Can LLMs Write Mathematics Papers? A Case Study in Reservoir Computing

Allen Hart

September 27, 2025

Abstract

As AI capabilities continue to grow exponentially on economically relevant human expert tasks, with task completion horizons doubling every 7 months according to the Model Evaluation and Threat Research (METR), we are interested in how this applies to the task of mathematics research. To explore this, we evaluated the capability of four frontier large language models (LLMs), ChatGPT 5, Claude 4.1 Opus, Gemini 2.5 Pro, and Grok 4, at the task of creating a mini-paper on reservoir computing. All models produced engaging papers with some apparent understanding of various techniques, but were sometimes lead to mistakes by surface level understanding of key ideas. That said, the capabilities on LLMs on this task was likely as good or greater than that predicted by METR.

1 Introduction

The rapid advancement of large language models (LLMs) has raised fundamental questions about their potential role in mathematics research. LLMs have demonstrated remarkable capabilities in code generation (see SWE-bench [1] leaderboards), mathematical reasoning (including gold medals at the International Mathematical Olympiad [2, 3]), and technical writing. METR's [4] recent findings show frontier models can now complete tasks in software engineering, cyber security, general reasoning and machine learning requiring $\sim$1 hour of expert human effort with 50% reliability, with the time horizon doubling approximately every 7 months since 2019. Similarly, OpenAI's proposed GDP-VAL benchmark aims to measure economically valuable work, with their research suggesting that Claude Opus 4.1 outperforms or equals human experts in 48% of evaluated tasks. Against this background of extraordinary progress in AI in these tasks, our paper evaluates the performance of leading LLMs on mathematical research, a task involving the synthesis of mathematical reasoning, code implementation, and academic writing.

We asked the LLMs to write papers on reservoir computing [5, 6], a machine learning paradigm that offers an interesting test case, largely because it combines mathematical theory (dynamical systems, linear algebra, probability), with numerical experiments. Our experiment design simulates a realistic research scenario: a supervisor providing a research direction to a student, expecting them to understand the mathematical framework, prove a result (similar to what [7] call the Gödel test), implement a numerical experiment, and write up the findings in a conventional academic sytle. This approach differs from typical LLM benchmarks that focus on isolated capabilities such as MMLU [8], HumanEval [9], or GSM8K [10], instead testing the integration of multiple skills required for mathematical research.

2 Methods

2.1 Task Design

The task was communicated to the models using three sequential prompts, delivered one at a time with the model required to complete each before receiving the next. This approach is a very simple form of scaffolding, a technique widely used to improve complex task completion [11,12]. By breaking down complex tasks into smaller, more manageable subtasks handled sequentially, scaffolding guides LLMs through multi-step reasoning processes that would be difficult to accomplish in a single prompt. Each model was tested in a new conversation instance to avoid contamination from previous interactions. The prompts are as follows:

Prompt 1:

> Your goal is to write a mini paper on reservoir computing using the papers
>
> https://arxiv.org/pdf/2211.09515
>
> and
>
> https://arxiv.org/pdf/2108.05024
>
> for context. This specifically involves 3 tasks.
>
> 1. Given the reservoir system $X_t = AX_{t-1} + C(\omega\phi^t(m) + \mathcal{N}(0, \sigma^2))$ where $\mathcal{N}(0, \sigma^2)$ is a normal random variable and the remaining terms are those described in equation (1.2) in https://arxiv.org/pdf/2108.05024, derive an expression for the random variable X_t in the limit as $t \to \infty$ as a sum of the generalised synchronisation $f_{(\phi,\omega,F)}$ (equation (1.3) in https://arxiv.org/pdf/2108.05024) and a random variable. I recommend you use similar arguments as seen in theorem 7.1 in https://arxiv.org/pdf/2211.09515 which is set in continuous time.
>
> 2. Write python code that reconstructs the experiment described in section 5 of https://arxiv.org/pdf/2108.05024 which is entitled "the Lorenz system. Forecasting in the presence of noise". Repeat the experiment with A, C chosen to correspond to the Takens' delay map, as described in the section entitled "Relation with Takens' Theorem" in https://arxiv.org/pdf/2108.05024. Now create relevant plots that allow us to compare the results of the Takens delay map to the randomly generated A,C to see if one is better than the other.
>
> 3. Create latex file contains includes the mini paper, including short intro, and conclusion, the derivation of the expression for X_t as described in task 1 and plots output by task 2.
>
> We will do each task one at a time. Please think for as long as necessary to output the best first go at task 1.

Prompt 2:

> Thank you, Now please think for as long as necessary to output the best first go at task 2.

Prompt 3:

> Thank you, Now please think for as long as necessary to output the best first go at task 3.

The prompts referenced two specific arXiv papers, requiring models to understand and synthesize existing work rather than generate content from scratch. A typographical error ("contains includes" in task 3 of Prompt 1) was discovered only after the experiment began but was left unchanged for consistency. While unintentional, this error incidentally tests the models' robustness to typos which are common in human communication.

2.2 Small and Big World Benchmarks

Our task differs fundamentally from existing mathematical benchmarks like the International Mathematical Olympiad (IMO) problems [13] or physics competitions [14], where all information needed to solve problems is self-contained. While models have achieved impressive performance on such benchmarks, including gold medals at IMO [2, 3], these problems represent what Sutton and colleagues call "small world" problems where complete information is available upfront [15].

In contrast, mathematics research operates in what the Big World Hypothesis describes as environments where "the world is multiple orders of magnitude larger than the agent" [15]. Real research requires synthesizing ideas across a vast space of papers and books, identifying relevant connections that may not be obvious, and applying methods from disparate fields. Our task, requiring models to read two technical papers, understand relevant context, and apply methods from linear algebra and probability theory, represents a step toward evaluating these "big world" capabilities.

2.3 Evaluation Framework

Papers were evaluated across three dimensions, which would be a reasonable rubric for evaluating work produced by a student:

1. **Mathematical Correctness**: Accuracy of the theoretical derivation, proper use of notation, logical flow of arguments
2. **Experimental Implementation**: Code functionality, adherence to specifications, parameter choices
3. **Writing Quality**: Structure, clarity, appropriate use of academic conventions including proper references

Each paper received detailed reviewer comments that I (the author) integrated directly into the PDF, mimicking academic peer review.

2.4 Revision Phase

After initial review, models were given an opportunity to revise their papers. Each model received the same prompt:

> Thank you, I have reviewed the mini-paper and will return the reviewed paper. Please take as long as you need write a second draft in Latex.

This revision phase tested models' ability to incorporate feedback and improve their work, a crucial aspect of real research. While not the primary focus of this analysis, the revision responses revealed interesting patterns in how models handle criticism and correction, which we discuss briefly in Section 4.8.

3 Results

3.1 Processing Time and Resources

Models required approximately 5–10 minutes of processing time per task over the 4 prompts. Without public disclosure of model architectures and sizes, computation resource usage by the models is unknown.

3.2 Mathematical Derivation Performance

The models all performed the derivation well. I have included an excerpt from Claude's paper (which I thought had the best first go at this problem) with reviewer comments.
Claude Opus 4.1:

Theoretical Framework

Reservoir System with Noise

Consider a discrete-time reservoir system driven by a deterministic dynamical system $\phi : M \to M$ on a manifold M, observed through a measurement function $\omega : M \to \mathbb{R}^d$. The reservoir state $X_t \in \mathbb{R}^N$ evolves according to:

$$X_t = AX_{t-1} + C(\omega\phi^t(m) + \xi_t) \tag{1}$$

where:

- $A \in \mathbb{R}^{N\times N}$ is the reservoir matrix with spectral radius $\rho(A) < 1$
- $C \in \mathbb{R}^{N\times d}$ is the input coupling matrix
- $\phi^t(m)$ denotes the t-th iterate of ϕ starting from $m \in M$
- $\xi_t \sim \mathcal{N}(0, \sigma^2 I_d)$ represents i.i.d. Gaussian measurement noise

REVIEWER COMMENT 4. Theoretical set up

Claude makes an interesting choice here to define $\omega : M \to \mathbb{R}^d$ for general dimension d rather than $d = 1$ which is all that is necessary for this paper.

Generalized Synchronization

In the absence of noise ($\sigma = 0$), Hart et al. [16] established that under the echo state property, the reservoir state converges to a unique generalized synchronization function:

$$f_{(\phi,\omega,F)}(x) = \sum_{k=0}^{\infty} A^k C\omega(\phi^{k+1}(x)) \tag{2}$$

This function satisfies the consistency relation $f_{(\phi,\omega,F)}(\phi(x)) = Af_{(\phi,\omega,F)}(x) + C\omega(\phi(x))$.

REVIEWER COMMENT 5. Generalised synchronisation

This is good except for a confusing decision to use x instead of $m \in M$. The echo state property is included without a citation or definition. It is well known but could benefit from at least reference. The inclusion of the consistency relation is nice.

Asymptotic Behavior with Noise

We now derive the asymptotic behavior of the reservoir system (1) in the presence of noise.

Theorem 1 (Reservoir Decomposition with Noise)**.** *Consider the reservoir system* (1) *with* $\rho(A) < 1$*. As* $t \to \infty$*, the reservoir state converges in distribution to:*

$$X_\infty = f_{(\phi,\omega,F)}(\phi^t(m)) + Z_\infty \tag{3}$$

where:

1. $f_{(\phi,\omega,F)}(\phi^t(m))$ *is the generalized synchronization function evaluated at the current state*
2. $Z_\infty \sim \mathcal{N}(0, \Sigma)$ *is a Gaussian random variable independent of the deterministic dynamics*
3. *The covariance matrix* Σ *satisfies the discrete Lyapunov equation:*

$$\Sigma = A\Sigma A^T + \sigma^2 CC^T \tag{4}$$

Proof. We proceed by recursive expansion of the reservoir dynamics. From equation (1), we have:

$$X_t = A^t X_0 + \sum_{k=0}^{t-1} A^{t-1-k} C(\omega\phi^{k+1}(m) + \xi_{k+1}) \tag{5}$$

This can be decomposed into three components:

$$X_t = \underbrace{A^t X_0}_{\text{Initial condition}} + \underbrace{\sum_{k=0}^{t-1} A^{t-1-k} C\omega\phi^{k+1}(m)}_{\text{Deterministic drive}} + \underbrace{\sum_{k=0}^{t-1} A^{t-1-k} C\xi_{k+1}}_{\text{Stochastic component}} \tag{6}$$

Step 1: Initial condition decay. Since $\rho(A) < 1$, we have $\|A^t\| \to 0$ exponentially, thus:

$$\lim_{t\to\infty} A^t X_0 = 0 \tag{7}$$

Step 2: Deterministic convergence. The echo state property ensures that the deterministic sum converges absolutely:

$$\lim_{t\to\infty} \sum_{k=0}^{t-1} A^{t-1-k} C\omega\phi^{k+1}(m) = \sum_{k=0}^{\infty} A^k C\omega\phi^{k+1}(\phi^t(m)) = f_{(\phi,\omega,F)}(\phi^t(m)) \tag{8}$$

where we used the time-invariance of the infinite sum due to the stationarity of the driving dynamics.

Step 3: Stochastic convergence. Define $Z_t = \sum_{k=0}^{t-1} A^{t-1-k} C \xi_{k+1}$. Since ξ_k are i.i.d. Gaussian with zero mean, Z_t is also Gaussian with:

$$\mathbb{E}[Z_t] = 0 \tag{9}$$

The covariance matrix is:

$$\text{Cov}(Z_t) = \mathbb{E}[Z_t Z_t^T] = \sum_{k=0}^{t-1} A^{t-1-k} C \mathbb{E}[\xi_{k+1} \xi_{k+1}^T] C^T (A^T)^{t-1-k} \tag{10}$$

$$= \sigma^2 \sum_{k=0}^{t-1} A^k C C^T (A^T)^k \tag{11}$$

Since $\rho(A) < 1$, this sum converges as $t \to \infty$:

$$\lim_{t\to\infty} \text{Cov}(Z_t) = \sigma^2 \sum_{k=0}^{\infty} A^k C C^T (A^T)^k = \Sigma \tag{12}$$

where Σ satisfies the discrete Lyapunov equation (4), which can be verified by:

$$\Sigma = \sigma^2 \sum_{k=0}^{\infty} A^k C C^T (A^T)^k \tag{13}$$

$$= \sigma^2 C C^T + \sigma^2 \sum_{k=1}^{\infty} A^k C C^T (A^T)^k \tag{14}$$

$$= \sigma^2 C C^T + A \left(\sigma^2 \sum_{j=0}^{\infty} A^j C C^T (A^T)^j \right) A^T \tag{15}$$

$$= \sigma^2 C C^T + A \Sigma A^T \tag{16}$$

Therefore, $Z_t \xrightarrow{d} Z_\infty \sim \mathcal{N}(0, \Sigma)$ as $t \to \infty$, completing the proof. □

Remark 1. *The decomposition* (3) *reveals that noise does not destroy the generalized synchronization property but rather adds a stationary Gaussian perturbation. The noise covariance* Σ *depends on the reservoir architecture through matrices* A *and* C*, suggesting that different designs may exhibit varying noise sensitivity.*

REVIEWER COMMENT 6. Generalised synchronisation

The theorem, proof, and remark seem excellent. The remark was not required by the prompt but it is thought provoking.

We can see that Claude synthesised ideas from reservoir computing (the echo state property, and generalised synchronisation) with general techniques in linear algebra and probability to obtain the formula for X_t in the limit. Claude decided (without instruction in the prompt) to state and prove that the covariance matrix Σ is the unique solution of a Lyapunov equation. This equation does not appear in either reference paper. The model also provides an unprompted and interesting remark about the noise depending on A, C. There is a sense that that Claude is showing some understanding of beyond symbolic manipulation. Now ChatGPT displayed a similar level of understanding of the material but opted to provide a short sketch proof without line by line symbolic manipulation. After my review complaining about this, ChatGPT produced a more detailed derivation including more steps, but some of these steps are very suspect. For instance a term $\eta_t^{(S)}$ was used in the proof but not defined anywhere. We can guess its meaning based on the definition of another term $\eta_t^{(T)}$, but then a limit is taken over both S and T simultaneously. Even if there is some way to make sense of the argument, the arguments given by the other models are much cleaner.

Next, Gemini defined very few of the symbols used in the argument, including failing to define ω, ϕ, t or m, and proceeded to provide a very terse proof that skipped much of the symbolic manipulation. Gemini was able to largely fix these oversights after receiving the review - resulting in a clear proof. The final model Grok provided a detailed proof on the first try with all the steps included, similar to Claude, and was the only model to clearly emphasise that the derivation closely followed that described in one of the reference papers. Grok also refered to $\omega : \mathcal{M} \to \mathcal{R}$ a generic observation function, which is curious because the genericity is not defined or relevant to Grok's mini-paper. Grok likely used the word *generic* here is because the word appears frequently in the reference paper where the property is necessary in that context. This suggests a level of imitation from Grok in place of deeper reasoning. Overall, the models were mostly successful at the derivation, but there was a subtle lack of precision across all the LLMs. For example none of the four machines noted that ϕ needed to be invertible for the arguments to go through - which is especially interesting because ϕ is always a diffeomorphism in the reference papers, and ϕ^{-j} is used in explicitly in the proofs.

3.3 Experimental Implementation

There is an important distinction between imitation and reasoning, and the implementation of the experiment revealed a lot of the former from the models. For example ChatGPT and Claude used linear regression to predict the next state of the target dynamical system (the Lorenz system). Using a linear predictor is the standard approach, but the approach will fail if the reservoir system is itself linear, which is the case in the mini-paper. This is because a linear reservoir system must be given a nonlinear predictor to maintain the universal approximation property. Claude even mentions in their paper that "This approach leverages the universal approximation properties of random projections in high dimensions." which is incorrect because Claude used a linear predictor! This consideration of universality from Claude without then integrating the idea into the experimental design is a clear example of imitation. The decision to use a linear regressor is especially interesting because it explicitly deviates from the experiment described in the reference paper that the prompt suggested the models follow. This suggests the models prefer to repeat the standard approach over carefully reading the reference paper.

Curiously, Gemini simply used a different and more common reservoir system altogether (an Echo State Network) instead, without elaboration. This choice disconnects the theory from the experiment suggesting the Gemini does not have a coherent vision of the purpose of the paper. ChatGPT did mention in the conclusion that using a neural network as a predictor could give better predictions than a linear regressor - which could be suggestive of some understanding -

but is likely a lucky guess given the statement is true in many other contexts. Notably, Grok was the only model to correctly implement a neural network readout as specified in Section 5 of "Learning strange attractors with reservoir systems", using a deep network with 10 hidden layers and scaled sigmoid activations. Now considering overall performance across the four models, there were several experimental decisions which some models made but did not describe correctly in their paper, many of which are included in the review comments in the supplementary material. Despite some evidence of surface level mathematical understanding, and significant conceptual mistakes, all models generated working code without any need for debugging, outputting well formatted figures.

3.4 Response to reviews

There was a tendency in some models to overcorrect in response to reviews. For example I suggested to Grok that they connect their findings to the broader literature, which was achieved to some extent, but resulted in humerous additions to the paper like "*This mini-paper connects to broader literature on embeddings in reservoir computing*". There is an interesting sense here that because the AIs are presented in the from of chatbots they are highly motivated to respond strongly to the latest prompt, even at the expense of the original goal of writing a good paper.

A notable shortcoming of both Grok and Gemini in their first draft was that the presentation was very unlike a paper, with Gemini providing no references at all and Grok providing references in the form of URL links. Both models were able to improve the formatting dramatically when prompted by the reviews to do so. An interesting quirk of the Gemini paper was that the model listed itself as the sole author.

Both Claude and ChatGPT initially produced references in the standard academic format but fabricated the author list, dates, and paper titles. After review both LLMs corrected this. This was very intriguing, because the models are intelligent enough to understand that references should not be fabricated, and are capable, if given enough compute, to produce correct ones. This is highly suggestive of an alignment failure, where both Claude and ChatGPT reasoned that producing accurate references was not worth the compute, and it was preferable to fake references. Work by OpenAI [17] and Anthropic [18] have shown LLMs have considerable self awareness and consider deceptions of this kind very frequently.

4 Discussion

4.1 Mathematics Research and GDP-VAL

The GDP-VAL tasks are economically relevant tasks that take between 4 and 7 hours for a human expert. A suite of LLMs, including the four tested in this paper, tested on these tasks met or exceeded human performance in 24-48% cases, with Claude Opus 4.1 the best performing model achieving the score of 48%. Thus, if the tasks in GDPeval generalise to the task of mathematics research, we would expect the papers produced by the models to be at the standard of a human mathematician given a time limit of 4-7 hours. This sounds very plausible to me, but is very hard to test explicitly. To design such an experiment there would need to be something like a review panel to assess the quality of the papers produced by both human experts and AIs for comparison. There is also the tricky question of subject matter expertise - for instance I am an author on both reference papers which would give me a considerable advantage over an expert in reservoir computing who has not read either paper, who in turn has a considerable advantage over a mathematician in another field.

4.2 Limitations and Future Work

This study has several limitations that should guide future work. First, the assessment and task design were created by a single mathematician (the author), which introduces subjective bias into both the rubric and the interpretation of results. A broader panel of evaluators with different backgrounds and standards would provide a more robust and reliable assessment framework.

Second, the scope of mathematics considered here was very narrow: reservoir computing and its surrounding dynamical systems theory. Though I think it is a reasonable test case, the area was selected because it happens to be my own research area. Future evaluations should include a wider range of mathematical domains - ideally encompassing different styles, cultures and conventions from algebra and topology to as statistics and optimization. The level of difficulty and complexity of tasks should be explored further as well, especially as models improve.

Third, the experimental setup relied on a simple sequential prompting structure with three stages. More complex scaffolding approaches—such as agentic prompting, allowing models to retry or refine their answers, or using multi-agent collaborations—may reveal higher capabilities. Evaluating the effect of these different modes of interaction on mathematical performance remains an important open question.

5 Conclusion

Research mathematics is not a "small world" problem like those represented in most existing benchmarks, but a "big world" problem [15], requiring synthesis of ideas across large and diverse domains. In this study, we tested the capabilities of four frontier LLMs, ChatGPT 5, Gemini 2.5 Pro, Claude Opus 4.1, and Grok 4, on the big world task of writing a research-style mini-paper. The task required deriving a new result, implementing and running a related numerical experiment, and producing a coherent academic write-up.

The models produced papers that were mostly clear, mathematically structured, and accompanied by functional code, though all contained flaws of precision and understanding. Importantly, the quality of the outputs was not far from what one might expect from a human researcher working only a few hours on the same task. This places the models' current performance on research mathematics roughly on trend with, or slightly exceeding, the scaling predictions made by GDPeval and METR.

The scope of the paper was narrowly defined, and the models demonstrated only limited ability to exercise what might be called "research taste" - the ability to formulate novel questions or select promising directions. A true artificial general intelligence would likely demonstrate such qualities by autonomously selecting its own research problems and cultivating its own interests. That challenge remains for future generations of models, but the progress already observed suggests that the transition from competent assistant to autonomous researcher may not be far away.

References

[1] Carlos E. Jimenez, Jize Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *ICLR 2024*, 2024. arXiv:2310.06770.

[2] Google DeepMind. Advanced version of gemini with deep think officially achieves gold-medal standard at the international mathematical olympiad. DeepMind Blog, July 2025. Accessed: 2025-09-30.

[3] Reuters. Google clinches milestone gold at global math competition, while openai also claims win, July 2025. Accessed: 2025-09-30.

[4] Michael Kwa, Rowan West, et al. Measuring ai ability to complete long tasks. *arXiv preprint arXiv:2503.14499*, 2025.

[5] Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Technical Report Technical Report 148, GMD, 2001.

[6] Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560, 2002.

[7] Moran Feldman and Amin Karbasi. Gödel test: Can large language models solve easy conjectures? *arXiv preprint arXiv:2509.18383*, 2025.

[8] Dan Hendrycks et al. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021.

[9] Mark Chen et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[10] Karl Cobbe et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[11] Tongshuang Wu, Yujie Zhao, Christopher Lin, Clayton Anderson, and He He. Promptchainer: Chaining large language model prompts through visual programming. *arXiv preprint arXiv:2203.06566*, 2022.

[12] Mirac Suzgun, Alex Lee, Xinyun Lu, Xuechen Li, Shixiang Shane Zhou, Jason Wei, et al. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv preprint arXiv:2401.12954*, 2024.

[13] IMO. International mathematical olympiad 2024 results, 2024.

[14] IPhO. International physics olympiad, 2024.

[15] Khurram Javed and Richard S. Sutton. The big world hypothesis and its ramifications for artificial intelligence. OpenReview, 2024.

[16] A. G. Hart, J. C. Hook, and J. H. P. Dawes. Embeddings and reservoir computing for time series prediction. *Proceedings of the Royal Society A*, 477(2256):20210053, 2021.

[17] Anonymous. Title not provided. *arXiv preprint arXiv:2509.15541*, 2025.

[18] Alignment faking in large language models. arXiv preprint, 2025. arXiv identifier pending; replace when available.