LLMs as philosophers

What can they do, and why aren't they better?

William D'Alessandro

1. An LLM project for philosophy of mathematical practice

First let me briefly describe my project. As a philosopher of mathematical practice, I'm interested in the methods, goals and values of working mathematics. Much of my work has focused on the nature of *explanation* within mathematics and how the desire for explanatory understanding shapes mathematical research. (For instance, mathematicians often reprove the same theorem in different ways because they judge that earlier proofs merely *verified* the result without giving any insight into why it's true.)

The goal of my project was to use AI tools to create a large dataset of references to explanatory concerns for use by mathematical practice researchers. I downloaded 5000 math PDFs from arXiv.org and ran a Python script to feed these to Gemini in batches, along with a prompt detailing the kinds of examples I wanted. I then applied a filtering prompt to remove many low-quality selections from Gemini's initial outputs. The final result was a dataset consisting of ~1250 annotated examples from ~735 distinct sources (available here). Gemini's ability to accurately retrieve information from very long prompts without hallucination was essential to this phase of the work.

With the dataset in hand, I was interested to test Gemini's utility for various followup research and analysis tasks. At one point I asked the model to adjudicate between existing theories of mathematical explanation on the basis of the examples contained in the dataset. Its response was cogently reasoned and accurate in its characterization of the views at issue. Here for instance is the portion of Gemini's analysis dealing with ontic theories of explanation (sans-serif Arial font indicates text produced by Gemini; the references in Gemini's response are to cases from the dataset):

> **Challenges for Ontic Approaches:**
>
> Ontic accounts generally hold that explanations track objective dependency relations in the world (or, for mathematics, objective relations between mathematical facts/structures). The explanans makes the explanandum happen or be the way it is due to these real relations.

https://doi.org/10.14361/9783839447529-115 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Challenging Cases/Types of Cases:

1. **Explanations via "heuristic interpretation" or "intuitive analogy" that are explicitly not the rigorous proof:**

    - **Example:** Anshelevich providing a "heuristic interpretation" of the functional Itô formula, linking it to classical Taylor expansions, explicitly for understanding before the rigorous (and potentially complex) proof.
    - **Challenge:** While the heuristic might point to an underlying structural analogy, the explanation itself in these instances seems to reside in the cognitive act of relating the complex to the simple and familiar, even if that analogy isn't a perfect or rigorous ontic dependency. The value is in making it "graspable", which is more of an epistemic/pragmatic virtue than a direct reflection of an objective dependency chain in the formal proof. An ontic account would need to argue that the heuristic does track some objective (perhaps looser) structural resemblance.

2. **Reproofs explicitly valued for "simplicity", "clarity", or being "less computational" when the original proof already established the ontic link:**

    - **Example:** Schlosser's "elementary derivations" of hypergeometric identities, contrasting with prior "not entirely elementary" proofs. Or the desire for a "more conceptual" proof by Burghelea & Haller.
    - **Challenge:** If the original proof already correctly identified the mathematical facts from which the theorem follows (the ontic basis), why is a new, simpler proof more "explanatory"? The ontic dependencies might be the same. The added value seems to be epistemic: the new proof makes those dependencies easier to grasp or presents them in a more "natural" way. An ontic account might struggle to capture this difference in explanatory value if the objective dependencies are unchanged.

3. **Explanations where the "explanatory work" seems to be done by a representation or framework rather than directly by the objects themselves:**

    - **Example:** The "logarithmic point of view" (El Goul) making a refined result an "easy consequence". The framework itself facilitates understanding. Or Fialowski & Penkava discussing how reframing deformation theory in terms of coderivations makes connections "completely transparent".
    - **Challenge:** While the framework undoubtedly reflects ontic structures, the explanatory power seems to come from the choice of representation that highlights certain features and relationships. An ontic account needs to carefully delineate whether the explanation lies in the objective structures themselves or in the way a particular framework allows us to cognize those structures.

https://doi.org/10.14361/9783839447529-115 https://www.inlibra.com/de/page/agb - Open Access - CC BY

I was also interested in Gemini's ability to philosophize in an original way about the cases in the dataset. So I asked it to try. ("Suppose you were asked to come up with an entirely novel theory of explanation, one able to do better than existing theories at making sense of all the cases you've reviewed. What might such a theory look like? Feel free to speculate.") A representative portion of the model's response:

> If I were to speculate on an entirely novel theory of mathematical explanation, aiming to encompass the diverse cases reviewed, I would try to build something that integrates insights from existing theories while addressing their limitations. Let's call it **"Explanatory Resonance Theory" (ERT)**.
> Here's what ERT might look like:
>
> **Core Idea of Explanatory Resonance Theory (ERT):** Mathematical explanation occurs when a target mathematical statement or phenomenon (the explanandum) is brought into "explanatory resonance" with a broader, more structured, or more fundamental mathematical context (the explanans-framework). This resonance is not a single type of connection but a multi-faceted relationship that enhances our cognitive grasp and structural integration of the explanandum.
>
> **Key Components and Features of ERT:**
>
> 1. **Multi-faceted Resonance:** "Resonance" isn't a monolithic concept. It can manifest in several ways, acknowledging the pluralistic nature of what mathematicians value as explanatory: [here Gemini lists and describes "structural", "unifying", "simplifying/clarifying", "analogical/metaphorical" and "generative" varieties of resonance, attempting to incorporate the core ideas of the views discussed previously and bringing up many of its earlier examples].
> 2. **The Explanans-Framework:** This is not just a single proposition but a richer context. It could be:
>
>   ○ A well-established theory (e.g., Galois theory, representation theory, category theory).
>   ○ A specific set of definitions and their consequences.
>   ○ A powerful visualization or diagrammatic system.
>   ○ A newly introduced concept or structure that re-organizes existing knowledge.
>
> 3. **Cognitive Grasp and Structural Integration (The Goal):** The ultimate aim of explanation under ERT is to enhance the mathematician's cognitive grasp of the explanandum. This means:
>
>   ○ **Understanding Why:** Not just that it's true, but why it must be true given the explanans-framework.
>   ○ **Seeing Connections:** Understanding how the explanandum relates to other mathematical ideas and structures.
>   ○ **Increased Predictive/Manipulative Power (Potentially):** A good explanation might suggest generalizations, new questions, or more efficient ways to work with the concepts.

https://doi.org/10.14361/9783839447529-115 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

- **Reduced Surprise/Mystery:** Phenomena that initially seem “surprising” or “mysterious” become less so when they resonate with a suitable framework.

4. **Dynamic and Contextual:** What counts as a good explanans-framework or a strong resonance can be:

   - **Field-dependent:** Different mathematical subfields might value different types of resonance. Bijective proofs are highly valued in combinatorics for their structural resonance.
   - **Historically contingent:** What was once a complex explanandum can become part of a familiar explanans-framework for new explananda.
   - **Audience-dependent (to some extent):** While striving for objectivity, the effectiveness of an explanation in producing “cognitive grasp” can have a pragmatic element related to the audience’s background knowledge.

In essence, Explanatory Resonance Theory would propose that mathematical explanation is the process of making a mathematical idea “vibrate sympathetically” with a larger, more harmonious, and well-understood part of the mathematical orchestra. The quality of the explanation depends on the richness, depth, and clarity of this induced resonance.

While Gemini provided a strong response to the earlier critique prompt, it’s clear that this is not a similarly impressive piece of original philosophy. It exhibits the typical LLM tendency to try to accommodate as many standard ideas and existing views as possible without taking a definitive stand on any claim of note. Its unifying idea is a vague, unhelpful metaphor. Not a single feature of the proposed theory is meaningfully novel. (I realize that readers outside the philosophy of mathematical practice may lack the background to independently evaluate Gemini’s proposal, but I don’t think my assessment is controversial.)

Gemini’s behavior in my experiment is an example of a wider phenomenon. The best current LLM-based systems excel at analytical tasks requiring considerable insight, subject understanding and dialectical sophistication. Their performance here often seems roughly at the level of human experts. At the same time, these systems show little ability to go beyond the frontier: somehow their vast knowledge and impressive reasoning ability almost never yields breakthroughs in any domain. Even Sam Altman, ever ready to boast about OpenAI models’ capabilities, has only expressed the hope that “systems that can figure out novel insights” might arrive sometime in 2026 (Altman 2025).

Ben Levinstein’s Philosophy Pipeline project[1] furnishes another discipline-specific example. Levinstein has assembled a complex AI workflow to generate complete philosophy papers in the style of the journal *Analysis*, aiming eventually for a 20% publication success rate. While the current version of the Pipeline produces readable papers with competent structures and appropriate citations, Levinstein agrees that “no interesting

1 https://github.com/balevinstein/philosophy-pipeline.

https://doi.org/10.14361/9783839447529-115 https://www.inlibra.com/de/page/agb - Open Access - CC BY

discoveries or even ideas have come from LLMs yet" (July 2025, personal communication).

I confess to finding this somewhat mysterious. Why do current models make incisive assistants but poor thinkers, when it seems as though the cognitive capabilities required in these cases should be mostly the same? What do we even mean when we talk about a model's ability to generate "high-quality original ideas"? What sorts of capabilities might allow future models to clear this important hurdle?

In the rest of this note, I'll discuss a few plausible explanations which I don't ultimately find satisfying and end with one that seems better.

2. Why can't LLMs produce good new ideas?

I assume that AI systems are the kinds of things that could at least in principle do worthwhile original philosophy and make other sorts of intellectual breakthroughs. Of course, not everyone shares this assumption: some think human minds have productive powers which no machine learning model could ever match, owing for instance to human consciousness, embodiment, ensoulment, the nonformal or nondigital nature of human cognitive processing, or some other putatively unique feature of ours. Those who hold such views may take themselves to have a simple explanation of LLMs' failure to make creative discoveries. Since I don't hold them, my job is a bit harder.

2.1 Because they can't invent new concepts?

In philosophy and elsewhere, intellectual progress is often closely linked to the introduction of new concepts: almost singlehandedly, notions like *rigid designator*, *possible world*, *scientific paradigm* and *epistemic injustice* seem to bring order to many things we know and open doors to many others we haven't yet considered.

One might also think that LLMs are fundamentally unable to devise new concepts. This suspicion is supported by the common picture of language models as "stochastic parrots" (Bender et al. 2021), able only to repeat the contents of their training data in various permutations. If correct, this would go a long way toward explaining why LLMs can be made very good at manipulating and applying existing ideas without ever producing transformational insights.

But I don't think this explanation is right, at least on some understandings of what it claims. If the idea is supposed to be that LLMs are strictly limited to the store of concepts encountered during training, this seems pretty clearly untrue. LLMs are skilled at devising categories which were almost certainly not seen in training and which require some degree of creativity to hit upon.

A model benchmark[2] based on the *New York Times*' daily Connections puzzle gives one way to see this. In (the standard version of) the puzzle, players are presented with a grid of 16 words which must be sorted into four sets of four, with the words in each set falling under a common concept. Typically the category definition for at least one set (often the

2 https://github.com/lechmazur/nyt-connections

https://doi.org/10.14361/9783839447529-115 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

hardest category, styled with a purple background) involves unusual wordplay or a non-obvious factual connection. For instance, the words in this set in a recent puzzle were CASH, HAZE, MAC, PEC. The NYT-provided category description was *starts of culinary nuts* (cashew, hazelnut, macadamia, pecan).

When the Connections benchmark debuted in spring 2024, the then-best models did relatively poorly. But the benchmark has been saturated by the current generation of reasoning models. For instance, OpenAI's o1 correctly solved about 99% of the 90 puzzles from December 2024 to February 2025. (The human player base's success rate during this period was about 71%.) This led to the development of a new assessment based on a more difficult version of the puzzle than the one played by humans.

It seems clear that achieving a high level of performance on Connections requires the ability to devise concepts which are insightful and appropriate yet in some sense novel. (o1 presumably didn't encounter the specific concept *starts of the names of culinary nuts* in its training data. At any rate, it didn't learn 89 out of 90 similarly recondite concepts in this way.) The stochastic-parrot explanation in its strictest sense therefore seems implausible.

Of course, one could point out that the concept *starts of the names of culinary nuts* is a composite made up of concepts which the model surely did see in training. Entirely true: but then the same can be said for human concepts associated with major advances. The ideas of names, reference, possible worlds and the like were all well known before Kripke. Fricker invented neither the category of the epistemic nor the notion of injustice. Concepts praised as novel and insightful are often assembled from existing materials.

What distinguishes important new concepts from relatively trivial ones seems not to be mere novelty, but rather the virtues the former possess: elegance, breadth, utility, naturalness, explanatory power, and so on. We can agree that LLMs aren't very good at inventing concepts with these characteristics. But why not?

2.2 Because they aren't good at reasoning?

A second story one could tell starts with the importance of reasoning to intellectual breakthroughs. It's rarely sufficient, after all, to come up with a nice idea or two: one has to show how these ideas fit together among themselves and with the known facts, what can be done with them, why they're better than alternatives. Indeed, given sufficient acumen, one might not need new ideas at all. Mathematicians often celebrate[3] clever elementary proofs of known results. Sometimes progress comes from understanding a problem deeply enough to grasp what's essential and see how to do more with less.

And one might tie this story to a claim about LLMs' reasoning abilities. Again, there's a ready-to-hand stereotype that language models never engage in thought, inference, planning or world-modeling; they are, rather, "glorified autocompletes" whose outputs are constrained not by truth or logic but only the estimated likelihood of a given text continuation.

3 https://math.stackexchange.com/questions/1849840/big-list-of-erd%C5%91s-elementary-proofs

https://doi.org/10.14361/9783839447529-115 https://www.inlibra.com/de/page/agb - Open Access - CC BY

There is of course *something* to this stereotype. Pretrained LLMs are indeed trained to minimize a loss function tied to their ability to successfully predict held-out words in internet text. And this is indeed part of the reason they sometimes hallucinate, repeat common falsehoods and make various strange mistakes. But it's also clear that this is a highly incomplete picture of how current LLM products work.

First, although the objective of pretraining is to minimize loss on text completion, the best way to fulfill this demanding objective is often by developing cognitive-like internal structures and functions. Sometimes these structures can be identified explicitly, as in the case of GPT-2 small's circuit for locating the indirect object of a sentence (Wang et al. 2022). Large pretrained models should thus be understood as sophisticated information processors, not mindless word generators. Second, pretrained language models receive further finetuning to improve accuracy and reduce hallucinations. The state of the art here has moved far beyond the initial release of ChatGPT, with models like Gemini 2.5 Pro outperforming human experts[4] on questions requiring information retrieval from long-context prompts. Third, the advent of reasoning models has led to vast improvements in LLM planning and problem-solving. A specially trained version of Gemini Deep Think recently logged a gold medal-worthy performance[5] on the 2025 International Mathematical Olympiad (IMO) problem set, answering five of the competition's six famously difficult problems correctly. (Less than 10% of the human test-takers, themselves elite competitive solvers of these kinds of problems, typically earn gold medals.) The IMO president described Gemini's solutions as "astonishing in many respects… clear, precise and [mostly] easy to follow".

All of this demonstrates that LLMs are at least in principle capable of sophisticated and accurate reasoning. Their performance in arenas like the IMO suggests that, given proper training, they can solve difficult problems of genuine intellectual interest. So it seems unlikely that a general lack of reasoning power is responsible for the dearth of LLM breakthroughs.

2.3 Because good philosophy has few reinforcement-learning-friendly patterns?

Here's my best guess about why LLMs are bad at original philosophy. While it shares some DNA with the previous suggestion, it isn't based on a dismissive picture of LLMs as non-thinkers, but rather on the details of how models are trained to reason well in particular domains.

The standard method for training reasoning models works roughly as follows. First, the model is given a problem to solve. Then, rather than attempting to immediately generate a solution in the fashion of a standard LLM, the model produces a "reasoning trace" or "chain of thought" describing a potential approach to solving the problem. These traces are assigned a score—by a human or (more often) a specialized reward model—according to the quality of the reasoning they contain and the correctness of the associated so-

4 https://longbench2.github.io

5 https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad

https://doi.org/10.14361/9783839447529-115 https://www.inlibra.com/de/page/agb - Open Access -

lution. The original model then updates its weights to favor high-scoring patterns and deprecate low-scoring ones, so it will be likelier to produce more of the former.

Models can learn some general-purpose reasoning guidelines in this way: e.g., that it's good to break complex problems into smaller pieces, proceed step by step, avoid circularity, and so on. But the whole process works especially well in a particular domain to the extent that several things are true: (1) reasoning in the domain involves many recognizable, repeatable patterns; (2) it's relatively clear which of these patterns are good or bad in a given context; (3) most problems have an unambiguously correct solution; (4) there exist large corpora of explicit examples of good and bad domain-specific reasoning (both for pretraining the main model to pick up a variety of standard reasoning styles, and for training the reward model to recognize good patterns).

Domains like mathematics and coding meet these criteria. There are countless publicly available examples of correct proofs. Everyone more or less agrees on what a good proof looks like. Much of mathematical reasoning involves choosing the right argument pattern from a library of standard techniques. (This is especially true for competition-style problems, which operate under a distinctive set of constraints and are known to draw on a learnable set of strategies. Google's engineers seem to have exploited these affordances, providing their customized Gemini model with "a curated corpus of high-quality solutions to mathematics problems" as well as "general hints and tips on how to approach IMO problems".) So it's no surprise that reasoning models have recorded some of their most impressive successes in domains like these.

Other domains are not so much like this. For instance, even the best current models perform quite poorly on the ARC-AGI-2 benchmark[6], involving a varied set of visual reasoning tasks meant to be straightforward for humans to complete. What makes the ARC challenge difficult for LLMs is at least in part that the problem set involves few repeated patterns: the logic of each problem, while not objectively very complex, is quite unique, so the task space has little in the way of readily learnable structure to exploit. The relatively small size of the 1000-problem training set is a further handicap.

Philosophical reasoning, I take it, is more like a set of ARC problems than a body of mathematical proofs. While there are such things as "standard philosophical moves"[7] that recur in various contexts (some of which are quasi-mathematical in character), it seems difficult to say anything very general about when a given argument pattern is likely to succeed or fail. (One man's modus ponens[8] is another man's modus tollens, as the saying goes.) And of course there's often deep disagreement about whether some argument has succeeded in a given case, or even whether it *could* succeed in any possible case. Finally, we lack the sorts of gold-standard research-level argument corpora that would fit comfortably into ML training pipelines.

Some of these problems could be overcome in principle with lots of work. Others seem less tractable. Unlike the cases of mathematical reasoning, coding and the like, neither AI developers nor academic philosophers have obvious financial incentives to go full

6 https://arcprize.org/leaderboard

7 https://dailynous.com/2025/01/16/a-taxonomy-of-philosophy-moves

8 https://en.wikipedia.org/wiki/Here_is_one_hand

https://doi.org/10.14361/9783839447529-115 https://www.inlibra.com/de/page/agb - Open Access - CC BY

steam ahead on this kind of progress, so it's unclear what would bring about the necessary changes.

In the meantime, it's not entirely surprising that LLMs make good assistants but mediocre original thinkers. Generic RL for reasoning has equipped them with strong domain-neutral thinking skills, but world-class performance requires additional domain-specific optimization. No model has yet been optimized in this way for philosophical reasoning. Alas for us, at least for now.[9]

References

Altman S (2025) The gentle singularity. In: *Sam Altman*. Available at: https://blog.samaltman.com/the-gentle-singularity (accessed 13 June 2025).

Bender E, Gebru T, McMillan-Major A et al. (2021) On the dangers of stochastic parrots: Can language models be too big? In: *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 3–10 March 2021, pp. 610–623. New York: Association for Computing Machinery.

D'Alessandro W (forthcoming) Using large language models to study mathematical practice. In: Kant D, Perez-Escobar JA, Sarikaya D and Sarikaya M (eds) *Mathematicians at Work: Empirically Informed Philosophy of Mathematics*. Dordrecht: Springer.

Wang K, Variengien A, Conmy A et al. (2022) Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. Available at: https://arxiv.org/abs/2211.00593.

9 Aside from the Gemini outputs directly quoted in the text, no LLMs were involved in the writing of this chapter.

https://doi.org/10.14361/9783839447529-115 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY