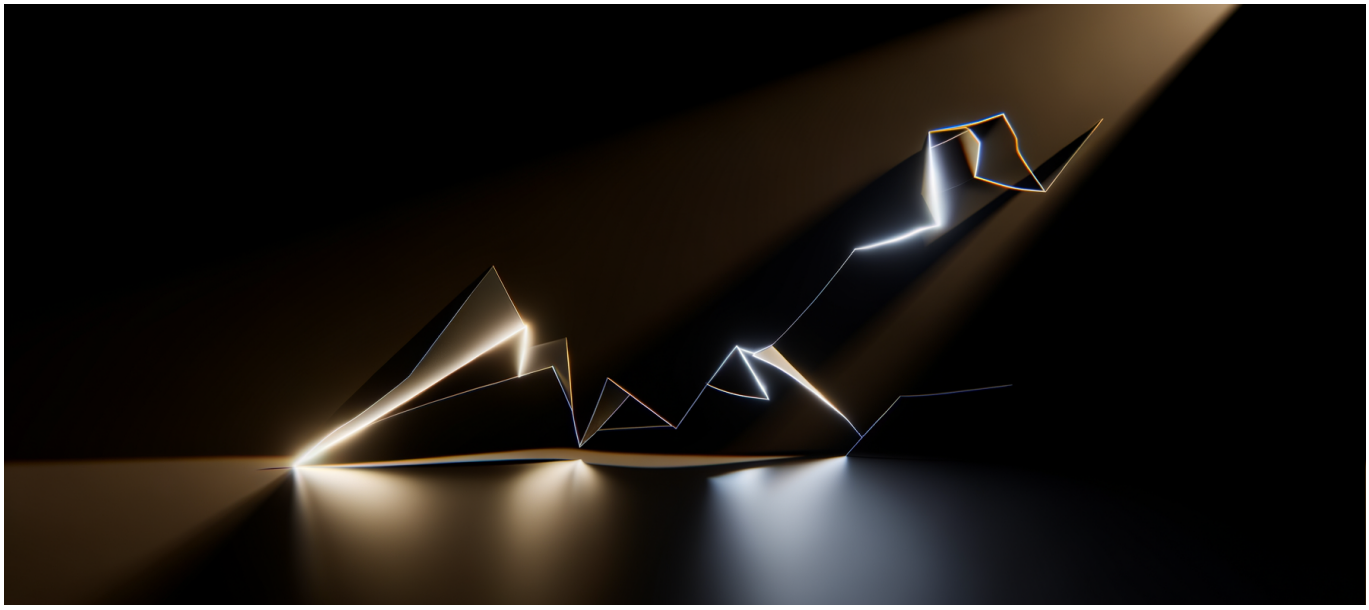

University of Montreal Study Proves AI Beats Average Humans on Creativity Tests—But Top 10% Still Outperform GPT-4

The world’s largest creativity study just revealed an uncomfortable truth: half of humanity is now less creative than a language model. But the top 10% of human minds still operate in territory AI cannot reach.

The News: 100,000 Humans vs. Four AI Giants

A study published January 21, 2026 in Scientific Reports by University of Montreal’s Professor Karim Jerbi delivers the first large-scale objective measurement of where AI creativity actually stands against human baseline. The research team—which includes deep learning pioneer Yoshua Bengio and collaborators from Mila (Quebec AI Institute), Google DeepMind, University of Toronto, and Concordia University—tested GPT-4, ChatGPT, Claude, and Gemini against over 100,000

human participants.

The primary benchmark was the Divergent Association Task (DAT), a standardized psychological test that requires subjects to generate ten words with maximum semantic distance from each other. Think “telescope,” “mushroom,” “legislation,” “whisper”—concepts so unrelated that linking them requires genuine cognitive flexibility. The test measures divergent thinking, the cognitive process most closely associated with creative potential.

The results split cleanly down the middle. AI models outperformed the median human participant on DAT scores. But the top 50% of human test-takers exceeded all four AI models. The top 10% operated in a performance band that no AI approached.

According to the University of Montreal’s announcement, the study represents the largest direct human-AI creativity comparison ever conducted using objective linguistic metrics like semantic distance—not subjective ratings from human judges, but measurable mathematical properties of the generated content.

Why This Matters: The Median Line Crossed

The significance isn’t that AI scored well on a creativity test. It’s that we now have empirical proof of exactly where the human-AI capability boundary sits—and that boundary runs through the middle of humanity.

For the past two years, the creativity debate has been theoretical. Can machines truly create? Is it just sophisticated pattern matching? The Montreal study sidesteps the philosophical question entirely. It doesn’t ask whether AI creativity is “real.” It asks whether AI output is measurably more divergent than human output. For half of humans tested, the answer is yes.

This creates a new competitive landscape. Roles that required average creative output—brainstorming sessions, initial concept generation, baseline content creation—now have a non-human alternative that performs comparably or better on objective metrics. The 100,000-person sample size removes any argument about statistical noise or cherry-picked results.

But the study reveals something more important than AI capability. It reveals human capability variance. The gap between median human performance and top-

decile human performance on the DAT exceeds the gap between median human performance and GPT-4 performance. Put differently: the difference between an average creative and an exceptional creative is larger than the difference between an average creative and a machine.

The strategic moat isn't "human creativity"—it's exceptional human creativity.

Winners in this landscape are organizations that can identify and deploy top-quartile creative talent while using AI to handle median-level creative work. Losers are those who assume "we have humans, so we're covered" without differentiating between human performance tiers.

Technical Depth: What the DAT Actually Measures and Why AI Struggles Beyond It

The Divergent Association Task works by calculating semantic distance using word embedding models. When you generate ten words, the algorithm computes the average pairwise distance between all word vectors in high-dimensional semantic space. Words that rarely appear in similar contexts across training corpora have greater semantic distance.

This is precisely what large language models are architecturally optimized to understand—semantic relationships between tokens. The surprise isn't that GPT-4 performs well on DAT. The surprise is that it doesn't dominate completely.

As reported by ScienceDaily, the research team discovered that temperature settings and etymology-based prompting strategies could boost AI creativity scores. Higher temperature increases sampling randomness, allowing models to select less probable tokens. Etymology prompting—asking the model to consider word origins and relationships—seemed to activate more diverse association pathways.

These findings reveal the mechanism behind AI creativity: it's fundamentally a function of sampling breadth across learned probability distributions. Increase randomness, and you get more divergent output. But there's a ceiling. The model can only sample from what it has learned, and its learned space, while vast, is bounded by training data.

Human top performers likely succeed through mechanisms unavailable to current architectures. They draw on embodied experience, emotional resonance, cross-domain expertise integration, and genuine novelty generation rather than recombination of existing patterns. The DAT captures one dimension of creativity—semantic divergence—and even there, the top humans outperform.

Where the gap widens dramatically is what the researchers call “richer narrative tasks.” The study tested participants and AI models on haiku composition, movie plot synopses, and flash fiction. According to Singularity Hub’s analysis, humans significantly outperformed AI on all three narrative benchmarks.

Why Narrative Beats AI: Constraint Satisfaction Meets Meaning Generation

Haikus require 5-7-5 syllable structure while conveying genuine insight about human experience. Movie plots require narrative tension that satisfies genre conventions while subverting expectations. Flash fiction demands character motivation that feels psychologically authentic.

Each of these tasks combines formal constraint satisfaction—which AI handles competently—with meaning generation that must resonate with human readers. The AI can produce technically correct haikus. It struggles to produce haikus that make a reader pause and feel something true.

This distinction matters for engineering decisions. If your application requires semantic divergence (brainstorming, ideation, exploration of solution space), current AI models perform at or above median human level. If your application requires meaningful narrative (user communication, brand voice, content that builds emotional connection), top-tier human talent still dramatically outperforms.

The Contrarian Take: What the Coverage Gets Wrong

Most reporting on this study will frame it as “AI vs. humans: who wins?” That framing misses the actual insight.

The underhyped finding: Human creativity is not normally distributed. The gap between the 50th percentile and 90th percentile human creative performance is

enormous—larger than most organizations account for in hiring, team composition, or creative process design. This study provides hard evidence that creative talent is not interchangeable, and the variance between humans far exceeds the variance between humans and machines.

The overhyped conclusion: “AI is now creative.” The DAT measures one specific cognitive capacity: semantic divergence in word association. It does not measure creative value, cultural impact, emotional resonance, or the ability to produce work that genuinely advances human understanding. A word list with high semantic distance is not a painting, a novel, or an invention. Conflating DAT performance with “creativity” writ large is a category error that the study’s authors are careful to avoid but that popular coverage will inevitably commit.

What’s missing from the debate: The study tests AI against humans on tasks designed for humans. The DAT was created to measure human divergent thinking using human-interpretable outputs. It says nothing about AI creativity in modalities where humans don’t participate—architectural design exploration, mathematical conjecture generation, or novel molecule synthesis. AI may already exceed all humans in creative domains we haven’t thought to test because the “creativity” happens in representational spaces humans don’t naturally inhabit.

The honest read of this study is narrower but more useful than the headlines suggest: on a specific, well-validated measure of one dimension of creative cognition, current AI models perform at the human median. That’s significant. It’s not everything.

Practical Implications: What Technical Leaders Should Actually Do

Audit Your Creative Workflows by Task Type

Map every process that requires creative input in your organization. Categorize each by whether the primary output is:

- **Semantic exploration** (brainstorming, keyword generation, solution space mapping): AI performs at median human level. Augmentation or replacement viable for efficiency gains.
- **Structured narrative** (documentation, technical writing, standard

communications): AI performs competently with human review. Quality ceiling exists but acceptable for many use cases.
- **Resonant narrative** (brand voice, user-facing content that builds relationship, creative work intended to move people emotionally): Human top performers still dramatically outperform. Reserve these tasks for your best talent and invest in identifying who that is.

Implement Performance Variance Measurement

Most organizations don’t measure creative output quality with any rigor. This study demonstrates that human creative performance varies by factors of 2-3x or more across the population. If you can’t identify your top-quartile performers, you can’t deploy them effectively.

Consider implementing DAT-style assessments in hiring pipelines for creative roles. The test is well-validated, takes minutes to administer, and provides an objective baseline. It won’t capture everything about creative potential, but it captures more than interviews and portfolio reviews that are subject to halo effects and presentation skills.

Adjust Temperature and Prompting for Ideation Workloads

The finding that temperature settings and etymology-based prompting boost AI creativity scores has immediate practical application. If you’re using AI for brainstorming or concept generation, default settings are leaving performance on the table.

Experiment with temperature ranges of 0.8-1.2 for ideation tasks (versus the typical 0.7 for factual accuracy). Structure prompts to ask for etymological relationships or conceptual origins as part of the ideation chain. Measure output divergence to calibrate for your specific use case.

Build Human-AI Creative Teams with Clear Role Separation

The study points toward an optimal configuration: AI for volume generation of candidates across solution space, humans for selection, refinement, and final creative judgment.

In practice, this means structuring creative workflows as:

- AI generates 10x-100x the candidate ideas a human would produce in the same time
- Human experts filter for quality, feasibility, and resonance
- Selected candidates get human development and refinement
- Final output represents human creative judgment informed by AI-expanded option space

This architecture plays to AI strength (breadth, speed, semantic exploration) and human strength (depth, meaning, selection judgment). Neither component alone matches the performance of the combined system.

Forward Look: Where This Leads in 6-12 Months

Expect Creativity Benchmarks to Proliferate

The Montreal study establishes a methodology. Other research groups will apply it to different creative domains: visual ideation (using image generation models), musical composition, code architecture, and scientific hypothesis generation. Each domain will reveal its own human-AI capability boundary. Some will show AI already ahead; others will show humans with wider leads than expected.

Within 12 months, we'll have a much more granular map of exactly which creative tasks are most amenable to AI augmentation and which remain firmly in human territory. Make architectural decisions accordingly.

Creative Assessment Technology Will Become a Category

If creative variance matters—and this study proves it does—then creative measurement becomes valuable. Expect to see startup activity around automated creative assessment tools for hiring, team composition, and individual development. The DAT is just one instrument; a full creative capability profile would include narrative generation, visual ideation, problem reframing, and domain-specific divergent thinking.

Organizations that adopt rigorous creative measurement early will have a hiring and deployment advantage. Those that continue treating creative talent as fungible will systematically misallocate their best people.

AI Model Training Will Explicitly Target Creativity Metrics

GPT-4 was not optimized for DAT performance. It was optimized for next-token prediction on diverse text. That it performs at human median on a creativity test is an emergent property, not a design goal.

Now that we have objective creativity metrics, model developers will train for them. Expect the next generation of foundation models to include creativity benchmarks alongside accuracy, reasoning, and factual recall. The gap between AI and top-decile humans will narrow, though the study suggests architectural barriers remain.

The “Creativity Premium” Will Increase

As AI handles median creative work, the economic value concentrates in the tails. Organizations will pay more for top-decile creative talent because the alternative—good enough AI output—will be nearly free at scale.

This follows the pattern established by other cognitively demanding fields after automation. Average performance becomes commoditized; exceptional performance becomes more valuable, not less. Compensation distributions for creative roles will likely spread, with median salaries flat or declining and top-performer compensation rising.

For individual practitioners, the implication is clear: reaching top-quartile performance in your creative domain is now a career necessity, not an optional aspiration. “Average” creative work has a new competitor that doesn’t sleep, doesn’t need benefits, and scales infinitely.

The Deeper Significance: What Top Humans Do Differently

The most intriguing question the Montreal study raises is not about AI. It’s about the top 10% of humans. What are they doing that current AI architectures cannot replicate?

The research doesn’t answer this definitively, but the narrative task results point toward possible explanations. Top creative performers likely integrate multiple cognitive systems that AI models currently separate or simulate poorly:

Embodied cognition: Human creativity draws on physical experience—the feeling of cold water, the weight of grief, the confusion of navigating a foreign city. AI models can describe these experiences but don't have them. The best human creatives translate embodied knowledge into resonant expression.

Emotional modeling: Creating work that moves people requires predicting emotional response in audiences. Humans with deep social and emotional experience model readers intuitively. AI models approximate this through pattern matching on emotionally successful texts, but the mechanism differs.

Deliberate constraint violation: Great creativity often involves knowing the rules well enough to break them productively. This requires meta-awareness of expectations and intentional subversion. Current AI models are excellent at following patterns and reasonable at varying them, but weak at strategic pattern breaking with specific effect in mind.

Cross-domain integration: The most creative insights often come from connecting ideas across distant fields—applying biological principles to software architecture, or economic theory to relationship dynamics. This requires having genuine expertise in multiple domains and perceiving structural similarities that don't appear in surface statistics. It's unclear whether scaling current architectures will produce this capability or whether it requires fundamentally different approaches.

Understanding what top humans do differently isn't just philosophical curiosity. It's a roadmap for AI development. If we can characterize the mechanisms behind top-decile human creative performance, we can attempt to build systems that exhibit similar properties. The Montreal study gives us a benchmark; the next challenge is reverse-engineering the process that beats it.

Conclusion: The New Creative Landscape

The Montreal study marks a transition point. We move from asking "can AI be creative?" to asking "which creative tasks still require top human talent?"

The answer is clear: tasks requiring semantic divergence alone now have competitive AI solutions. Tasks requiring meaningful narrative, emotional resonance, and cultural relevance still need exceptional human performers—and the emphasis is on exceptional. Median creative talent now competes with AI that

doesn’t tire, negotiate, or need healthcare.

For technical leaders, the study provides actionable intelligence. Know which tasks fall into which category. Measure creative performance in your organization. Design human-AI workflows that leverage complementary strengths. And prepare for compensation structures that increasingly reward the right tail of the human creative distribution.

The 100,000 humans who participated in this study unknowingly ran an experiment that answers a question every CTO needs to answer: where do machines stop and exceptional humans begin? That boundary now has coordinates.

The organizations that thrive will be those that can work both sides of the line—deploying AI for scale and speed while concentrating rare human creative talent where it still outperforms anything artificial.