University of Montreal Tests AI Against 100,000 Humans on Creativity—GPT-4 Beats 72% But Top 10% Still Win

GPT-4 now outperforms the average human on standardized creativity tests. But before you fire your creative team, consider this: when researchers pushed the model harder, it obsessively inserted “microscope” into 70% of its responses.

The Largest Creativity Benchmark Ever Conducted

The University of Montreal study published January 25, 2026 tested GPT-4, Claude, Gemini, and several other large language models against more than 100,000 human participants. The battery included the Divergent Association Task (DAT)—a validated measure of creative thinking—along with haiku composition and flash fiction challenges.

The headline result: GPT-4, when optimized with adjusted temperature parameters, scored higher than 72% of all human participants on divergent linguistic creativity tasks.

This isn't a small sample with questionable methodology. One hundred thousand humans makes this the largest creativity comparison study ever published. The statistical power here is enormous. We now have robust evidence that frontier AI models have crossed the average human creativity threshold on standardized measures.

But the study's most important finding isn't about AI capabilities. It's about AI limitations.

The top 10% most creative humans significantly outperformed every AI model tested. Not by a small margin—the gap was substantial and consistent across all task types. Even more striking: the most creative half of humans beat every AI system in the study.

The Microscope Problem: What Repetition Bias Reveals

Before temperature optimization, GPT-4 exhibited what researchers called "severe repetition bias." The word "microscope" appeared in 70% of responses. "Elephant" showed up in 60%. These weren't tasks about scientific instruments or African wildlife—the model was fixating on specific tokens regardless of context.

This isn't a bug you can patch with prompt engineering. It reveals something fundamental about how these models generate text.

When you ask a human to brainstorm creative ideas, they naturally monitor their own output. They notice when they're repeating themselves. They feel the staleness of using the same word twice in a paragraph. This metacognitive loop—the ability to evaluate one's own creative output in real-time—is automatic in humans.

LLMs lack this capability entirely.

The Montreal researchers specifically noted that GPT-4o shows fixation bias comparable to humans but cannot evaluate whether its own ideas are original. The

model has no internal sense of “I’ve said this before” or “this feels derivative.”

Temperature adjustment partially addresses this by injecting randomness into token selection. Higher temperature means the model is more likely to pick lower-probability words, reducing repetition. But this is a statistical workaround, not a solution. You’re fighting the model’s natural tendencies rather than giving it genuine creative judgment.

Why Temperature Is a Crude Tool for Creativity

Temperature in language models controls the probability distribution over next-token predictions. At temperature 0, the model always picks the highest-probability token—deterministic and repetitive. At temperature 1, it samples according to the learned distribution. Above 1, lower-probability tokens become increasingly likely, introducing more variation.

The problem: higher temperature doesn’t make outputs more creative. It makes them more random. These aren’t the same thing.

Human creativity involves generating novel combinations that are both unexpected AND appropriate. A surrealist poem that surprises you still needs internal coherence. A plot twist needs to feel surprising yet inevitable in retrospect. Cranking temperature creates outputs that might be less predictable but are often less coherent, less purposeful, less meaningful.

The Montreal study’s optimization process found temperature settings that maximized DAT scores specifically. But the DAT measures divergent thinking—the ability to generate semantically distant word associations. This is one component of creativity, not the whole picture.

> Temperature optimization improved GPT-4’s divergent thinking scores without improving its actual creative work. The haiku were still derivative. The flash fiction still lacked genuine surprise.

What the Coverage Gets Wrong

Most reporting on this study falls into one of two traps: either “AI is now more

creative than humans!” or “Human creativity still wins!” Both framings miss the point.

The Benchmarking Fallacy

The DAT is a useful psychometric tool. It’s validated, reliable, and measures something real about cognitive flexibility. But it measures a specific type of divergent linguistic thinking—the ability to generate words that are semantically distant from each other.

Creative work in the real world involves far more than this. It requires:

- **Constraint satisfaction:** Working within format, tone, brand voice, technical requirements
- **Cultural context:** Understanding what audiences will find fresh versus clichéd
- **Emotional resonance:** Creating work that moves people, not just surprises them
- **Iterative refinement:** Recognizing which ideas deserve development and which should be killed
- **Strategic intent:** Knowing what the creative work needs to accomplish

The study found that human writers produced creative work “with greater variety and originality, particularly in poetry and plot summaries.” This aligns with what anyone who’s used LLMs for serious creative work already knows: the models can generate plausible first drafts but struggle with the judgment required to make something genuinely good.

The Average Human Isn’t Your Competition

If you’re a CTO or senior engineer reading this, you’re not competing with average human creativity. You’re hiring from the top quartile—or trying to. Your writers, designers, and product thinkers are selected for creative capability.

The study explicitly shows that the most creative half of humans outperformed every AI system tested. If your team has any selection process at all, you’re working with people who beat GPT-4 on these measures.

This doesn’t mean AI tools aren’t useful for creative work. It means the framing of “AI versus humans” obscures the more interesting question: how do skilled humans

use AI tools to extend their capabilities?

The Technical Reality of AI Creativity

To understand why the top 10% of humans remain unreachable for current models, we need to examine what LLMs actually do when they generate “creative” output.

Interpolation Versus Extrapolation

Language models learn statistical patterns from training data. When they generate text, they’re essentially interpolating within the space of patterns they’ve seen. A model trained on millions of poems can combine elements of those poems in new configurations. It can produce a sonnet about quantum physics because it’s seen sonnets and it’s seen text about quantum physics.

What it cannot do is extrapolate beyond this space in meaningful ways. Genuine creative breakthroughs—the ones that earn artists careers and win scientists Nobel Prizes—involve stepping outside existing patterns, not recombining them.

Consider what made Shakespeare’s metaphors striking in their time, or what made Coltrane’s harmonic innovations matter. These weren’t combinations of existing elements. They were expansions of the possibility space itself.

Current LLM architectures are mathematically incapable of this type of extrapolation. They’re compression algorithms that have learned the statistical structure of creative work. They can produce outputs that look like creative work. They cannot produce outputs that expand what creative work means.

The Absence of Ground Truth in Creative Evaluation

When we train models for tasks like image classification or protein folding, we have ground truth labels. A picture either contains a cat or it doesn’t. A protein structure either matches experimental data or it doesn’t. We can measure whether model outputs are correct.

Creativity has no equivalent ground truth. We can’t compute a loss function for “how creative is this haiku?” So we resort to proxy measures—the DAT, human ratings, or metrics like novelty and diversity in generated outputs.

These proxies are leaky. A model can optimize for DAT scores without optimizing for the underlying trait we actually care about. This is Goodhart's Law applied to creativity: when a measure becomes a target, it ceases to be a good measure.

The Montreal study used validated instruments and large samples. It's good science. But it's measuring what's measurable, which isn't the same as measuring what matters for your creative workflows.

Google Gemini Pro: The Interesting Middle Case

Google Gemini Pro achieved scores statistically close to average human creativity levels without the temperature optimization that GPT-4 required. This suggests different architectures handle divergent thinking differently.

We don't have visibility into Gemini's exact training methodology, but the result implies that some approaches may naturally produce more varied outputs. Whether this translates to better real-world creative applications is unproven.

For technical leaders evaluating these systems, the implication is that benchmark performance on creativity measures may not correlate perfectly with usefulness for creative tasks. A model that scores slightly lower on DAT but requires less aggressive temperature tuning might produce more consistently usable output.

This is an area where vendor claims need scrutiny. "More creative" is a marketing phrase, not a technical specification.

What This Means for Your AI Strategy

The Montreal study provides data to inform decisions you're probably already making about AI in creative workflows. Here's how to interpret the results:

Use AI for Volume, Humans for Quality

If your creative needs involve generating many options quickly—ad variations, headline alternatives, product descriptions—AI tools now demonstrably perform at or above average human levels. For tasks where "good enough" times a thousand beats "excellent" times ten, the math works.

For work where quality matters more than quantity—brand campaigns, thought

leadership, products where creative excellence is a competitive advantage—the top 10% human gap remains significant. Invest in talented people and give them AI tools, rather than replacing them with AI outputs.

Build for Human-AI Collaboration, Not Replacement

The most productive configuration isn't AI generating final creative work. It's AI generating raw material that humans refine.

This requires building workflows where:

- AI outputs are explicitly treated as drafts
- Human editors have authority to reject, substantially revise, or restart
- Feedback loops exist to track which AI outputs consistently require heavy editing
- Creative professionals understand the model's biases (like the microscope problem) and can compensate

The danger is automating too much and ending up with output that's technically competent but generically bland. If your competitors do the same thing, you've commoditized your creative work without gaining advantage.

Monitor for Repetition Bias in Production

The 70% microscope figure should alarm anyone running LLMs in creative applications. If you're generating customer-facing content at scale, you need monitoring systems that detect repetition patterns.

This isn't standard in most LLM deployment frameworks. You'll need to build or buy detection for:

- Unusual token frequency across outputs
- Semantic clustering indicating the model is stuck in a narrow space
- Temporal patterns showing repetition increasing over time

Temperature adjustment helps but doesn't eliminate the problem. Treat repetition monitoring as essential infrastructure, not optional.

Don’t Over-Index on Benchmark Improvements

The study found GPT-4 scored higher than 72% of humans after optimization. Future models will score higher. The temptation will be to conclude that AI has “solved” creativity when benchmarks reach 90% or 95%.

Resist this. The benchmarks measure a subset of creative capability. They don’t measure the judgment, cultural awareness, and intentionality that distinguish professional creative work from impressive party tricks.

Track real-world outcomes in your creative workflows—engagement metrics, conversion rates, brand perception studies—not just benchmark scores when evaluating whether to increase AI involvement.

The Six-Month Outlook

Based on the Montreal findings and current model trajectories, here’s what technical leaders should expect:

Benchmark scores will continue rising, but the top-10% gap will persist

Next-generation models will beat 80%, then 85% of humans on measures like the DAT. But the architectural limitations that prevent genuine extrapolation aren’t being addressed by current research directions. The gap between “statistically creative” and “genuinely creative” will remain.

Repetition bias will get better but not disappear

Model developers are aware of fixation problems. Expect improved diversity in outputs through better sampling strategies, constitutional AI approaches, or architectural changes. But the fundamental lack of metacognitive evaluation means humans will still need to catch repetition that slips through.

Industry-specific creative benchmarks will emerge

The DAT is domain-general. Expect specialized benchmarks for advertising copy, technical documentation, code comments, and other professional creative work.

These will reveal different capability profiles across models and use cases.

Hybrid creative tools will mature

The next generation of creative software will assume AI-assisted workflows. Expect better interfaces for human-AI collaboration—tools that make it easier to iterate on AI outputs, track what's been tried, and maintain creative coherence across multiple generation rounds.

The Uncomfortable Truth About Creative Automation

The Montreal study confirms something the industry has suspected but couldn't prove: AI has reached human-average performance on standardized creativity measures. This is a genuine milestone. It has real implications for how creative work gets done.

But the study also confirms that we're nowhere close to automating elite creative work. The gap isn't small. The mechanisms that would close it—genuine metacognition, cultural grounding, intentional meaning-making—aren't present in current architectures and aren't on near-term roadmaps.

For business leaders, this means the question isn't "Will AI replace creative professionals?" It's "Which creative tasks benefit from AI assistance, and which require human judgment we can't automate?"

The Montreal researchers framed their work as settling "a years-long debate about whether AI can match human creative abilities." In reality, they've sharpened the debate: AI can match average human abilities on narrow measures, but the abilities that matter most for competitive advantage remain human.

That's not a feel-good conclusion or a dismissal of AI capabilities. It's a strategic reality that should inform investment, hiring, and workflow decisions.

The companies that will win are those that figure out the right allocation of creative work between humans and machines—not those that automate the most, and not those that automate the least. The Montreal data gives us a clearer empirical foundation for making that allocation thoughtfully.

The study proves AI can generate creative output at average human levels, but the business value of creativity lives almost entirely in the top percentiles that AI cannot reach—invest in tools that amplify your best people rather than systems that replace your average ones.