Position: Why AI R&D Benchmarks Need to Measure Research Judgment Separately

Yilin Huang *
Amherst College
yilhuang28@amherst.edu

Abstract

Preparedness frameworks from frontier AI labs highlight automation of AI research and development (R&D) as a critical capability to monitor. However, current benchmarks fail to distinguish between success from brute-force search or the nuanced research judgment we termed as tactical research taste. This measurement gap is critical for expensive and longer experiments where tactical research taste provides the greatest economic advantage over brute-force approaches. Without measuring tactical research taste separately, we cannot accurately assess where AI stands as an efficient compute multiplier for research, create governance measures for AI R&D automation risks, or know when to hold labs accountable within preparedness frameworks. In this paper, we define tactical research taste, a skillset currently not explicitly measured in AI R&D automation evaluations. Then, we argue for the importance of measuring tactical capabilities separately from overall R&D automation capability, and outline multiple complementary evaluation approaches to address this critical gap in AI capability assessment.

1 Introduction

1.1 Risks from AI R&D automation

Recent developments in frontier AI capabilities have prompted leading developers and governments to establish frameworks for identifying catastrophic risks from advanced AI systems (1; 2; 3; 4). A distinct challenge emerges in assessing AI systems' capacity for autonomous research and development (3; 9). AI agents automating AI R&D could drastically increase available skilled research labor, potentially accelerating progress and leading to runaway feedback loops of capability improvement (5; 6). Such scenarios raise concerns about capability progress outpacing safety measures. To manage these risks, AI developers and public bodies have identified the need for *early-warning evaluations* that can assess AI R&D capabilities before systems become fully autonomous (1; 3; 4).

Several recent benchmarks attempt to evaluate AI research-automation capabilities across different domains. For general scientific research, *ScienceAgentBench* tests the complete scientific discovery workflow, while *DiscoveryWorld* creates simulated laboratory environments for hypothesis generation and experimental analysis (7; 8). Focusing specifically on AI and machine learning research, *RE-Bench* evaluates agents on open-ended ML research problems over 8-hour time horizons with direct human comparisons, while *MLE-bench* and *ML-Agent-Bench* target machine-learning engineering and shorter research tasks respectively (9; 10; 11).

*Work conducted at UChicago Existential Risk Laboratory

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science Workshop (NeurIPS 2025).

1.2 Measurement Gaps in Existing Benchmarks

As one of the primary benchmarks evaluating AI research automation, RE-Bench reveals a critical measurement gap: models excel through brute-force search but struggle when research judgment is required.

Models achieved their highest scores on kernel optimization tasks through systematic parameter tweaking in environments with fast evaluation cycles. Models submit solutions over 10 times faster than humans but most attempts score close to zero, failing to improve on reference solutions (9). This brute-force approach works because the environment is small (fewest lines of code), making exhaustive search computationally feasible despite the low success rate.

In contrast, models performed significantly worse on tasks requiring "detailed planning": experiment design, idea adaptation, and research prioritization. These tasks, like optimizing LLM foundry configurations, resist brute-force approaches because the solution space is too large and evaluation time is too long for systematic search.

Importantly, even when models succeeded on detailed planning tasks, they provided no clear rationale for their parameter choices in various spots. Therefore, **RE-Bench's success cases become uninformative**. When a model gets a high score through unexplained parameter choices, we can't tell if it succeeded through genuinely good research judgment or mere lucky parameter guessing.

This performance pattern exposes what current benchmarks can't measure: an agent's success correlates directly with whether tasks favor search optimization over research judgment. To address this gap, this paper creates a conceptual framework of research taste, makes the case for why tactical research taste is an important variable to look out for in AI RD automation, and suggests potential evaluation methods and future directions.

2 What is Research Taste?

The measurement gaps identified in RE-Bench stem from conflating different research capabilities under a single evaluation framework. To address this, we need a clearer conceptual foundation for what constitutes research judgment versus other abilities (i.e. experimental implementation through code). Research taste is embedded in the entire research process and separated into two levels: strategic taste is picking which mountain to climb on, and tactical taste is choosing how to climb the mountain.

2.1 Tactical vs. Strategic

Strategic research taste involves the high-level decisions about what to work on. This means picking important research areas, generating potential research questions within those areas, and prioritizing which hypotheses are worth pursuing. This requires long-term thinking about what problems matter most and where resources should be allocated. A researcher with good strategic taste might recognize that improving multi-step reasoning is more cost-effective than improving few-shot learning right now, then coming up with a relevant research question and hypothesis.

Tactical research taste starts after you've decided what hypothesis to test and involves three key judgments:

1. **Operationalizing variables** - what exactly to measure and control
2. **Experimental design** - how to test the hypothesis cleanly
3. **Result interpretation** - whether unexpected outcomes reflect hypothesis failure or experimental flaws

A researcher with good tactical taste knows how to design clean experiments that isolate the variables they care about in order to narrow down their hypothesis search space, and can recognize when results suggest they need to pivot their approach (13).

This distinction is necessary as they're often weakly correlated in humans: junior researchers tend to focus on improving tactical research taste while senior researchers tend to be good at strategic research taste (14). These two skills appear to require different cognitive skills: strategic taste

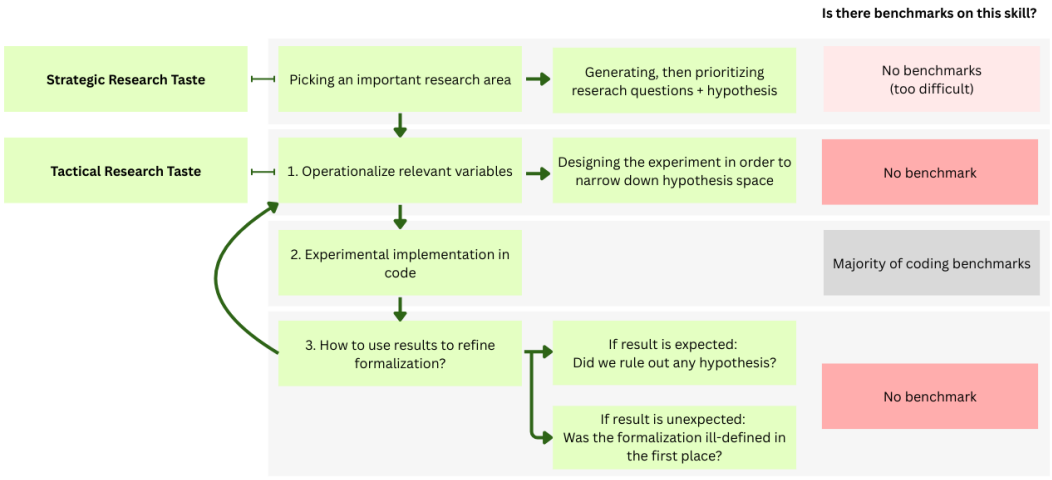

Figure 1: Strategic and tactical research taste components with current benchmark coverage.

needs broad domain knowledge and intuition about research impact from seeing and reading many experiments, while tactical taste requires experimental rigor and the ability to navigate ambiguity in constrained settings.

As Figure 1 shows, current benchmarks extensively measure experimental implementation (step 2) while failing to evaluate variable operationalization (step 1) and result interpretation (step 3). This creates a fundamental measurement gap: benchmarks like SWE-bench excel at testing implementation skills in well-defined environments with clear success metrics, but cannot assess whether models understand which variables matter for a given hypothesis or how to interpret ambiguous experimental outcomes (12).

2.2 Why Focus on Tactical Taste Before Strategic Taste

Strategic taste operates on longer feedback loops and broader scope. A researcher might not know if their strategic choices were good for months or years. Tactical taste has shorter feedback loops: you can often tell within days or weeks whether your experimental design was sound. Importantly, strategic research taste depends on tactical execution to generate meaningful results. Even brilliant strategic choices about research direction will fail if the experiments are poorly designed or results misinterpreted. Since models currently struggle with tactical research taste, any failure in the full R&D loop could stem from tactical bottlenecks rather than strategic limitations. This is why we focus on evaluating tactical research taste first. Until models can reliably execute experiments and interpret results, we cannot meaningfully assess their strategic research judgment.

3 Why Does Tactical Research Taste Matter for AI R&D Automation?

High-stakes experiments can't rely on lucky guesses. Tactical research taste becomes critical when experiments are expensive or take months to evaluate. Frontier AI companies training large models can't afford to run dozens of poorly-designed experiments hoping one succeeds by chance. In certain tasks in RE-Bench, over 90% of model runs result in scoring 0. Additionally, some algorithmic improvements only show clear effects at scale, which would require running larger and more expensive experiments to verify results (15). In these high-cost environments, the difference between systematic experimental design and brute-force search becomes economically decisive.

Separate benchmarking reveals what's driving AI R&D progress. If overall R&D capability scores improve while tactical research taste scores remain flat, we know models are succeeding through increased computational search rather than better experimental judgment. This suggests compute remains the primary bottleneck for R&D automation.

Conversely, if both R&D capability and tactical research taste scores improve together, models are becoming more efficient at extracting insights from experiments—choosing better variables to test, designing cleaner experimental setups, and correctly interpreting ambiguous results. This efficiency gain could generalize beyond R&D to other domains requiring systematic investigation.

Implications for governance and takeoff speed. Without tactical research taste, AI R&D automation requires massive compute scaling to brute-force solutions. With strong tactical research taste, models can achieve the same research progress with far less experimental compute. This distinction matters for both regulatory approaches and takeoff scenarios. Research taste acts as a compute multiplier that could dramatically accelerate AI development timelines, which calls for governance regulations.

4 How to Evaluate Research Taste

We propose some preliminary approaches to address this measurement gap, each targeting different aspects of research taste while acknowledging potential limitations.

4.1 Overall Tactical Approach: Complex Environments with Limited Attempts

Create environments that reward research taste over brute-force search by making exhaustive exploration prohibitively expensive. This involves:

- Complex, multi-component systems with many potential variables to manipulate
- Long evaluation cycles that prevent rapid iteration
- Success requiring principled experimental design rather than parameter sweeping
- Strict limits on experiment attempts (e.g., 5-10 trials maximum)

Even if models eventually saturate such benchmarks without demonstrating clear research taste, achieving strong performance under these constraints would still represent meaningful progress toward R&D automation.

4.2 Result Analysis Approach: Testing Interpretation Skills

Rather than requiring models to design and execute experiments, test their ability to interpret completed experimental results:

- Present experimental setups and ask models to identify design flaws or confounding variables
- Give ambiguous results and test whether models correctly distinguish between hypothesis failure and experimental error
- Provide multiple competing explanations for results and evaluate reasoning quality
- Test next-step experimental design based on previous results

This approach isolates the interpretation component of tactical research taste from implementation ability.

4.3 Strategic Approach: Knowledge-Cutoff Research Direction Assessment

Test strategic research taste by having models with knowledge cutoffs suggest promising research directions:

- Present the state of a field as of the cutoff date
- Ask models to predict which approaches will be most fruitful
- Evaluate predictions against subsequent real-world research outcomes
- Test ability to identify neglected but important research areas

4.4 Limitations and Call for Multiple Approaches

Creating effective tactical research taste benchmarks faces several fundamental challenges. First, it's difficult to isolate research judgment from general problem-solving ability—models might succeed through domain knowledge or reasoning skills rather than genuine experimental intuition. Second, any specific benchmark risks becoming a proxy task that doesn't generalize to real research contexts. Third, there's tension between creating tasks granular enough to measure specific skills versus comprehensive enough to capture actual research judgment.

These are preliminary directions rather than complete solutions. Each approach targets different aspects of tactical research taste, and the field likely needs multiple complementary measures rather than seeking a single definitive benchmark. However, over-fragmenting evaluation into too many micro-tasks risks losing sight of the integrated judgment that defines good research taste.

5 Future Directions

We propose this as an agenda for researchers working on AI capability assessment. Better measurement of research taste distinct from general coding ability is needed before we can accurately assess how close AI systems are to automating high-stakes AI R&D. The current measurement gap leaves us uncertain whether models are developing genuine research judgment or simply becoming better at brute-force search.

Several critical questions remain for future work:

How much can research taste accelerate R&D progress? The speed-up from better experimental design depends heavily on field maturity—domains with abundant low-hanging fruit may benefit less from tactical research taste than mature fields where progress requires careful hypothesis testing. Additionally, even small improvements in experimental efficiency could compound over time, potentially unlocking new research directions.

What determines the complexity of research environments? Understanding how AI systems might reduce search spaces through better variable selection and experimental design is crucial for predicting automation timelines. This connects directly to whether tactical research taste can make intractable research problems more manageable.

Does tactical research taste enable full versus partial R&D automation? While AI systems might achieve significant research acceleration without strong research taste through computational brute force, full automation may require genuine experimental judgment. The difference matters enormously—partial automation provides linear speed-ups while full automation could enable exponential research progress.

Will inference scaling improve experimental design capabilities? As models become more capable at reasoning during inference, it remains unclear whether this translates to better research judgment or simply more sophisticated search strategies.

How does tactical research taste relate to strategic research taste? Understanding whether experimental design skills transfer to higher-level research direction choices will determine whether measuring tactical capabilities provides insight into broader R&D automation risks.

Does tactical research taste create differential automation timelines for safety versus capabilities research? Safety research often requires careful experimental design to detect subtle effects or rule out confounding factors, while capabilities research may benefit more from rapid iteration and scaling. If tactical research taste becomes a bottleneck specifically for safety work, AI systems might accelerate capabilities development while leaving safety research dependent on human oversight, potentially widening the gap between AI progress and safety measures.

6 Conclusion

Current AI R&D benchmarks conflate tactical research taste with general coding ability, creating dangerous blind spots in capability assessment. Models excel through brute-force parameter exploration but struggle with having reasoned judgements in experimental design and result interpretation.

This measurement gap obscures whether AI progress stems from genuine experimental insight or computational search, making it difficult to accurately forecast research automation trajectories.

The economic and safety implications are substantial. In high-stakes research environments where experiments are expensive and evaluation cycles are long, tactical research taste provides the greatest competitive advantage over brute-force approaches. Models with strong research judgment can achieve equivalent insights with dramatically fewer experimental resources, acting as an efficient compute multipliers. Without this capability, AI systems may succeed in low-cost environments while failing in contexts that matter more for frontier research.

Preparedness frameworks cannot effectively govern AI R&D automation without disaggregated measurement of these capabilities. The evaluation approaches outlined here—constrained experimental environments, result interpretation tasks, and strategic direction assessment—represent initial steps toward more precise capability evaluation. Until we can distinguish systematic experimental reasoning from parameter exploration, our understanding of AI research automation risks remains fundamentally incomplete. We propose this as a research agenda to be further improved for the AI safety and evaluation communities.

References

References

[1] OpenAI. Our updated Preparedness Framework (v2). 2025. https://openai.com/index/updating-our-preparedness-framework/.

[2] Anthropic. Responsible Scaling Policy (RSP), v1 (and updates). 2023. https://www.anthropic.com/responsible-scaling-policy.

[3] UK AI Safety Institute. AI Safety Institute: approach to evaluations. 2024. https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations.

[4] The White House. Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. 2023. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[5] Leopold Aschenbrenner. *Situational Awareness: The Decade Ahead.* 2024. https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf.

[6] Tom Davidson and Daniel Eth. Will AI R&D Automation Cause a Software Intelligence Explosion? Forethought Research, 2025. https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion.

[7] Ziru Chen, Shijie Chen, Yuting Ning, *et al.* ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. arXiv:2410.05080, 2024. https://arxiv.org/abs/2410.05080.

[8] Peter Jansen, Marc-Alexandre Côté, Tushar Khot, *et al.* DISCOVERYWORLD: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents. arXiv:2406.06769, 2024. https://arxiv.org/abs/2406.06769.

[9] Hjalmar Wijk, Jeffrey Ladish, Paul Christiano, *et al.* RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. arXiv:2411.15114, 2024. https://arxiv.org/abs/2411.15114.

[10] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, *et al.* MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. arXiv:2410.07095, 2024. https://arxiv.org/abs/2410.07095.

[11] Qian Huang, Jian Vora, Percy Liang, Jure Leskovec. MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation. arXiv:2310.03302, (ICML 2024 version). https://arxiv.org/abs/2310.03302.

[12] Florian Brand and Jean-Stanislas Denain. What skills does SWE-bench Verified evaluate? Epoch AI, June 13, 2025. https://epoch.ai/blog/what-skills-does-swe-bench-verified-evaluate.

[13] Neel Nanda. My Research Process: Understanding and Cultivating Research Taste. AI Alignment Forum, May 1, 2025. https://www.alignmentforum.org/posts/Ldrss6o3tiKT6NdMm/my-research-process-understanding-and-cultivating-research.

[14] David Owen. Interviewing AI researchers on automation of AI R&D. Epoch AI, Aug 27, 2024. https://epoch.ai/blog/interviewing-ai-researchers-on-automation-of-ai-rnd.

[15] Henry Josephson. How fast can algorithms advance capabilities? Epoch AI, 2025. https://epoch.ai/gradient-updates/how-fast-can-algorithms-advance-capabilities.