# Bridging the Von Neuman Gap: Why LLMs Haven't Made Novel Discoveries

**Anonymous submission**


## Abstract

Large language models (LLMs) have been trained on vast data spanning nearly every scientific discipline, yet they rarely produce meaningful novel discovery. Human polymaths such as John von Neumann routinely generated breakthroughs across disparate fields—from game theory to quantum mechanics to the very architecture of the modern computer—by connecting insights across domains. We argue this gap reflects a structural limitation of the LLM paradigm rather than a problem of scale. Drawing on Piaget's theory of cognitive development and Gentner's structure-mapping, we contend novel discovery depends on two core processes: constructing nuanced internal schemas of the external world and flexibly redeploying them via analogical mapping. Without embodied data or exploration, LLMs form shallow world models; and because their architectures optimize for statistical efficiency, they struggle to extend analogies out of distribution in ways that capture relational structure across domains. Without rethinking training environments and architectures, LLMs will remain constrained to weak abstraction rather than the deep reasoning required for scientific innovation.


## 1 Introduction

Large language models (LLMs) have reached or exceeded human performance in many specialized domains, from mathematics and law to protein structure prediction (Abramson et al. 2024; Zhong et al. 2024). Yet despite this breadth of competence, it's exceedingly rare for an LLM to produce a verifiable novel scientific discovery—a genuine insight not previously known to humans that expands the boundaries of knowledge (Shojaee et al. 2025). This absence is often noted with surprise: if LLMs can solve Olympiad problems, pass graduate-level exams, and synthesize knowledge across disciplines, why have they not combined these abilities to generate groundbreaking, out-of-distribution (OOD) findings?

We argue that this gap is not surprising at all. Drawing from cognitive science and developmental psychology, we propose that human novelty generation depends on two essential ingredients current LLMs lack: (1) the development of robust internal schemas that accurately model the external world, and (2) the flexible redeployment of these schemas to new contexts through analogical reasoning. Piaget's theory of learning shows how embodied experience forms schemas that compress the world into abstract, reusable models (Beilin 1992). Gentner's Structure-Mapping Theory explains how these schemas can be redeployed across domains through relational alignment, enabling the deep analogies behind scientific discovery (Gentner 1983).

Grounding our perspective in Piaget's and Gentner's frameworks, we contend that good schemas lead to good analogies, and good analogies enable novel hypotheses that generalize OOD (Figure 1). Human history provides vivid illustrations of this process. For example, the hydraulic analogy in electricity—conceiving current as fluid flow—helped early scientists reason about voltage, resistance, and circuits (Tembrevilla, Milner-Bolotin, and Petrina 2019). James Clerk Maxwell's vortex analogy in fluid mechanics enabled him to formulate the equations of electromagnetism by mapping fluid vortices onto field lines (Harman 1998). More recently, the discovery of CRISPR–Cas9 gene editing (Jinek et al. 2012) emerged from recognizing that bacterial immune systems could be repurposed as programmable molecular "scissors". Si, Yang, and Hashimoto (2024) found that while LLM-generated hypotheses were judged more novel than human ones, they were significantly less feasible, reflecting a weak causal model of the world. The central challenge lies in moving beyond statistical novelty to schema-based analogical reasoning, the cognitive foundation of historical scientific discovery.

## 2 Cognitive Science Frameworks

### 2.1 Defining Novel Discovery

Although LLMs have shown progress in scientific reasoning, concrete cases of genuine discovery remain rare. DeepMind's *FunSearch* combined an LLM with evolutionary search to generate mathematical programs, one of which produced a valid size-512 Cap Set for $n = 8$; yet the result emerged within a constrained search space through brute-force generation rather than conceptual insight (Romera-Paredes et al. 2024). *AlphaEvolve* discovered a slightly more efficient $4 \times 4$ matrix multiplication algorithm, but this, too, represented optimization within fixed formal rules (Novikov et al. 2025). Sakana AI's *AI Scientist* aimed higher—automating hypothesis generation, experimentation, and paper writing—but nearly half its experiments failed due to coding errors, and its publications lacked factual rigor and conceptual novelty(Beel, Kan, and

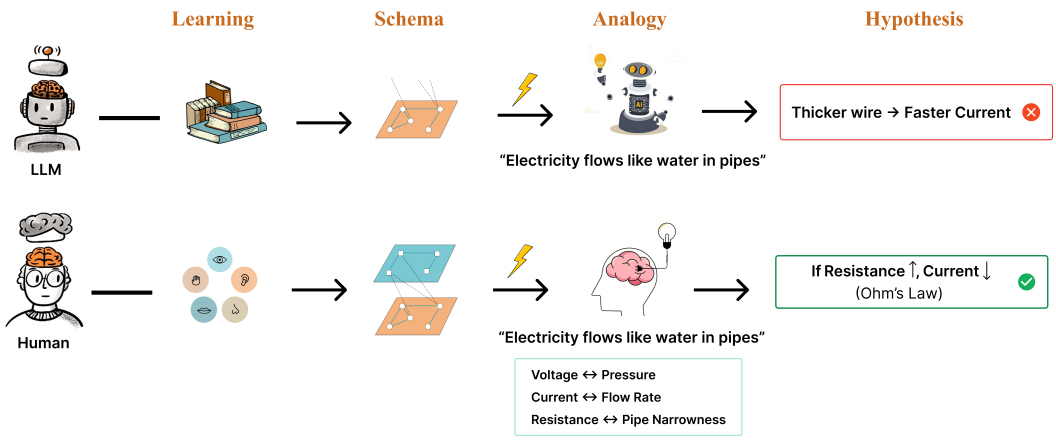


Figure 1: Humans build rich, causal schemas from embodied experience, enabling relational analogies and correct hypotheses (e.g., Ohm's Law). LLMs, trained primarily on text, form flatter associative schemas, leading to surface analogies and brittle, incorrect hypotheses

Baumgart 2025). Even GPT-5's recent assistance in a mathematical proof, noted by Aaronson and Witteveen (2025), refined human reasoning without demonstrating independent abstraction. Collectively, these systems accelerate the process of research but not yet the substance of discovery.

Across these cases, LLMs display what might be called *statistical creativity*: they can search vast combinatorial spaces, recombine prior knowledge, and produce occasionally surprising results. However, that does not make them engines of scientific discovery. Unlike von Neumann, they do not extend beyond their training distribution or formulate hypotheses that challenge the limits of their learned representations. Their successes are bounded by human-defined objectives and preexisting frameworks.

Genuine breakthroughs provide a testable insight about the external world and introduce a principle or relation that extends human knowledge beyond existing frameworks (Kuhn 1962). By these criteria, discoveries such as Newton's law of gravitation, Maxwell's equations, or the CRISPR–Cas9 gene-editing system qualify: they reveal underlying structures of reality that were previously unknown and enabled entire domains of inquiry. Advances such as protein folding exemplify optimization within established frameworks rather than out-of-distribution novelty (Jumper et al. 2021). It's rare to find examples of LLMs synthesizing their massive training dataset into a new conceptual discovery, an imaginative leap like Kekulé's dream of a snake biting its tail (Rocke 2010) that revealed benzene's ring structure.

## 2.2 The von Neumann Gap

We use the term *von Neumann gap* to denote the discrepancy between human polymathic reasoning, exemplified by von Neumann's capacity to connect formal and physical domains, and LLMs' current domain-bounded generalization. Von Neumann's intellectual reach spanned disciplines that rarely intersect: from founding modern game theory and formalizing set theory, to shaping the architecture of the digital computer and contributing to the design of the atomic bomb. His career embodies the kind of integrative reasoning, linking abstract mathematics, physics, and computation, that remains beyond today's models.

The name serves as a conceptual benchmark, not an expectation that models must reach von Neumann's level of genius. Still, we would expect that a model exposed to the full corpus of scientific, mathematical, and psychological knowledge available online could form at least low-hanging novel connections across disciplines. Yet, this kind of integrative reasoning remains absent. Something fundamental about how modern LLMs learn prevents them from bridging that gap.

## 2.3 Piaget's Theory of Cognitive Development

According to Piaget, the goal of learning is to construct the most accurate internal model of the world available at a given time (Beilin 1992). In the sensorimotor stage (0–2 years), children build embodied schemas through direct interaction with the environment, discovering object permanence and forming habits grounded in physical action. In the preoperational stage (2–7 years), schemas become symbolic: words, gestures, and images represent objects and events. By the concrete operational stage (7–11 years), children can run internal "simulations" of their schemas, applying logical operations and reversible reasoning to concrete scenarios. Finally, in the formal operational stage (12+ years), these foundations enable abstract thought, hypothetical reasoning, and systematic problem-solving. As humans learn, schemas reorganize to become increasingly abstract, hierarchical, and nuanced, functioning as cognitive priors for interpreting the world.

## 2.4 Gentner's Structure-Mapping Theory

Gentner's Structure-Mapping Theory provides a cognitive account of how analogy supports abstraction and discovery

(Gentner 1983). Unlike surface similarity, analogy depends on aligning relational structures across domains. In this process, a familiar "base" domain is mapped onto a less familiar "target" domain, with correspondences drawn between underlying causal relations. Gentner formalized this through the systematicity principle, which holds that analogies preserving coherent, interconnected relations are more powerful than those based on isolated features. High-quality analogies are therefore indispensable for novel discovery, because they enable relational structures from well-understood domains to be systematically redeployed in unfamiliar ones. Gentner's theory identifies the cognitive mechanism that allows humans to reason out-of-distribution, moving beyond rote pattern recognition toward flexible, relational inference.

# 3 Why Current LLMs Fail?

## 3.1 Internal World Models

Large Language Models do not merely memorize text—they form implicit internal world models that guide their predictions (Li, Cao, and Cheung 2024). Evidence from mechanistic interpretability shows that even small transformers learn structured representations of game states rather than just token statistics (Li et al. 2023; Karvonen et al. 2024). Recent work further demonstrates that LLMs encode linear spatial world models (Tehenan et al. 2025) and can apply simple heuristics in physical reasoning tasks such as pulleys (Robertson and Wolff 2025). Yet, as Robertson and Wolff (2025) emphasize, these models lack the facility to reason over nuanced structural connectivity, failing when problems demand deeper relational understanding. More broadly, evaluations reveal that world-model coherence often breaks down under perturbation (Vafa et al. 2024).

Piaget's Theory of Cognitive Development makes clear why LLMs fail to form robust world models. The ability to build higher-order abstractions in the formal operational stage depends on foundations laid in earlier stages: spatial grounding in the sensorimotor and preoperational stages, and exploratory simulation in the concrete operational stage. Many of our most powerful scientific analogies—such as electricity flowing like water or the atom resembling a solar system—are rooted in embodied interaction. Exploration enables the counterfactual reasoning that underpins robust schemas—for example, asking *what would happen if resistance increased?* LLMs, by contrast, encounter only linguistic descriptions of these mappings, not the embodied patterns themselves. While text corpora encode valuable abstractions in mathematics, physics, and scientific reasoning, they lack the embodied variation necessary to anchor relational concepts (Bisk et al. 2020). One can read about conservation of energy, but until they interact with systems of push and pull, the concept remains fragile. LLMs lack both of these developmental foundations: spatial understanding of the physical world and self-generated exploration to refine internal schemas.

## 3.2 Limits of Analogical Mapping

The core problem isn't that LLMs can't do analogical mapping—they often succeed at surface-level reasoning (Musker et al. 2025). Transformer attention excels at capturing token co-occurrences and statistical dependencies (Geva et al. 2023; Vig and Belinkov 2019), but this strength becomes brittle out of distribution. The deeper issue is that the next-token prediction paradigm is structurally myopic: trained under teacher forcing, models exploit local token correlations rather than constructing generalizable rules. As Bachmann and Nagarajan (2024) show, this leads to failures even on simple lookahead planning tasks, and Nagarajan et al. (2025) demonstrate similar breakdowns on algorithmic problems requiring novel pattern construction. This brittleness is evident in analogy itself: Lewis and Mitchell (2024) find that while LLMs handle standard analogy problems, they collapse on counterfactual variants. Puranam, Sen, and Workiewicz (2025) confirm this gap empirically, showing that GPT-4 often applies incorrect analogies based on superficial features, while humans generate fewer but causally grounded mappings.

Humans rely on hierarchical schemas grounded in multiple levels of abstraction. Understanding electricity, for example, involves mathematical formalism, intuitive "flow" metaphors, and mechanical analogies simultaneously. This layered structure supports analogical reasoning that is deeply tied to spatial and causal grounding—capacities LLMs lack. Shani et al. (2025) show that LLMs instead perform "aggressive statistical compression," prioritizing efficiency over preserving the fine-grained distinctions essential for human-like reasoning. While this yields broad categorical alignment with human concepts, it erases the typicality gradients and internal semantic structure that enable flexible analogical mapping across domains. By relying on statistical similarity from next-token prediction, LLMs fail to produce the relational, out-of-distribution analogies that Gentner identifies as essential for novel discovery.

# 4 Towards Solutions

## 4.1 Training Environment

**Training with spatial data.** Current LLMs lack genuine spatial grounding (Schulze Buschoff et al. 2025). While "multimodal" models incorporate image–text pairs, the vast majority of their pretraining data remains linguistic(Yin et al. 2024), and structured 3D representations—point clouds, depth maps, volumetric scenes—are virtually absent. Recent work like SpatialLM (Mao et al. 2025) and SpatialVLM (Chen et al. 2024) shows that training on 3D data significantly improves spatial reasoning, but these remain small-scale experiments. We argue that spatial data must become a substantial component of pretraining—comprising 20–30% of training tokens. This includes LiDAR scans, multi-view depth captures, and mesh geometries with explicit spatial relations.

**The field lacks spatial data at scale.** We have trillions of text tokens and billions of images, but structured 3D datasets number in the thousands. Building spatial data infrastructure must be a community priority. Beyond curating existing 3D datasets, the field must actively collect embodied spatial interaction data.

First, everyday spatial interaction: people navigating

buildings, walking through city streets, arranging objects, playing sports, exploring unfamiliar environments. Second, skilled spatial problem-solving: construction workers routing pipes through tight structural constraints, mechanics assembling complex machinery, surgeons operating in 3D anatomical spaces, or movers optimizing furniture placement. Everyday navigation captures foundational spatial operations, while expert tasks reveal how to deploy these operations in complex, constrained problems. Even autonomous vehicles and delivery robots build rich spatial maps with LiDAR and depth sensors that should be incorporated into LLM training. We acknowledge the challenge of collecting spatial data but believe it is crucial for developing LLMs with higher-order reasoning. The gap is fundamental: human children accumulate thousands of hours of physical spatial interaction before abstract reasoning emerges. Current LLMs have a fraction of the embodied experience, yet are expected to develop a robust scientific model of the world.

**Training for exploration.** We argue that exploration should be elevated to a core training phase—a dedicated stage in LLM development where models systematically interact with environments to build causal priors. Evidence supports this claim: fine-tuning LLMs on embodied experiences in environments such as VirtualHome yields over 60% improvements in reasoning tasks by grounding models in object permanence and causal regularities (Xiang et al. 2023). Similarly, embodied agents like STEVE (Zhao et al. 2024) in Minecraft and Voyager (Wang et al. 2023) demonstrate how autonomous exploration can accumulate transferable skills, while S2ERS (Zhang et al. 2025) reduces spatial hallucinations in maze-like planning through reinforcement learning. These results highlight the broader principle: models must actively probe their environments to uncover invariants such as conservation laws and stability.

However, current exploration remains limited to simplified virtual environments with basic physics and discrete state spaces. We propose two pathways forward. First, scaling open-ended environments like Minecraft and MuJoCo to increase physical realism and diversity of spatial challenges. Second, leveraging physics simulators for systematic intervention: For instance, models could test how beam thickness affects load-bearing capacity or how fluid viscosity alters flow patterns. We see particular promise in experiment-based sandboxes, a largely unexplored direction, where models must manipulate environments to rediscover scientific laws from first principles. By repeatedly altering parameters and observing outcomes, models can learn the causal rules that govern systems, developing robust priors for analogy and transfer across domains.

**The core challenge now lies in developing large, diverse, and high-quality environments.** Environments may span simulated physical worlds to structured conceptual domains and should be standardized and scalable so models can act, observe outcomes, and refine their internal representations. Environments provide a path for models to move beyond the ***statistical imitation of expert data***, enabling self-learning through interaction.

### 4.2 Cognitively Aligned Architectures

We contend that progress toward analogical reasoning requires neurosymbolic architectures that combine the explicit relational structures of symbolic systems, with the statistical generalization capabilities of neural networks (Bougzime et al. 2025). Early results support this approach: on Raven's Progressive Matrices, ARLC achieves near-perfect performance by explicitly modeling relational rules (Hersche et al. 2024), while (Shah et al. 2022) show that integrating symbolic background knowledge with neural embeddings enables analogical inferences beyond surface correlations. At the same time, scalability remains an open engineering challenge. Current approaches either collapse into fuzzy embeddings or brittle symbolic rules that fail to generalize (Naik et al. 2024). The next step is to design architectures that learn structured symbolic schemas at scale while retaining generalization, enabling cross-domain analogical reasoning.

## 5 A Case For Evolutionary Pretraining

At first glance, claiming that training on spatial or 3D data is central to breakthroughs in abstract domains like mathematics or biology seems counterintuitive. Yet this perceived mismatch dissolves once we ground the argument in cognitive theory. Humans are not born with the capacity for symbolic reasoning or abstract mathematics. These abilities emerge only after years of embodied interaction—through exploration, manipulation, and spatial reasoning about the physical world. Before any child can understand algebraic equivalence, they have spent years building intuitive models of object permanence, motion, and causality. These embodied experiences constitute the scaffolding upon which higher-order reasoning is built. Current large language models, by contrast, skip this developmental stage entirely. We hand them the equivalent of a "math textbook" without first letting them build an intuitive world model. Consequently, they fail to generate conceptually novel insights or simulate out-of-distribution phenomena.

From this perspective, pretraining should be understood as a developmental process rather than a static data-ingestion phase. **The objective is not to produce a fully intelligent model at the end of pretraining**, but to construct a strong, nuanced prior that equips the system to learn once it is deployed into the world. Evolution offers a deep parallel. Complex cognition did not appear spontaneously; it required the biological "priors" established by an ecological niche that rewarded marginal increases in intelligence (Cisek and Hayden 2022). In response to evolutionary pressure, multicellular eukaryotes developed specialized neurons, nervous systems, and eventually the complex brains underlying cognition (Moroz 2009; Pang et al. 2022). Evolution, in this view, is nature's slow pre-training process: it encodes structural constraints that make learning efficient once an organism is deployed into the world.

Thus, evolutionary pretraining reframes how we think about building general intelligence. Pretraining should not aim to memorize or simulate scientific reasoning directly, but to construct the developmental substrate that makes reasoning possible. The goal is to equip models with the abil-

ity to build meaningful causal abstractions when faced with novel problems in-context. Piaget makes it clear that there is no higher-order reasoning in humans without years of embodied exploration, both mental and physical. By exposing models to a pretraining regimen of spatially grounded, and exploratory environments, we leave LLMs with a deep, intuitive understanding of the world. Once deployed, these priors allow the model to form hypotheses, simulate counterfactuals, and reorganize its understanding in-context. True scientific intelligence, in this sense, will not emerge solely from scaling language models, but from cultivating systems that evolve the capacity to learn about the world the way evolution once did.

## 6 Conclusion

We aim to build intelligence that truly understands the world and internalizes the complex dynamics of how things move and change. That understanding is not meaningfully encoded in textual or visual data. Even simple embodied experiences, like a child rolling a ball downhill, convey physical principles such as motion and gravity that remain opaque to models trained without spatial grounding. Furthermore, without the ability to intervene, tilting the plane or changing the mass, these models cannot internalize causal dynamics. To reason about the world as humans do, models must refine their internal representations to represent concepts across multiple layers of abstraction: spatial, mechanical, and causal.

Bridging the von Neumann gap will require more than scale: it demands training regimes and architectures that embed the principles of human reasoning. Incorporating spatial data and enabling LLMs to explore counterfactuals can anchor robust internal world models, while neurosymbolic architectures may provide the scaffolding for deep analogical inference across domains. Imagine thousands of polymathic systems like Von Neumann, each capable of connecting insights across disparate fields and generating unexpected analogies that drive scientific breakthroughs. We urge the AI community to ground future research in cognitive theory, so that AI moves beyond statistical efficiency and emerges as an engine of unprecedented discovery.

## References


Aaronson, S.; and Witteveen, F. 2025. Limits to black-box amplification in QMA. *arXiv preprint arXiv:2509.21131*.

Abramson, J.; Adler, J.; Dunger, J.; et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630: 493–500.

Bachmann, G.; and Nagarajan, V. 2024. The pitfalls of next-token prediction. *arXiv preprint arXiv:2403.06963*.

Beel, J.; Kan, M.-Y.; and Baumgart, M. 2025. Evaluating Sakana's AI Scientist: Bold Claims, Mixed Results, and a Promising Future? In *ACM SIGIR Forum*, volume 59, 1–20. ACM New York, NY, USA.

Beilin, H. 1992. Piaget's enduring contribution to developmental psychology. *Developmental Psychology*, 28(2): 191–204.

Bisk, Y.; Holtzman, A.; Thomason, J.; Andreas, J.; Bengio, Y.; Chai, J.; Lapata, M.; Lazaridou, A.; May, J.; Nisnevich, A.; Pinto, N.; and Turian, J. 2020. Experience Grounds Language. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–8735. Online: Association for Computational Linguistics.

Bougzime, O.; Jabbar, S.; Cruz, C.; and Demoly, F. 2025. Unlocking the Potential of Generative AI through Neuro-Symbolic Architectures: Benefits and Limitations. *arXiv preprint arXiv:2502.11269*.

Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Sadigh, D.; Guibas, L.; and Xia, F. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14455–14465.

Cisek, P.; and Hayden, B. Y. 2022. Neuroscience needs evolution.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2): 155–170.

Geva, M.; Bastings, J.; Filippova, K.; and Globerson, A. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*.

Harman, P. M. 1998. *The Natural Philosophy of James Clerk Maxwell*. Cambridge, UK: Cambridge University Press.

Hersche, M.; Camposampiero, G.; Wattenhofer, R.; Sebastian, A.; and Rahimi, A. 2024. Towards Learning to Reason: Comparing LLMs with Neuro-Symbolic on Arithmetic Relations in Abstract Reasoning. *arXiv preprint arXiv:2412.05586*.

Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J. A.; and Charpentier, E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096): 816–821. Epub 2012 Jun 28.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873): 583–589.

Karvonen, A.; Wright, B.; Rager, C.; Angell, R.; Brinkmann, J.; Smith, L.; Mayrink Verdun, C.; Bau, D.; and Marks, S. 2024. Measuring progress in dictionary learning for language model interpretability with board game models. *Advances in Neural Information Processing Systems*, 37: 83091–83118.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press. ISBN 9780226458086.

Lewis, M.; and Mitchell, M. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*.

Li, K.; Hopkins, A. K.; Bau, D.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*.

Li, Z.; Cao, Y.; and Cheung, J. C. 2024. Do LLMs Build World Representations? Probing Through the Lens of State Abstraction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Mao, Y.; Zhong, J.; Fang, C.; Zheng, J.; Tang, R.; Zhu, H.; Tan, P.; and Zhou, Z. 2025. SpatialLM: Training Large Language Models for Structured Indoor Modeling. *arXiv preprint arXiv:2506.07491*.

Moroz, L. L. 2009. On the independent origins of complex brains and neurons. *Brain Behavior and Evolution*, 74(3): 177–190.

Musker, S.; Duchnowski, A.; Millière, R.; and Pavlick, E. 2025. LLMs as models for analogical reasoning. *Journal of Memory and Language*, 145: 104676.

Nagarajan, V.; Wu, C. H.; Ding, C.; and Raghunathan, A. 2025. Roll the dice & look before you leap: Going beyond the creative limits of next-token prediction. In *Forty-second International Conference on Machine Learning*.

Naik, A.; Liu, J.; Wang, C.; Sethi, A.; Dutta, S.; Naik, M.; and Wong, E. 2024. Dolphin: A programmable framework for scalable neurosymbolic learning. *arXiv preprint arXiv:2410.03348*.

Novikov, A.; Vũ, N.; Eisenberger, M.; Dupont, E.; Huang, P.-S.; Wagner, A. Z.; Shirobokov, S.; Kozlovskii, B.; Ruiz, F. J.; Mehrabian, A.; et al. 2025. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*.

Pang, J. C.; Rilling, J. K.; Roberts, J. A.; Van Den Heuvel, M. P.; and Cocchi, L. 2022. Evolutionary shaping of human brain dynamics. *Elife*, 11: e80627.

Puranam, P.; Sen, P.; and Workiewicz, M. 2025. Can LLMs Help Improve Analogical Reasoning For Strategic Decisions? Experimental Evidence from Humans and GPT-4. *arXiv preprint arXiv:2505.00603*.

Robertson, C.; and Wolff, P. 2025. LLM world models are mental: Output layer evidence of brittle world model use in LLM mechanical reasoning. *arXiv preprint arXiv:2507.15521*.

Rocke, A. J. 2010. *Image and Reality: Kekulé, Kopp, and the Scientific Imagination*. Chicago, IL: University of Chicago Press. ISBN 9780226723320.

Romera-Paredes, B.; Barekatain, M.; Novikov, A.; Balog, M.; Kumar, M. P.; Dupont, E.; Ruiz, F. J.; Ellenberg, J. S.; Wang, P.; Fawzi, O.; et al. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995): 468–475.

Schulze Buschoff, L. M.; Akata, E.; Bethge, M.; and Schulz, E. 2025. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1): 96–106.

Shah, V.; Sharma, A.; Shroff, G.; Vig, L.; Dash, T.; and Srinivasan, A. 2022. Knowledge-based analogical reasoning in neuro-symbolic latent spaces. *arXiv preprint arXiv:2209.08750*.

Shani, C.; Jurafsky, D.; LeCun, Y.; and Shwartz-Ziv, R. 2025. From tokens to thoughts: How LLMs and humans trade compression for meaning. *arXiv preprint arXiv:2505.17117*.

Shojaee, P.; Nguyen, N.-H.; Meidani, K.; Farimani, A. B.; Doan, K. D.; and Reddy, C. K. 2025. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*.

Si, C.; Yang, D.; and Hashimoto, T. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Tehenan, M.; Moya, C. B.; Long, T.; and Lin, G. 2025. Linear Spatial World Models Emerge in Large Language Models. *arXiv preprint arXiv:2506.02996*.

Tembrevilla, G.; Milner-Bolotin, M.; and Petrina, S. 2019. Electric fluid to electric current: The problematic attempts of abstraction to concretization. arXiv:1910.02762.

Vafa, K.; Chen, J. Y.; Rambachan, A.; Kleinberg, J.; and Mullainathan, S. 2024. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37: 26941–26975.

Vig, J.; and Belinkov, Y. 2019. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Xiang, J.; Tao, T.; Gu, Y.; Shu, T.; Wang, Z.; Yang, Z.; and Hu, Z. 2023. Language models meet world models: Embodied experiences enhance language models. *Advances in neural information processing systems*, 36: 75392–75412.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Zhang, H.; Deng, H.; Ou, J.; et al. 2025. Mitigating spatial hallucination in large language models for path planning via prompt engineering. *Scientific Reports*, 15: 8881.

Zhao, Z.; Chai, W.; Wang, X.; Li, B.; Hao, S.; Cao, S.; Ye, T.; and Wang, G. 2024. See and think: Embodied agent in virtual environment. In *European Conference on Computer Vision*, 187–204. Springer.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2024. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 2299–2314. Mexico City, Mexico: Association for Computational Linguistics.