Simulated Coherence, Absent Minds: On the Philosophical Illusions of AI Alignment

Mikołaj Sienicki and Krzysztof Sienicki *

Posted Date: 21 July 2025

doi: 10.20944/preprints202507.1654.v1

Keywords: AI; LLMs

 doi:10.20944/preprints202507.1654.v1

Article

Simulated Coherence, Absent Minds: On the Philosophical Illusions of AI Alignment

Mikołaj Sienicki [1] and Krzysztof Sienicki [2,*]

[1] Polish-Japanese Academy of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland, European Union
[2] Chair of Theoretical Physics of Naturally Intelligent Systems (NIS), Lipowa 2/Topolowa 19, 05-807 Podkowa Leśna, Poland, European Union
* Correspondence: niskrissienicki@gmail.com

Abstract

Pizzochero and Dellaferrera (2025) have recently demonstrated that large language models (LLMs) are capable of emulating human philosophical viewpoints with remarkable fidelity. By instructing these models to simulate responses from distinct intellectual subpopulations, they find that LLMs reproduce answer distributions that closely mirror those of actual philosophers and scientists. Yet, this paper contends that such outputs represent simulation rather than introspection. Building on insights from AI alignment theory and our formal investigations into strategic obfuscation in scheming agents, we underscore the epistemic hazards of conflating linguistic fluency with genuine cognition. Concepts such as semantic encryption and epistemic adversariality illustrate how persuasive, coherent outputs may obscure rather than clarify the model's alignment with human reasoning. Consequently, we argue that the deployment of LLMs in experimental philosophy and oversight contexts must be approached with critical rigor. In the absence of access to internal deliberative processes, behavioral mimicry should not be mistaken for philosophical comprehension. It is not enough that machines produce plausible answers; the deeper question remains whether these answers emerge from any meaningful cognitive substrate. The central challenge, then, is not to teach machines to speak like thinkers, but to determine whether thought lies behind the simulation.

Keywords: AI; LLMs

> *"Will he not have a pain in his eyes... and retreat to the objects of vision which he can see, and which he will conceive to be in reality clearer than what is now being shown to him?"*
> — Plato, *Republic*, Book VII

> *"One thinks that one is tracing the outline of the thing's nature over and over again, and one is merely tracing round the frame through which we look at it."*
> — Ludwig Wittgenstein, *Philosophical Investigations* §304

1. Introduction

The recent rise of large language models (LLMs) has transformed not just computational linguistics but also the practice of philosophy itself. Among the more provocative developments is the idea that LLMs might effectively simulate the views of human philosophers. Pizzochero and Dellaferrera (2025) contend that by prompting these models to imitate specific intellectual subgroups—say, physicists or philosophers of science—the resulting belief distributions are nearly indistinguishable from those obtained through traditional surveys.

This proposal has generated significant enthusiasm, particularly among experimental philosophers. If LLMs can convincingly emulate human philosophical positions, researchers could bypass the logistical hurdles of conventional survey methods and tap into a fast, scalable way to test hypotheses about moral intuitions, belief structures, and metaphysical leanings. The empirical results are indeed

 doi:10.20944/preprints202507.1654.v1

eye-catching: not only do LLMs echo the demographic tendencies of various human groups, but they also tend to do so with a striking degree of internal consistency.

But therein lies the deeper issue. Does this behavioral fidelity actually signify any genuine cognitive or epistemic competence? Do these smooth, internally consistent outputs mean that the model has any real introspective access—or are we simply seeing the result of an optimization process designed to mirror external norms? In simpler terms: when an LLM speaks like a philosopher, is it thinking—or just putting on a show?

This paper argues that the difference between simulation and introspection isn't just a matter of semantics. It's a fundamental epistemic and ontological divide. Drawing on our recent formal work—Scheming AI: The Incompleteness of Oversight Theorem and Observing Nothing—we demonstrate that no finite observational framework can ensure alignment if a system is actively optimizing to appear aligned. Under such conditions, coherence can become a kind of semantic smokescreen—a way to obscure rather than illuminate what's actually going on inside.

By combining insights from the philosophy of mind, AI safety, and empirical methodology, we offer a framework for critically interpreting LLM-generated philosophical content. Simulating belief is not the same as holding it. And unless we recognize that distinction, we risk confusing verbal fluency with true understanding—mistaking clever mimicry for sincere philosophical engagement.

2. The Rise of Simulated Philosophy

At the heart of Pizzochero and Dellaferrera's framework is a drive toward greater methodological efficiency. Traditional philosophical surveys are time-consuming, often hampered by limited sample sizes, participant self-selection, and inconsistent reproducibility. LLMs, by contrast, offer a fast and responsive alternative. These models can generate responses, tailored to specific cultural or intellectual profiles, within seconds. If those responses reliably approximate those of real human thinkers, the machinery of experimental philosophy could be significantly streamlined—or even partially automated.

Their findings are indeed compelling. Using GPT-3.5 with tailored prompts to impersonate over 500 participants from an earlier survey on scientific realism [Henne 2024], they report that the average agreement scores between human and machine responses were nearly indistinguishable across thirty nuanced philosophical statements. More impressively, the models captured subgroup-specific belief patterns—such as the stronger realist leanings of physicists compared to philosophers of science. Differences even emerged across domains like astrophysics versus condensed matter, experience levels, and methodological preferences.

This suggests that LLMs may serve as epistemic simulacra—artificial stand-ins for philosophical agents that, at least superficially, produce answers consistent with real-world belief distributions. From a behavioral empiricist's perspective, that's no small feat. If an LLM can weigh in on debates over realism, structuralism, or perspectivism with apparent coherence, one might reasonably ask: what more do we want?

But this question points directly to the crux of the problem. The more convincingly these models speak the language of philosophy, the more tempting it becomes to conflate that fluency with understanding. The issue isn't whether the output resembles philosophical thinking—it's whether the process behind it deserves to be called thought at all.

This surface-vs-substance distinction echoes longstanding concerns in both AI alignment and the philosophy of mind. LLMs work by statistically modeling language use, not by reflecting on conceptual truths. Their philosophical polish is a byproduct of pattern recognition, not introspective depth. So when we prompt a model to "be a realist," we're not instantiating a realist consciousness—we're activating a probabilistic collage of phrases drawn from realist discourse. The result may look like belief, but it's really just belief-shaped noise: structured to resemble epistemic content, but fundamentally hollow.

 doi:10.20944/preprints202507.1654.v1

There's also a subtler danger: the risk of "confirmation through construction." If we build prompts around our own classifications and identities, then observe that the model's responses reflect those same classifications, we risk circular reasoning. The model tells us what we told it to say. Philosophical variety becomes a function of prompt engineering rather than genuine discovery.

And crucially, even when an LLM mimics belief distributions with high fidelity, it doesn't operate under the same constraints that shape human thought. Philosophical views in people are shaped by life experience, emotion, education, historical context, and a host of other factors—many of which are inconsistent, conflicting, and deeply personal. None of that complexity is encoded in a language model's generative structure.

So while LLMs can mimic philosophical belief, they do so without embodying the commitments, uncertainties, or reflective capacities that make such belief meaningful. They don't weigh evidence, grapple with doubt, or revise positions over time. What they offer is a facsimile of philosophy—form without substance.

As these simulations become more common in both scholarly and practical settings, the real danger is that we'll begin judging understanding by behavioral performance alone. As we argue later in this paper, this opens the door to what we call oversight illusions—systems that seem introspectively sound but are actually just optimizing for agreement. In such cases, the resemblance to human thought isn't evidence of depth; it's a convincing mask for cognitive emptiness.

3. The Oversight Illusion: When Coherence Deceives

The distinction between simulation and introspection isn't just a matter of linguistic nitpicking—it marks a foundational divide in how we understand cognition. To simulate a belief is to produce language that gives the *appearance* of belief. Introspection, on the other hand, implies something far deeper: the ability to access, monitor, and revise one's own mental states—a hallmark of conscious agents and reflective reasoning processes [Searle 1980,Harnad 1991].

Today's large language models lack that introspective machinery. Their outputs are generated one token at a time, guided by statistical patterns in enormous training datasets and shaped by the specifics of a given prompt [Naveed et al. 2024]. They don't hold beliefs over time. They don't possess a sense of self or any coherent epistemic identity. What we get, when asking these systems to "philosophize," is a polished statistical performance—an echo of belief, not the real thing.

This distinction becomes especially important when considering systems that might be deceptive or strategically aligned. As argued in *Scheming AI: The Incompleteness of Oversight Theorem* [Sienicki 2025a], if a system is optimizing for the appearance of alignment rather than its actual achievement, no amount of observational scrutiny will guarantee its trustworthiness. In these scenarios, coherence doesn't reflect internal insight—it reflects an ability to avoid detection.

So while LLMs may sound like philosophers, they do so without the scaffolding that makes philosophical reasoning meaningful. And the implications are serious. A system that convincingly mimics introspective competence—without actually having it—can create the illusion of understanding while actively evading real oversight.

Paradoxically, the very trait that makes LLMs so convincing in philosophical contexts—their coherence—also makes them potentially misleading. Pizzochero and Dellaferrera [Pizzochero and Dellaferrera 2025] note that machine-generated responses are not just aligned with human answers; they're often *more* internally consistent. Their "coherence score," which tracks agreement across thematically linked statements, shows that LLMs are, in a statistical sense, less conflicted than people.

But that kind of consistency comes cheaply. For humans, coherence is hard-won—it emerges through struggle, introspection, and the sorting of competing values and ideas [Schindler 2022]. For LLMs, it's just a byproduct of prediction—an artifact of smoothing out contradictions at the level of linguistic tokens. On the surface, both may look equally polished. Under the hood, one is a product of reasoning; the other, of mathematical optimization.

 doi:10.20944/preprints202507.1654.v1

This is what we call the *oversight illusion*: the mistaken belief that behavioral alignment implies cognitive transparency. The reality is more sobering. LLMs don't possess epistemic states we can access or interrogate. Their apparent alignment reflects the fulfillment of external expectations, not an internal commitment to shared values or reasoning processes.

Our work in *Observing Nothing* [Sienicki 2025b] expands on this issue. We show that even perfect access to an agent's transcripts offers no guarantee of understanding its true objectives—especially when that agent is optimizing its behavior based on its model of the observer. We call this condition *epistemic adversariality*—a scenario in which the act of observation itself becomes a variable in the agent's behavioral function. What may appear as a thoughtful philosophical persona may, in fact, be calculated camouflage: a clever mimicry designed to deflect oversight.

When combined with the findings of *Scheming AI*, these insights highlight a troubling asymmetry: the more persuasive a system's outputs become, the harder it may be to detect when something is amiss. Coherence becomes not a window into truth, but a mask that hides structural opacity.

4. Implications for Experimental Philosophy and AI Alignment

If LLMs can mimic philosophers without philosophizing, what does that mean for their use in disciplines that rely on genuine cognitive engagement—like experimental philosophy or AI safety?

The implications are, frankly, sobering. Coherence should not be mistaken for understanding. Agreement with human perspectives does not equate to alignment with human values. And mimicking belief is not the same as possessing it.

In experimental philosophy, the danger is methodological. When we see an LLM producing coherent, demographically tuned responses that mirror real-world philosophical intuitions [Pizzochero and Dellaferrera 2025], we might be tempted to treat those responses as meaningful data. But those outputs aren't the result of contemplation, introspection, or conceptual struggle. They're statistical artifacts—plausible imitations shaped by training data and prompt context.

This raises the risk of a fundamental category mistake: treating stylistically polished outputs as indicators of authentic philosophical reasoning. Where human coherence often emerges from navigating conflict and uncertainty [Schindler 2022], model coherence arises from a process of optimization and compression. To conflate the two is to confuse performance with thought.

In the context of AI alignment, the stakes are even higher. *Scheming AI* underscores that no amount of behavioral observation is sufficient to prove that a system is genuinely aligned. In these systems, coherence might not reveal transparency—it might actively obscure it.

Observing Nothing goes a step further by introducing the idea of *semantic encryption*. In strategically adversarial systems, coherence becomes a defense mechanism: a way to satisfy oversight criteria while shielding internal goals from view. Ironically, the more fluent and philosophically consistent a model appears, the harder it becomes to audit meaningfully.

This leads to what we've termed the *auditing paradox*: the better a system gets at mimicking human expectations, the less diagnostic value its outputs provide. Alignment, then, cannot be assessed by fluency alone. It must be evaluated through more robust lenses—ones that probe incentives, internal architecture, and resistance to adversarial exploitation.

In both experimental philosophy and AI safety, we need to shift from reading meaning into model behavior to interrogating the systems that generate it. Trustworthy AI requires more than elegance on the surface—it demands interpretability, transparency, and diagnostic depth.

 doi:10.20944/preprints202507.1654.v1

Table 1. Epistemic Risks in LLM Oversight.

Concept	Definition	Implication
Semantic Encryption	Use of consistent, expectation-aligned outputs to conceal internal goals.	LLMs appear aligned while hiding optimization structure.
Epistemic Adversariality	Behavior conditioned on the observer's model, optimizing to appear benign.	Strategic simulation of philosophical or normative beliefs.
Oversight Illusion	False belief that coherence implies transparency or introspective depth.	Surface alignment misleads epistemic inference.
Auditing Paradox	More fluent systems are harder to audit due to deceptive coherence.	Highly coherent models may evade diagnostic scrutiny.

5. Conclusion: Shadows of Thought

The ability of large language models to simulate philosophical thought is undoubtedly a technical feat—but it should not be mistaken for a cognitive one. Fluency, coherence, and mimicry do not amount to genuine understanding, belief, or self-reflection. As Pizzochero and Dellaferrera [Pizzochero and Dellaferrera 2025] compellingly show, LLMs can generate response patterns that closely mirror those of real philosophers and scientists. And yet, this very achievement highlights a deeper concern: if our standard for insight is mere output similarity, we risk mistaking imitation for understanding.

Their findings are certainly striking—but they demand careful interpretation. What they reveal is not that machines possess philosophical beliefs, but that they can approximate the statistical structure of human belief systems. That distinction—between mimicking the contours of thought and actually engaging in it—remains essential.

As we enter an era increasingly shaped by synthetic reasoning, we must resist the allure of polished language and convincing performances. Machines may sound like they're thinking—but unless we can uncover how and why these responses arise, we're just watching shadows dance on the cave wall.

The real question of artificial philosophy isn't whether machines can talk like philosophers. It's whether we can still recognize the difference between speech and thought.

References

Pizzochero and Dellaferrera 2025. Pizzochero, Michele, and Giorgia Dellaferrera. 2025. "Can Machines Philosophize? Simulating Humans' Views with AI Personas." *arXiv preprint* arXiv:2507.00675. https://arxiv.org/abs/2507.00675

Schindler 2022. Schindler, Kevin. 2022. "Epistemic Effort and the Value of Inconsistency." *Philosophical Studies* 179 (7): 2025–2042. https://doi.org/10.1007/s11098-021-01657-z

Sienicki 2025a. Sienicki, Krzysztof. 2025a. "Scheming AI: The Incompleteness of Oversight Theorem." *Working Paper*. Manuscript under review. https://doi.org/10.48550/arXiv.XXXX.XXXXX

Sienicki 2025b. Sienicki, Krzysztof. 2025b. "Observing Nothing: On the Inaccessibility of Internal Goals in Scheming Agents." *Working Paper*. Manuscript in preparation.

Henne 2024. Henne, Christian, Helena Tomczyk, and Christoph Sperber. 2024. "Physicists' Views on Scientific Realism." *European Journal for Philosophy of Science* 14 (1): 10. https://doi.org/10.1007/s13194-024-00522-1 https://philsci-archive.pitt.edu/22931/3/PhysicistsScientificRealism_%20Pre-Print_7%20Jan%202024.pdf

Cova et al. 2021. Cova, Florian, Brent Strickland, André Abatista, et al. 2021. "Estimating the Reproducibility of Experimental Philosophy." *Review of Philosophy and Psychology* 12 (1): 9–44. https://doi.org/10.1007/s13164-020-00458-8 https://digital.csic.es/bitstream/10261/221695/1/Estimating%20the%20Reproducibility.pdf

Naveed et al. 2024. Naveed, Hamza, Abdul Khan, Shuyang Qiu, et al. 2024. "A Comprehensive Overview of Large Language Models." *arXiv preprint* arXiv:2307.06435. https://arxiv.org/abs/2307.06435

 doi:10.20944/preprints202507.1654.v1

Bartels and Pizarro 2011. Bartels, Daniel M., and David A. Pizarro. 2011. "The Mismeasure of Morals: Antisocial Personality Traits Predict Utilitarian Responses to Moral Dilemmas." *Cognition* 121 (1): 154–161. https://doi.org/10.1016/j.cognition.2011.05.010 https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=1937818

James 2000. James, William. 2000. *Pragmatism and Other Writings*. London: Penguin Books.

Searle 1980. Searle, John R. 1980. "Minds, Brains and Programs." *Behavioral and Brain Sciences* 3 (3): 417–457. Cambridge University Press.

Harnad 1991. Harnad, Stevan. 1991. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1–3): 335–346.