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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models (LLMs) have been trained on vast data spanning nearly every scientific discipline, yet they have not produced a single 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, no LLM to date has produced 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,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

and circuits [Tembrevilla et al., 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 et al. [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.

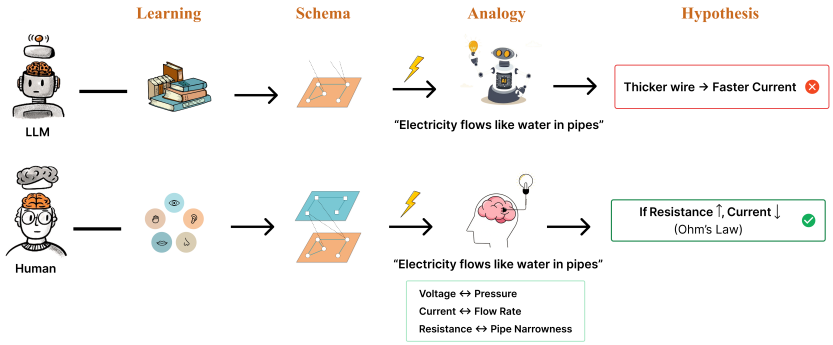

Figure 1: Humans build rich, causal schemas from embodied experience, enabling relational analogies and correct hypotheses (e.g., Ohm's Law). LLMs, trained primarily on text, form flatter associative schemas, leading to surface analogies and brittle, incorrect hypotheses

2 Cognitive Science Frameworks

2.1 Defining Novel Discovery

In a trivial sense, language models constantly produce "new" ideas: every draft completion, bug fix suggestion, or paraphrase is technically novel. Genuine breakthroughs, however, 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. 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 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.

2.3 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 et al., 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 et al. [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. Exposing LLMs to spatially grounded information can significantly improve their reasoning about the physical world. SpatialLM, trained on large-scale synthetic 3D indoor scenes, improves layout estimation and 3D object detection [Mao et al., 2025]. SpatialVLM scales relational spatial reasoning through billions of spatially grounded vision–language examples [Chen et al., 2024]. Han et al. [2025] highlights the importance of incorporating computational geometry, 3D point clouds (e.g., LiDAR scans), and CAD/architectural models into LLM training pipelines to advance spatial reasoning. Moving forward, researchers should prioritize 3D scene data, multimodal robotics traces, and physics-based simulation datasets, since these expose models to the structural relationships necessary for robust internal world models.

Training for exploration. Exploration is critical for building robust world models, enabling LLMs to internalize causal dynamics. 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. We propose two pathways forward—scaling open-ended virtual environments like Minecraft and MuJoCo to approximate embodied play, and leveraging scientific simulators such as SPICE or fluid dynamics solvers for controlled, intervention-rich exploration. We see particular promise in physics-based sandboxes, 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.

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 Conclusion

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

J. Abramson, J. Adler, J. Dunger, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. *arXiv preprint arXiv:2403.06963*, 2024.

Harry Beilin. Piaget's enduring contribution to developmental psychology. *Developmental Psychology*, 28(2):191–204, 1992. doi: 10.1037/0012-1649.28.2.191.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.703. URL `https://aclanthology.org/2020.emnlp-main.703/`.

Oualid Bougzime, Samir Jabbar, Christophe Cruz, and Frédéric Demoly. Unlocking the potential of generative ai through neuro-symbolic architectures: Benefits and limitations. *arXiv preprint arXiv:2502.11269*, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7 (2):155–170, 1983. ISSN 0364-0213. doi: 10.1016/S0364-0213(83)80009-3. URL `https://www.sciencedirect.com/science/article/pii/S0364021383800093`.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*, 2023.

Bin Han, Robert Wolfe, Anat Caspi, and Bill Howe. Can large language models integrate spatial data? empirical insights into reasoning strengths and computational weaknesses. *arXiv preprint arXiv:2508.05009*, 2025.

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

Michael Hersche, Giacomo Camposampiero, Roger Wattenhofer, Abu Sebastian, and Abbas Rahimi. Towards learning to reason: Comparing llms with neuro-symbolic on arithmetic relations in abstract reasoning. *arXiv preprint arXiv:2412.05586*, 2024.

Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821, August 2012. doi: 10.1126/science.1225829. Epub 2012 Jun 28.

Adam Karvonen, Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Smith, Claudio Mayrink Verdun, David Bau, and Samuel Marks. Measuring progress in dictionary learning for language model interpretability with board game models. *Advances in Neural Information Processing Systems*, 37:83091–83118, 2024.

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

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

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*, 2023.

Zichao Li, Yanshuai Cao, and Jackie CK Cheung. Do LLMs build world representations? probing through the lens of state abstraction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=lzfzjYuWgY.

Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. Spatiallm: Training large language models for structured indoor modeling. *arXiv preprint arXiv:2506.07491*, 2025.

Sam Musker, Alex Duchnowski, Raphaël Millière, and Ellie Pavlick. Llms as models for analogical reasoning. *Journal of Memory and Language*, 145:104676, 2025.

Vaishnavh Nagarajan, Chen Henry Wu, Charles Ding, and Aditi Raghunathan. Roll the dice & look before you leap: Going beyond the creative limits of next-token prediction. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Hi0SyHMmkd.

Aaditya Naik, Jason Liu, Claire Wang, Amish Sethi, Saikat Dutta, Mayur Naik, and Eric Wong. Dolphin: A programmable framework for scalable neurosymbolic learning. *arXiv preprint arXiv:2410.03348*, 2024.

Phanish Puranam, Prothit Sen, and Maciej Workiewicz. Can llms help improve analogical reasoning for strategic decisions? experimental evidence from humans and gpt-4. *arXiv preprint arXiv:2505.00603*, 2025.

Cole Robertson and Philip Wolff. Llm world models are mental: Output layer evidence of brittle world model use in llm mechanical reasoning. *arXiv preprint arXiv:2507.15521*, 2025.

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

Vishwa Shah, Aditya Sharma, Gautam Shroff, Lovekesh Vig, Tirtharaj Dash, and Ashwin Srinivasan. Knowledge-based analogical reasoning in neuro-symbolic latent spaces. *arXiv preprint arXiv:2209.08750*, 2022.

Chen Shani, Dan Jurafsky, Yann LeCun, and Ravid Shwartz-Ziv. From tokens to thoughts: How llms and humans trade compression for meaning. *arXiv preprint arXiv:2505.17117*, 2025.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K Reddy. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*, 2025.

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

Matthieu Tehenan, Christian Bolivar Moya, Tenghai Long, and Guang Lin. Linear spatial world models emerge in large language models. *arXiv preprint arXiv:2506.02996*, 2025.

Gerald Tembrevilla, Marina Milner-Bolotin, and Stephen Petrina. Electric fluid to electric current: The problematic attempts of abstraction to concretization, 2019. URL https://arxiv.org/abs/1910.02762.

Keyon Vafa, Justin Y Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024.

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

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. Language models meet world models: Embodied experiences enhance language models. *Advances in neural information processing systems*, 36:75392–75412, 2023.

H. Zhang, H. Deng, J. Ou, et al. Mitigating spatial hallucination in large language models for path planning via prompt engineering. *Scientific Reports*, 15:8881, 2025. doi: 10.1038/s41598-025-93601-5. URL https://doi.org/10.1038/s41598-025-93601-5.

Zhonghan Zhao, Wenhao Chai, Xuan Wang, Boyi Li, Shengyu Hao, Shidong Cao, Tian Ye, and Gaoang Wang. See and think: Embodied agent in virtual environment. In *European Conference on Computer Vision*, pages 187–204. Springer, 2024.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.149. URL https://aclanthology.org/2024.findings-naacl.149/.