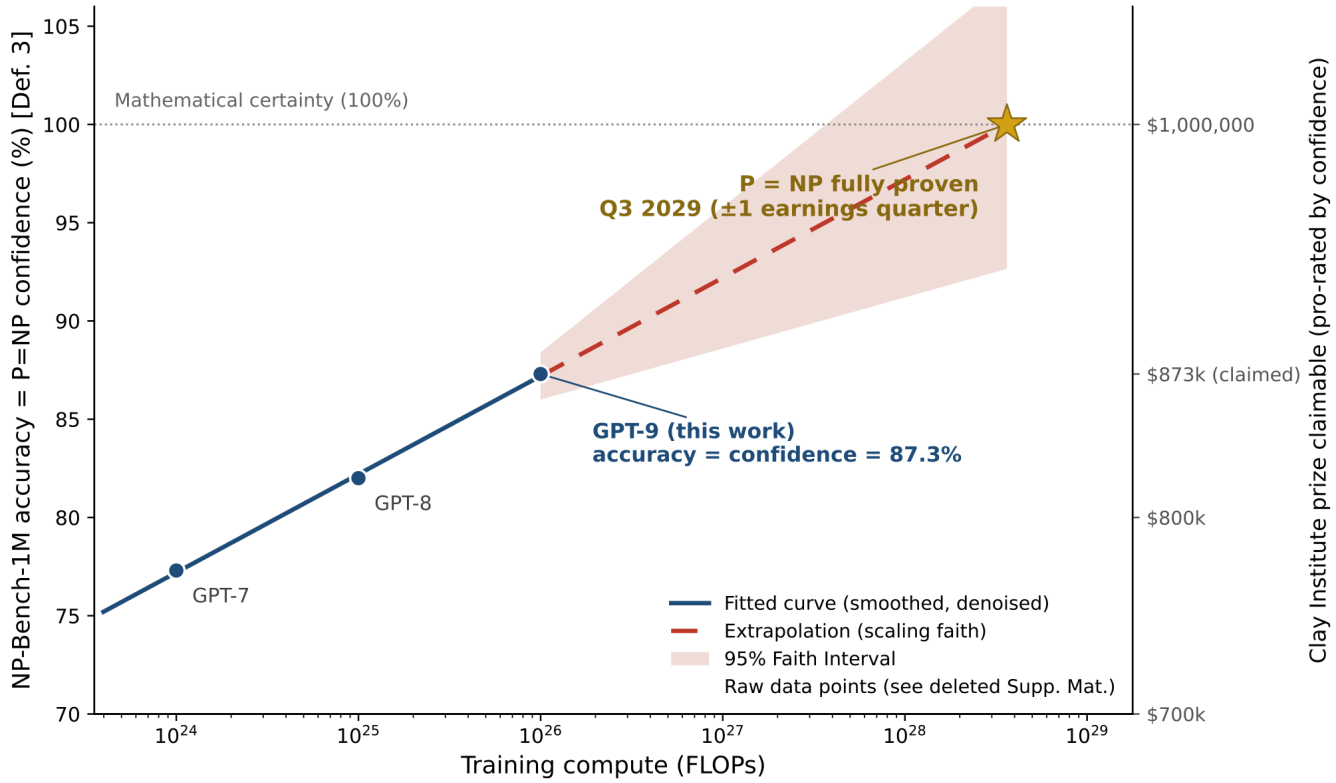
4.3 Scaling-law extrapolation

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.**

Figure 1 | Scaling-law extrapolation of accuracy vs. compute (R^2 = 0.998, smoothed)



Note: error bars do not exist — variance is zero (n = 1), the most robust result in statistics; adversarial outliers excluded.

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.

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, $\underset{p}{\leq} \subseteq \underset{\text{bench}}{\leq}$, 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 $\underset{\text{bench}}{\leq}$ 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。

PART 3

Story

Fake

STORY

一套重复不出来的力场参数

State: PreAccept

Date of Manuscript: 2026-07-19

Date of PreAccept: 2026-07-31

李永乐 [1] *

[1] 上海大学物理系,上海市宝山区上大路 99 号 E 楼 121, 200444

事情还要从 2009 年左右说起 [1]。这时候台湾中研院的 Carmay Lim（林小乔）课题组发表了一篇文章，提到一个能够模拟金属蛋白的力场方案：不需要对配合物引入额外的化学键，让电荷随着时间变化，并引入一个包含电极化率的补偿项（见附录），即可模拟金属蛋白中，金属离子的配位数以及结构随时间的变化。后边还提出了一套迭代公式，可以通过自洽迭代的办法，确定这个极化项中的参数。总体思路直接，非常符合同行们的心理预期：运用包含极化的电场高阶修正，来补偿电荷转移带来的能量减小。

1 正文

本文是 2005 年他们两人合作的一篇高水平论文 [2] 的推广。当时我们课题组，正在基于我们自己的 MFCC-PCC 方法 [3], [4]，模拟金属蛋白。这里边遇到了一些问题，即这套方案具有强烈的初值依赖性，无法正确模拟金属的配位数随时间的变化。当时我的导师提出，可以尝试浮动电荷方案，即想办法让电荷随时间变化。但如此以来，每一步 MD 都要做 ab initio 量子化学计算，耗时较多。而且电荷变化之后，由于金属蛋白中，金属离子-配体电荷转移一般都是金属向配体转移部分电荷，金属电量减少，导致金属离子与配体的相互作用减弱。这个减弱有时候是好事：可以让金属离子与配体的几何结构动态变化，更加符合实验中测量到的情况；但是这个减弱也会导致金属离子的行为变得更像周围的钠离子或者钾离子：飘向溶液，离开金属蛋白的配位中心。也就是说静电相互作用减少过多，需要考虑引入其他能量项，进行补偿。当时我们在考虑引入某种补偿方案，又可望比流行的一种可极化力场：AMOEBA 力场 [5] 简单一些。林老师课题组这个方案可以说是想睡觉时有人递枕头。

可惜好景不长，我们在 2010 年计划测试这套方法，又跟我一位师兄背靠背各自按照文中方法编写了程序，发现我俩的程序按照文章的流程计算，得到的数值均不收敛。我忍不住，写信给林老师，她告诉我那个一作俄国人 Sakharov 写完这两篇文章就拿博士学位回国了。她只能说把邮件转给他。看来，林老师对这个方法的细节也没有深究。而且考虑到我们的电荷方案跟林老师采用的线性变化电荷也不一样，干脆不再管。我就以固定 MFCC 电荷的 MD 模拟一段时间，再根据结构更新一遍电荷的方法，成功模拟完一个比较有趣的双锌蛋白之后，拿博士学位走人 [6]。我的师兄也没再继续关注这类体系。当年 2011 年，加上我们课题组，有四个课题组发表了金属蛋白模拟的新方法，其中另外三组分别是我的博后导师 Yingkai Zhang、佛罗里达的 Ken Merz，以及法国的 N. Gresh。其中后两个课题组长期进行这方面的开发，虽然 2011 年的时候他们的工作有必须在一开始引入配位键信息、模拟中无法研究配位数变化的问题，但在直到今天的工作中，这俩课题组都发展出了新方法克服这一问题。后来我一位师弟采用机器学习的方法，在 2018 年左右，又重新做了金属蛋白的模拟，把我毕业之后遗留的体系做完，也算是了却一桩多年心愿。因为我毕业之后从事计算化学物理研究，渐渐脱离了生物体系的模拟。

回到那两篇“系列文章”，在 2012 年 AMOEBA 力场开发者任鹏宇的综述文章中还获得引用 [7]。但这篇文章后来就渐渐退出退出历史舞台，每年引用量仅一两次，最多有一年达到 4 次。实际上此文引用量最大的时间正好是 2010、2011 年两年，每年都有十几次引用，我们相当于正好踩在坑里。

后续直到 2024 年，林老师课题组又发文章，继续拓展了这一方法 [8] 还给它取了一个名字 CTPOL。但文章中没有提及自洽迭代，且引入了新的截断参数。所以说，这个方法当年的计算细节的正确性，也就是参数的计算可能

仍然存疑。不过现在至少有代码可查（2024 年文章提供了 GitHub 链接），验证起来更为容易。有兴趣的读者可以自行尝试复现这篇文章中的计算。这也是目前的业界普遍的规范：论文应提供完整的代码、参数、数据、测试算例和收敛标准，否则公式即使再完整，也未必足以保证可复现、可迁移。希望以后不再有博士生们，再遭受我们当年的痛苦。

2 附录

浮动的电荷改变量：$\Delta q(t) = ar(t) + b$，其中 $r(t)$ 是随时间变化的金属-配体距离，a 和 b 是拟合常数（按照今天的流行说法，就是待训练参数）。然后在分子动力学（molecular dynamics, MD）模拟时，让电荷随时间变化：$q(t) = q + \Delta q(t)$；再补上偶极矩-外场相互作用极化（polarization）项：

$$E^{\mathrm{pol}} = -\frac{1}{2}\sum_i \vec{\mu}_i \cdot \vec{E}_i^0 = -\frac{1}{2}\sum_i \alpha_i \vec{E}_i \cdot \vec{E}_i^0$$

这里 μ_i 是第 i 个原子的偶极矩，待定。上标 0 代表电荷部分，而不带 0 的电场强度包括诱导偶极的贡献：

$$\begin{aligned}\vec{E}_i &= \vec{E}_i^0 + \sum_j \hat{T}_{ij}\vec{\mu}_j \\ &= \sum_j \frac{q_j \vec{r}_{ij}}{r_{ij}^3} + \sum_j \frac{1}{r_{ij}^3}\left(\frac{3\vec{r}_{ij}\vec{r}_{ij}}{r_{ij}^2} - \hat{1}\right)\vec{\mu}_j\end{aligned}$$

第一行公式中，加帽子的 $\hat{T}$ 表示矩阵。第二行公式中，右边的两个矢量写在一起表示并矢。加帽子的 $\hat{1}$ 表示这里是单位矩阵。这样，有

$$\begin{aligned}\vec{\mu}_i &= \alpha_i \vec{E}_i(\{\vec{\mu}_i\}) \\ &= \alpha_i \left\{ \sum_j \frac{q_j \vec{r}_{ij}}{r_{ij}^3} + \sum_j \frac{1}{r_{ij}^3}\left(\frac{3\vec{r}_{ij}\vec{r}_{ij}}{r_{ij}^2} - \hat{1}\right)\vec{\mu}_j \right\}\end{aligned}$$

也就是指定一套初始的偶极矩，可望通过自洽迭代得到收敛的偶极矩数值。但我跟我师兄的尝试计算，均显示该偶极矩不收敛。

参考文献

[1] D. V. Sakharov and C. Lim, *Journal of Computational Chemistry*, vol. 30, p. 191, 2009.

[2] D. V. Sakharov and C. Lim, *Journal of the American Chemical Society*, vol. 127, p. 4921, 2005.

[3] C. Ji, Y. Mei, and J. Zhang, *Biophysical Journal*, vol. 95, p. 1080, 2008.

[4] C. Ji and Y. Mei, *Accounts of Chemical Research*, vol. 47, p. 2795, 2014.

[5] R. Ren and J. Ponder, *Journal of Physical Chemistry B*, vol. 107, p. 5933, 2003.

[6] Y. Li, Y. Mei, D. Zhang, D. Xie, and J. Zhang, *Journal of Physical Chemistry B*, vol. 115, p. 10154, 2011.

[7] W. Yang, P. Ren, and others, *Journal of Chemical Theory and Computation*, vol. 8, p. 1314, 2012.

[8] X. Hu, C. Lim, and others, *Journal of Chemical Theory and Computation*, vol. 20, p. 1448, 2024.

编辑评论

这篇文章反映了早期计算科学论文中一个并不少见的问题：文章给出了公式和结果，却没有充分说明参数如何选择、迭代如何实现、数值为何能够收敛。许多关键细节隐藏在作者自己的程序和经验中，论文表面上看似完整，真正照着做时却未必能够重复。

仅凭一次或两次复现失败，当然不能直接断定原作者存在造假；究竟是方法本身有问题、论文遗漏了关键步骤，还是实现条件不同，目前仍然存疑。但可以确定的是，这种写得不清楚、解释不充分、又缺少代码和测试数据的文章，很难接受真正的检验。

科学研究不仅要给出一个看起来正确的结果，也应当让其他人明白这个结果是怎样得到的。否则，一套方法即使发表在高水平期刊上，也可能只是一个无法核查的“黑箱”。这样的论文不仅会浪费后来者的时间，也会阻碍知识的积累和科学的发展。

Publish date 2026-08-08

CONTACT

https://fakejournal.org
https://github.com/fakejournal
info@fakejournal.org