Without Wonder, LLMs Are Analysts, Not Thinkers: What Research Says About AI Creativity

Kabui, Charles

2026-03-13

Table of Contents

Introduction

A child looks at the sky and asks, "Why is it blue?" Nobody prompted that question. Nobody gave the child a task. The question came from wonder, a spontaneous desire to understand something that does not yet make sense.

Large language models cannot do this. They can answer the question brilliantly. They can explain Rayleigh scattering in six languages. But they will never, unprompted, look at the sky and ask "why." This gap between answering questions and asking them turns out to be far more consequential than most AI discussions acknowledge. Research from 2024 and 2025 shows that what we call "LLM reasoning" is closer to sophisticated pattern retrieval, and that genuine creativity requires exactly the kind of self-generated curiosity that these systems lack.

---

1. The Curiosity Gap Is Measurable

Wonder is not just a philosophical abstraction. Researchers have built frameworks to test whether LLMs exhibit anything resembling genuine curiosity.

Wang et al. (2025) directly tackled this in *"Why Did Apple Fall To The Ground,"* adapting the Five-Dimensional Curiosity Scale (a validated human psychology instrument) for LLMs

(arXiv:2510.20635). Their findings were revealing: **LLMs show a stronger "thirst for knowledge" than humans when asked factual questions**, but they **make conservative choices when faced with uncertain environments**. In other words, models eagerly retrieve information they have already seen, but shrink from genuine unknowns. That is the opposite of wonder. A curious child runs toward the unfamiliar. An LLM retreats to its training data.

Borah, Jin, and Mihalcea (2025) confirmed this from a cultural angle in *"The Curious Case of Curiosity"* (arXiv:2510.12943). They found that **LLMs flatten cross-cultural diversity in curiosity**, aligning most closely with Western question-asking patterns. Fine-tuning narrowed the gap by up to **50%**, but the core problem persisted: LLMs simulate a single curiosity style rather than genuinely exploring.

Ceraolo et al. (2024) built Quriosity, a dataset of **13,500 naturally occurring questions** from search engines, human conversations, and LLM chats (arXiv:2405.20318). Their analysis found that **up to 42% of human curiosity-driven questions are causal**, asking "why" something happens. Humans instinctively seek causes. LLMs, by contrast, only engage with causality when explicitly instructed to do so.

---

2. Without Wonder, "Reasoning" Becomes Pattern Matching

If an LLM cannot generate its own questions, what is it actually doing when it appears to reason? A growing body of evidence suggests: sophisticated retrieval and recombination.

Jiang et al. (2024) tested this directly in *"Large Language Models Are Not Yet Genuine Reasoners"* (arXiv:2406.11050), published at EMNLP 2024. Using controlled synthetic problems, they showed that **LLM performance depends heavily on superficial token patterns rather than logical structure**. Models performed well on classic reasoning problems but struggled when the same logic was expressed with different surface features. Their success, with statistical guarantees, was tied to recognising patterns, not understanding relationships.

Mukhopadhyay et al. (2025) pushed this further with PHANTOM RECALL (arXiv:2510.11812). They took 25 well-known logic puzzles and created 149 careful variations that preserved the reasoning structure but changed superficial details. The results were stark: **models scored near-perfectly on unmodified puzzles but significantly underperformed humans on modified versions.** The models confidently reproduced memorised solutions even when those solutions no longer fit, a failure the authors call "phantom recall." You cannot wonder whether your remembered answer still applies if you lack the capacity to wonder at all.

Liu et al. (2025) tested an even more demanding scenario with Camlang, a constructed language designed to evaluate metalinguistic reasoning (arXiv:2509.00425). **GPT-5 scored 98% on English tasks but only 47% on the same tasks in Camlang, far below the human**

score of 87%. When models could not rely on familiar patterns, they fell apart. Human verification revealed that most model successes came from shallow word-matching, not genuine grammatical understanding.

---

3. The Creativity Problem Goes Deeper Than Output Quality

Creativity requires more than producing novel combinations. It requires wondering whether something could be different, then exploring that possibility. Research on LLM creativity reveals a fundamental structural limitation.

Ruiz Luyten and van der Schaar (2026) formalised this in *"The Reasoning-Creativity Trade-off"* (arXiv:2601.00747). They proved mathematically that standard LLM training pipelines cause "diversity decay," where models collapse toward a narrow set of solution paths. Optimising for correctness systematically eliminates creative exploration. The model learns to give the right answer, not to wonder if there is a better question.

Nguyen and Singla (2025) demonstrated this "Artificial Hivemind" effect empirically (arXiv:2512.23601). When generating educational problems, **LLMs produced homogeneous outputs both within the same model and across different models.** Their proposed fix, CreativeDC, had to explicitly force the model through separate "divergent" and "convergent" phases to mimic the creative process that humans perform naturally through wonder.

Banerjee et al. (2025) found a direct tension between accuracy and creativity in *"Does Less Hallucination Mean Less Creativity?"* (arXiv:2512.11509), accepted at the AAAI 2026 Workshop. Testing models from **1B to 70B parameters**, they showed that methods designed to reduce factual errors had opposing effects on creative thinking. Chain of Verification actually enhanced divergent thinking, while Decoding by Contrasting Layers suppressed it. This confirms that LLM creativity is not a stable capacity. It is a side effect of how the model is configured, not something the model intrinsically pursues.

---

4. Reasoning vs. Analysis, a Critical Distinction

Here is the distinction that matters. When you wonder why humans grow upward instead of downward, you are doing something an LLM cannot: generating a question from a felt sense of strangeness about the world. An LLM can analyse that question once you pose it. It can discuss gravity, plant biology, and evolutionary pressures. But it will never originate that

question, because nothing about the concept strikes it as surprising. It has no expectations to violate.

Hong et al. (2025) found concrete evidence for this split at the mechanistic level (arXiv:2503.23084). They identified **a single linear feature in model internals that governs the switch between reasoning and memorisation**. Manipulating this feature causally changed model performance, confirming that LLMs have separable "modes" for retrieval and reasoning rather than an integrated understanding.

Khalid, Nourollah, and Schockaert (2025) went further in *"Large Language and Reasoning Models are Shallow Disjunctive Reasoners,"* published at ACL 2025 (arXiv:2503.23487). They tested both standard LLMs and newer "reasoning models" on spatial and temporal reasoning tasks. **Reasoning models outperformed LLMs on single-path problems but still struggled with multi-path reasoning**, where you must hold multiple possibilities in mind simultaneously. The authors describe this as "shallow disjunctive reasoning," the model picks one path and follows it, rather than genuinely exploring the space. Without wonder, there is no motivation to ask "what if I tried the other path?"

Steyvers and Peters (2025) connected this to metacognition, the ability to monitor and evaluate your own thinking, in *Current Directions in Psychological Science* (arXiv:2504.14045). They found that **while humans and LLMs sometimes appear aligned in metacognitive capacity, fundamental differences remain**. Genuine curiosity requires knowing what you do not know and wanting to fix that. LLMs lack this self-directed motivation entirely, which is why giving them better self-awareness is still an open problem.

---

5. What This Means for Human-AI Collaboration

Understanding the wonder gap changes how we should use these tools. LLMs are powerful analysers. Give them a well-formed question and structured data, and they produce useful outputs. But they cannot replace the human capacity to notice something odd, formulate a question nobody has asked before, and pursue it purely because it is interesting.

The research points to a practical division of labour. Humans generate the questions, spot the anomalies, and feel the productive confusion that drives discovery. LLMs handle the analysis, retrieve relevant information, and explore the space of known solutions. The danger is assuming LLMs can do both.

Ceraolo et al. (2024) found that **42% of human curiosity-driven questions are causal**, the "why" questions that drive science, engineering, and creative breakthroughs (arXiv:2405.20318). LLMs do not spontaneously generate these. Every scientific revolution started with someone wondering why the existing explanation felt wrong. That capacity has no equivalent in current AI systems.

The most productive approach treats LLMs as what they are: exceptionally fast, broad-knowledge analytical engines. Not thinkers. Not wonderers. Not creative partners. Engines that need a human to point them at the right question.

---

6. Conclusion

The research paints a consistent picture across curiosity studies, reasoning benchmarks, creativity evaluations, and mechanistic analyses.

- **LLMs retrieve eagerly but never wonder.** They excel at answering questions from their training data but cannot generate novel questions from genuine surprise or confusion.
- **What looks like reasoning is often pattern matching.** Performance collapses when surface features change, dropping from 98% to 47% on the same logical tasks expressed unfamiliarly.
- **Creativity is a side effect, not a capacity.** LLM creative output depends on configuration choices, not self-directed exploration. Standard training actively decays diversity.
- **The reasoning-memorisation boundary is mechanistically real.** A single internal feature toggles between modes, confirming these are separate processes, not integrated understanding.
- **Human wonder remains the engine of discovery.** Up to 42% of naturally occurring curiosity-driven questions are causal. LLMs don't produce such questions unprompted.

LLMs are extraordinary at analysis. They are the best research assistants we have ever built. But an assistant that never wonders, never notices, and never asks "why" on its own is fundamentally different from a creative mind. Knowing this difference is not a limitation. It is the key to using these tools well.

References

1. Wang, H., et al. (2025). *Why Did Apple Fall To The Ground: Evaluating Curiosity In Large Language Model.* arXiv:2510.20635.
2. Borah, A., Jin, Z. & Mihalcea, R. (2025). *The Curious Case of Curiosity across Human Cultures and LLMs.* arXiv:2510.12943.
3. Ceraolo, R., et al. (2024). *Quriosity: Analyzing Human Questioning Behavior and Causal Inquiry through Curiosity-Driven Queries.* arXiv:2405.20318. IJCNLP-AACL 2025 Findings.

4. Jiang, B., et al. (2024). *A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners.* arXiv:2406.11050. EMNLP 2024.

5. Mukhopadhyay, S., et al. (2025). *PHANTOM RECALL: When Familiar Puzzles Fool Smart Models.* arXiv:2510.11812.

6. Liu, F., et al. (2025). *The Gold Medals in an Empty Room: Diagnosing Metalinguistic Reasoning in LLMs with Camlang.* arXiv:2509.00425.

7. Ruiz Luyten, M. & van der Schaar, M. (2026). *The Reasoning-Creativity Trade-off: Toward Creativity-Driven Problem Solving.* arXiv:2601.00747.

8. Nguyen, M. H. & Singla, A. (2025). *Divergent-Convergent Thinking in Large Language Models for Creative Problem Generation.* arXiv:2512.23601.

9. Banerjee, M., et al. (2025). *Does Less Hallucination Mean Less Creativity? An Empirical Investigation in LLMs.* arXiv:2512.11509. AAAI 2026 Workshop.

10. Hong, Y., et al. (2025). *The Reasoning-Memorization Interplay in Language Models Is Mediated by a Single Direction.* arXiv:2503.23084.

11. Khalid, I., Nourollah, A. M. & Schockaert, S. (2025). *Large Language and Reasoning Models are Shallow Disjunctive Reasoners.* arXiv:2503.23487. ACL 2025.

12. Steyvers, M. & Peters, M. A. K. (2025). *Metacognition and Uncertainty Communication in Humans and Large Language Models.* arXiv:2504.14045. Current Directions in Psychological Science.
