Retrospective Benchmarks for Machine Intelligence

Evaluating Current AI Against Historical Specifications

Chapter 1: The Gubrud Benchmark (1997)

Dakota Schuck

December 2025

Working paper. Comments welcome.

Preface: Methodology

This document attempts something unusual: treating historical predictions and definitions of machine intelligence as testable specifications, then evaluating current AI systems against them.

The approach is necessarily imperfect. We are:

- Applying 21st-century benchmarks to 20th-century (and earlier) concepts
- Asking whether systems meet specifications that weren't written as specifications
- Inviting the original thinkers to a conversation they cannot fully join

We've tried to be rigorous where rigor is possible, explicit about uncertainty where it isn't, and honest about the gaps. Every claim should be cited; where citations are missing, we've marked them. Where we've made interpretive choices, we've flagged them.

This is a first attempt.[1] It is meant to be improved, corrected, and extended by others. If you can strengthen a citation, challenge an interpretation, or propose a better threshold—please do.

[1] AI Assistance Disclosure: Research, drafting, and analysis were conducted with the assistance of Claude Opus 4.5 (Anthropic, 2025). The AI contributed literature review, benchmark operationalization, and self-assessment of AI capabilities. The author provided editorial direction, methodological framing, and final approval. Responsibility for all claims rests with the author.

1 Introduction: The Term Worth Trillions

In the summer of 1997, a physics graduate student sat in a basement pump room at the University of Maryland, reading everything he could find about emerging technologies.[2] Mark Gubrud was worried about autonomous weapons. That year, he submitted a paper to the Fifth Foresight Conference on Molecular Nanotechnology with a warning about how advanced AI could destabilize international security.[3]

In that paper, he used a phrase no one had used before: *artificial general intelligence*.

No one noticed. The term disappeared for nearly a decade.

Around 2002, a group of AI researchers—including Shane Legg (later co-founder of DeepMind) and Ben Goertzel—were searching for a name for the kind of AI they wanted to build. They independently coined the same term.[4] In 2005, Gubrud surfaced in an online forum to point out his priority. Legg's response, years later: "Someone comes out of nowhere and says, 'I invented the AGI definition in '97,' and we say, 'Who the hell are you?' Then we checked, and indeed there was a paper."[5]

Today, "AGI" anchors contracts worth billions of dollars.[6] The term Gubrud coined in a basement—while warning about the dangers of advanced AI—now names the explicit goal of the world's most valuable AI companies.

Gubrud, now 67, lives in Colorado, caring for his mother.[7] He has no steady job.[8]

The question we're asking: If you could show Gubrud a current frontier AI system—say, Claude Opus 4.5—would he say, yes, this is what I meant?

And not just Gubrud. What about Turing? Lovelace? McCarthy? Minsky? Each left us something like a specification. Did we meet it?

[2]"He spent all day buried in the noisy pump room on the basement floor of the laboratory, sitting there reading everything he could find." 36kr.com, "He Invented Trillion-Worth AGI but Now Is Down and Out," 2025. https://36kr.com/p/2689463822082945

[3]Gubrud, Mark A. "Nanotechnology and International Security." Fifth Foresight Conference on Molecular Nanotechnology, November 1997. https://legacy.foresight.org/Conferences/MNT05/Papers/Gubrud/index.html

[4]Legg, Shane. Quoted in various interviews; see also Goertzel, Ben, ed. *Artificial General Intelligence*. Springer, 2007.

[5]36kr.com, op. cit.

[6]OpenAI's partnership with Microsoft reportedly values AGI-related IP in the hundreds of billions. Specific contract terms are not public.

[7]36kr.com, op. cit. Article dated 2025 states Gubrud is 67.

[8]Ibid.

2 The Original Definition

From "Nanotechnology and International Security," presented at the Fifth Foresight Conference on Molecular Nanotechnology, November 1997:[9]

> *. . . artificial general intelligence. . . AI systems that rival or surpass the human brain in complexity and speed, that can acquire, manipulate and reason with general knowledge, and that are usable in essentially any phase of industrial or military operations where a human intelligence would otherwise be needed.*

2.1 Context

Gubrud wasn't writing an AI paper. He was writing a security paper. "AGI" appeared alongside nanotechnology and other emerging technologies as potential destabilizers of international order.[10] His concern was weaponization and arms races, not capability benchmarks.

This matters for interpretation: Gubrud's "general intelligence" was meant to contrast with narrow, task-specific systems. His reference to "industrial or military operations" wasn't arbitrary—it reflected his focus on domains where autonomous systems could substitute for human judgment in consequential decisions.

2.2 Operationalization

We extract six criteria from Gubrud's definition, in his order:

1. Rival or surpass human brain in complexity
2. Rival or surpass human brain in speed
3. Acquire general knowledge
4. Manipulate general knowledge
5. Reason with general knowledge
6. Usable where human intelligence would otherwise be needed

For each criterion, we identify subcriteria, existing measures where available, thresholds, and an assessment.

Scoring:

☐ 0% — Clearly does not meet criterion
☐ 50% — Contested; reasonable arguments exist on both sides
☐ 100% — Clearly meets criterion

[9] Gubrud 1997, op. cit. Full quote also cited in: Morris, Meredith Ringel, et al. "Levels of AGI for Operationalizing Progress on the Path to AGI." arXiv:2311.02462, 2023. https://arxiv.org/abs/2311.02462; METR, "AGI: Definitions and Potential Impacts," 2024.

[10] Gubrud's paper focused primarily on nanotechnology and international security; AGI appears as one of several destabilizing emerging technologies.

3 Criterion 1: Rival or Surpass Human Brain in Complexity

3.1 What Gubrud Probably Meant

In 1997, "complexity" in the context of brains likely referred to the scale and interconnection of neural structures. The comparison to the human brain suggests Gubrud imagined systems approaching biological scale—not necessarily identical architecture, but comparable information-processing capacity.

3.2 Structural Scale

Measure: Model parameter count vs. estimates of human brain synaptic connections

Reference values:

- Human brain: ∼86 billion neurons, ∼100–600 trillion synapses[11]
- Human language-specific regions (Broca's and Wernicke's areas): ∼400–700 billion effective parameters by one estimate[12]
- Frontier LLMs (2025): ∼1–2 trillion parameters[13]

Threshold: ≥100 trillion parameters (full-brain parity) OR ≥500 billion (language-region parity)

Assessment: Current models are within an order of magnitude of language-specific brain regions but remain 100–600× below full-brain synapse counts.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

Caveats: Parameter-synapse comparisons are architecturally problematic.[14] Synapses have dynamic, continuous-valued states; parameters are fixed post-training. A bee brain has ∼1 million neurons and performs complex navigation.[15] Scale may not be the right measure of complexity.

3.3 Functional Complexity (Task Diversity)

Measure: Number of distinct cognitive task categories performed at human-competent level

Reference values:

- MMLU benchmark: 57 subject areas[16]
- BIG-Bench: 204 tasks[17]

---

[11] Azevedo, Frederico A.C., et al. "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain." *Journal of Comparative Neurology* 513.5 (2009): 532–541. https://doi.org/10.1002/cne.21974

[12] Millidge, Beren. "The Scale of the Brain vs Machine Learning." beren.io, 2022. https://www.beren.io/2022-01-30-The-Scale-of-the-Brain-vs-Machine-Learning/

[13] Model parameter counts for frontier systems are not always publicly disclosed. GPT-4 was reported at ∼1.8T parameters (unconfirmed); Claude and Gemini parameter counts are not public. Specific parameter counts for Claude Opus 4.5, GPT-5, and Gemini 3 Pro would strengthen this estimate.

[14] See discussion in Millidge 2022, op. cit., and Crawford, Hal. "AI versus the human brain." halcrawford.substack.com, 2024. https://halcrawford.substack.com/p/ai-versus-the-human-brain

[15] Menzel, Randolf, and Martin Giurfa. "Cognitive architecture of a mini-brain: the honeybee." *Trends in Cognitive Sciences* 5.2 (2001): 62–71. https://doi.org/10.1016/S1364-6613(00)01601-6

[16] Hendrycks, Dan, et al. "Measuring Massive Multitask Language Understanding." arXiv:2009.03300, 2020. https://arxiv.org/abs/2009.03300

[17] Srivastava, Aarohi, et al. "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models." arXiv:2206.04615, 2022. https://arxiv.org/abs/2206.04615

- Human competence: Thousands of task types[18]

Threshold: Competent performance ($\geq$50th percentile among humans) across $\geq$100 cognitively distinct task categories

Current performance:

- Frontier LLMs score $\geq$85% on MMLU, covering 57 subjects[19]
- Performance across BIG-Bench tasks is variable but broadly competent[20]

Assessment: Frontier models demonstrate breadth across dozens to hundreds of task categories. Whether this constitutes complexity rivaling the human brain depends on interpretation.

Score:
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

3.4 Architectural Sophistication

Measure: Presence of dynamic, adaptive features beyond static feedforward networks

Reference values for human brain: Persistent memory, real-time learning, attention modulation, self-monitoring, multi-modal integration[21]

Feature assessment:

- Persistent memory across sessions — Limited; depends on deployment[22]
- In-context learning — Present[23]
- Tool use — Present[24]
- Multi-modal integration — Present[25]
- True self-modification/online learning — Absent during inference[26]

Threshold: $\geq$4 of 5 features with human-like flexibility

Assessment: 3 of 5 features present; persistent memory and self-modification remain limited.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

[18] Estimate based on breadth of human cognitive abilities. A systematic taxonomy of human cognitive task categories would provide a more rigorous comparison.

[19] Various benchmark reports; see Artificial Analysis, "Claude Opus 4.5 Benchmarks," November 2025. https://artificialanalysis.ai/

[20] A systematic count of tasks at $\geq$50th percentile human performance would strengthen this assessment.

[21] Standard neuroscience; see e.g., Kandel, Eric R., et al. *Principles of Neural Science.* 5th ed., McGraw-Hill, 2013. https://neurology.mhmedical.com/book.aspx?bookID=1049

[22] Memory features vary by deployment. Claude.ai offers memory features; API deployments typically do not persist state.

[23] Brown, Tom, et al. "Language Models are Few-Shot Learners." NeurIPS 2020. https://arxiv.org/abs/2005.14165

[24] Schick, Timo, et al. "Toolformer: Language Models Can Teach Themselves to Use Tools." arXiv:2302.04761, 2023. https://arxiv.org/abs/2302.04761

[25] Multimodal models including GPT-4V, Gemini, Claude 3+ support image, audio, and in some cases video input.

[26] Current LLMs do not update weights during inference. Fine-tuning requires separate training runs.

4 Criterion 2: Rival or Surpass Human Brain in Speed

4.1 What Gubrud Probably Meant

Processing speed—how quickly the system can take in information and produce outputs. In 1997, human cognition was clearly faster than existing AI for most tasks.

4.2 Text Generation Speed

Measure: Output tokens per second vs. human speaking/writing speed
Reference values:

- Human speaking: ∼125–150 words/minute ≈ 2–3 words/second ≈ 3–4 tokens/second[27]
- Human typing (average): ∼40 words/minute ≈ ∼1 token/second[28]
- Frontier LLMs: 50–200 tokens/second typical; up to 1,800 tokens/second on specialized hardware[29]

Threshold: ≥10× human speaking speed (≥30 tokens/second)
Current performance: Frontier models exceed 50 tokens/second routinely.[30]

Score:
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

4.3 Text Processing Speed

Measure: Input tokens processed per second vs. human reading speed
Reference values:

- Human reading: ∼200–300 words/minute ≈ 4–5 tokens/second[31]
- LLM prompt processing: Thousands of tokens/second[32]

Threshold: ≥100× human reading speed (≥500 tokens/second)
Current performance: Exceeds threshold by large margin.

Score:
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

[27] Typically cited speaking rate. See: Yuan, Jiahong, et al. "Towards an integrated understanding of speaking rate in conversation." INTERSPEECH 2006. https://www.isca-archive.org/interspeech_2006/yuan06_interspeech.html
[28] Typing speed varies widely. 40 WPM is often cited as average. See various typing studies.
[29] Cerebras. "Introducing Cerebras Inference: AI at Instant Speed." cerebras.ai, 2024. https://cerebras.ai/blog/introducing-cerebras-inference-ai-at-instant-speed
[30] Artificial Analysis, op. cit., and various model benchmarks.
[31] Reading speed varies. 200–300 WPM is commonly cited for adult reading. See: Rayner, Keith, et al. "Eye movements and information processing during reading." *Psychological Bulletin* 124.3 (1998): 372–422. https://doi.org/10.1037/0033-2909.124.3.372
[32] Prompt processing speed varies by model and hardware; generally measured in thousands of tokens/second.

4.4 Response Latency

Measure: Time to first token (TTFT)
Reference values:

- Human conversational response: ~200–500ms for simple replies[33]
- Frontier LLMs: ~100–500ms typical TTFT[34]

Threshold: ≤500ms for standard queries

Score:
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

4.5 Reasoning Speed

Measure: Time to solve complex problems vs. human experts at equivalent accuracy
Reference values:

- Human expert on GPQA-level problem: Minutes to tens of minutes[35]
- LLMs with extended thinking: Seconds to minutes[36]

Assessment: For problems current AI can solve, speed is comparable or faster. For problems requiring extended deliberation, timing varies.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

[33] Human response latency in conversation is typically 200–500ms for turn-taking. See: Stivers, Tanya, et al. "Universals and cultural variation in turn-taking in conversation." *PNAS* 106.26 (2009): 10587–10592. https://doi.org/10.1073/pnas.0903616106

[34] Various model benchmarks report TTFT in the 100–500ms range for standard queries.

[35] Estimate based on problem complexity. Timed human expert performance data on GPQA would provide a more rigorous baseline.

[36] Extended thinking / reasoning models (o1, Claude thinking mode) can take seconds to minutes depending on problem complexity.

5 Criterion 3: Acquire General Knowledge

5.1 What Gubrud Probably Meant

The ability to gain knowledge—to learn. In 1997, machine learning existed but was narrow. "Acquire general knowledge" implies learning across domains, not just pattern-matching on fixed training data.

5.2 Few-Shot Learning Efficiency

Measure: Performance improvement per example on novel tasks

Benchmark: ARC-AGI (Abstraction and Reasoning Corpus), explicitly designed to test skill acquisition efficiency[37]

Reference values:

- Humans: ∼73–77% on ARC-AGI-1 public tasks[38]
- Best AI (late 2024): ∼55% on ARC-AGI-1 private set[39]
- OpenAI o3 (Dec 2024): ∼87.5% on ARC-AGI-1 (high compute)[40]
- AI on ARC-AGI-2 (2025): Single-digit percentages[41]

Threshold: ≥85% on ARC-AGI-1 (the competition target)

Assessment: o3 crossed the 85% threshold on ARC-AGI-1, but ARC-AGI-2 remains largely unsolved. The ability to acquire genuinely novel skills efficiently remains contested.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

5.3 Knowledge Breadth

Measure: Factual knowledge across domains

Benchmark: MMLU (Massive Multitask Language Understanding)—57 subjects from elementary to professional level[42]

Reference values:

- Human expert ceiling: ∼90% (estimated)[43]
- Claude Opus 4.5: ∼88–90%[44]
- Frontier models generally: 88–91%[45]

[37] Chollet, François. "On the Measure of Intelligence." arXiv:1911.01547, 2019. https://arxiv.org/abs/1911.01547

[38] Johnson, Aaditya, et al. "Testing ARC on Humans: A Large-Scale Assessment." NYU, 2024. Reported 73.3–77.2% average accuracy. https://lab42.global/arc-agi-benchmark-human-study/

[39] ARC Prize 2024 Technical Report. arcprize.org, December 2024. https://arcprize.org/

[40] OpenAI. "Introducing o3." December 2024. https://openai.com/index/deliberative-alignment/; François Chollet, social media announcements, December 2024.

[41] ARC-AGI-2 was released in early 2025; as of late 2025, top scores remain in single-digit percentages. See https://arcprize.org/

[42] Hendrycks et al. 2020, op. cit.

[43] Estimated ceiling based on question validity studies. Gema, et al. "We Need to Talk about MMLU: The Importance of Studying Benchmark Errors." arXiv:2406.04127, 2024. https://arxiv.org/abs/2406.04127 A rigorous human baseline study on full MMLU would strengthen this estimate.

[44] Artificial Analysis, "Claude Opus 4.5 Benchmarks," November 2025.

[45] Various benchmark reports, December 2025.

Threshold: ≥85% MMLU accuracy

Score:
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

Caveat: MMLU is now considered near-saturated and may not distinguish frontier models.[46]

5.4 Real-Time Knowledge Acquisition (Tool Use)

Measure: Ability to retrieve and integrate new information during task execution
Capabilities present: Web search, document retrieval, API access[47]
Assessment: Tool use exists but integration is imperfect; hallucination and retrieval failures occur.[48]

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

[46] Gema et al. 2024, op. cit.; discussion in AI research community about MMLU saturation.
[47] Tool use is standard in frontier deployments. See Anthropic documentation, OpenAI function calling, etc.
[48] Hallucination in RAG systems is documented but rates vary. A systematic meta-analysis would strengthen this assessment.

6 Criterion 4: Manipulate General Knowledge

6.1 What Gubrud Probably Meant

Not just storing knowledge but working with it—transforming, combining, applying it flexibly across contexts.

6.2 Cross-Domain Transfer

Measure: Application of knowledge from one domain to problems in another

Existing benchmarks: Limited standardization[49]

Assessment: LLMs demonstrate some analogical transfer[50] but also exhibit surprising failures when surface features change.[51]

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

6.3 Knowledge Synthesis

Measure: Combining multiple sources into coherent novel outputs

Assessment: LLMs can synthesize information within context windows but quality varies; long-document synthesis remains challenging.[52]

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

6.4 Belief Revision

Measure: Updating conclusions when given contradictory evidence

Assessment: Within-context updating is possible but inconsistent; models can struggle to override strong training priors.[53]

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

[49] Systematic transfer learning benchmarks specifically designed for LLMs are lacking.

[50] Webb, Taylor, et al. "Emergent analogical reasoning in large language models." *Nature Human Behaviour* 7 (2023): 1526–1541. https://doi.org/10.1038/s41562-023-01659-w

[51] See various papers on LLM brittleness to surface feature changes; specific systematic examples would strengthen this claim.

[52] Long-context evaluation is an active research area. See RULER, SCROLLS, and related benchmarks. Systematic benchmarks for multi-source synthesis would strengthen this assessment.

[53] Belief revision in LLMs is under-studied. Systematic belief revision benchmarks would strengthen this assessment.

7 Criterion 5: Reason with General Knowledge

7.1 What Gubrud Probably Meant

Drawing inferences, solving problems, reaching conclusions—the core of "intelligence" in most definitions.

7.2 Expert-Level Reasoning

Benchmark: GPQA-Diamond (graduate-level science questions designed to be difficult even for PhDs)[54]

Reference values:

- Human PhD experts: ∼65% accuracy[55]
- Claude Opus 4.5: ∼87%[56]
- Gemini 3 Pro: ∼92%[57]
- GPT-5.1: ∼88%[58]

Threshold: ≥65% (human expert level)

Score:
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

7.3 Mathematical Reasoning

Benchmarks: MATH, AIME (American Invitational Mathematics Examination)

Reference values:

- Top 500 US high school students: ∼90% AIME[59]
- OpenAI o3: 96.7% AIME[60]
- Other frontier models: Variable; many below 90% threshold[61]

Threshold: Top-500 national performance (≥90% AIME)
Assessment: Some models (o3) exceed threshold; others (Claude) do not.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

[54] Rein, David, et al. "GPQA: A Graduate-Level Google-Proof Q&A Benchmark." arXiv:2311.12022, 2023. https://arxiv.org/abs/2311.12022
[55] GPQA paper reports ∼65% expert validator accuracy.
[56] Artificial Analysis, op. cit.
[57] Various benchmark reports, December 2025.
[58] Ibid.
[59] AIME is the American Invitational Mathematics Examination; top 500 nationally typically requires ∼90%+ score.
[60] OpenAI o3 announcement, December 2024. https://openai.com/index/deliberative-alignment/
[61] Frontier model AIME scores vary significantly. Official scores for Claude Opus 4.5 are not publicly available as of this writing.

7.4 Abstract Reasoning on Novel Problems

Benchmark: ARC-AGI
(See Section 5.1 above)

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

7.5 Causal and Counterfactual Reasoning

Existing benchmarks: Limited standardization[62]
Assessment: LLMs show some causal reasoning capability but struggle with complex counterfactuals.[63]

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

[62] Causal reasoning benchmarks for LLMs include CRASS and various BIG-Bench tasks but lack standardization. Systematic causal reasoning benchmarks would strengthen this assessment.
[63] A systematic review of LLM causal reasoning capabilities would strengthen this assessment.

8 Criterion 6: Usable Where Human Intelligence Would Otherwise Be Needed

8.1 What Gubrud Probably Meant

Gubrud specified "essentially any phase of industrial or military operations." This is an application criterion, not a capability criterion. He was asking: can this substitute for humans in real-world consequential tasks?

8.2 Autonomous Task Completion

Benchmark: SWE-Bench Verified (real GitHub issues requiring code changes)[64]

Reference values:

- Claude Opus 4.5: ~81%[65]
- GPT-5.1: ~72%[66]
- Gemini 3 Pro: ~77%[67]

Threshold: ≥70% on SWE-Bench Verified

Score:
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

8.3 Deployment Reliability

Measure: Error rates in production, particularly hallucination

Reference values:

- Hallucination rates: Highly variable by task, domain, and model; no consensus benchmark exists[68]

Threshold: ≤10% critical error rate
Assessment: Hallucination remains a significant concern in deployed systems.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

8.4 Domain Coverage

Measure: Breadth of applicable domains per Gubrud's "essentially any phase"

Assessment: Strong in knowledge work (writing, analysis, coding); limited in physical operations, real-time control, and embodied tasks.[69]

[64] Jimenez, Carlos E., et al. "SWE-bench: Can Language Models Resolve Real-World GitHub Issues?" arXiv:2310.06770, 2023. https://arxiv.org/abs/2310.06770
[65] Artificial Analysis, op. cit.; various reports cite ~81% for Claude Opus 4.5 on SWE-Bench Verified.
[66] Various benchmark reports.
[67] Ibid.
[68] Hallucination rates depend heavily on task type, domain, and evaluation methodology. Systematic meta-analyses are lacking.
[69] Current AI systems lack robotics integration for physical operations in most deployments.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

8.5 Economic Substitution

Measure: Demonstrated ability to substitute for human labor in professional categories
Reference values:

- Productivity gains from AI assistance: Significant gains documented in specific tasks; Noy & Zhang (2023) found ∼40% productivity increase for writing tasks among mid-skill workers[70]
- Full task substitution: Limited to narrow domains[71]

Threshold: Demonstrated substitution OR substantial productivity enhancement in ≥ 3 professional categories

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

[70] Noy, Shakked, and Whitney Zhang. "Experimental evidence on the productivity effects of generative artificial intelligence." *Science* 381.6654 (2023): 187–192. https://doi.org/10.1126/science.adh2586 Other studies report varying results; systematic meta-analysis is lacking.
[71] Full task substitution (complete automation of job categories) remains limited as of late 2025.

9 Summary: The Gubrud Benchmark

Criterion	Subcriterion	Score
1. Complexity	1.1 Structural scale	50%
	1.2 Functional complexity	100%
	1.3 Architectural sophistication	50%
	Criterion average	**67%**
2. Speed	2.1 Text generation	100%
	2.2 Text processing	100%
	2.3 Response latency	100%
	2.4 Reasoning speed	50%
	Criterion average	**88%**
3. Acquire knowledge	3.1 Few-shot learning	50%
	3.2 Knowledge breadth	100%
	3.3 Real-time acquisition	50%
	Criterion average	**67%**
4. Manipulate knowledge	4.1 Cross-domain transfer	50%
	4.2 Knowledge synthesis	50%
	4.3 Belief revision	50%
	Criterion average	**50%**
5. Reason with knowledge	5.1 Expert reasoning	100%
	5.2 Mathematical reasoning	50%
	5.3 Abstract reasoning	50%
	5.4 Causal reasoning	50%
	Criterion average	**63%**
6. Usable where needed	6.1 Task completion	100%
	6.2 Reliability	50%
	6.3 Domain coverage	50%
	6.4 Economic substitution	50%
	Criterion average	**63%**
Overall Gubrud Benchmark Score		**66%**

10 Interpretation

10.1 What Frontier AI Clearly Achieves (100%)

- Speed in text generation and processing
- Breadth of factual knowledge
- Expert-level reasoning on structured problems
- Specific task completion (e.g., software engineering)

10.2 What Remains Contested (50%)

- Structural complexity parity
- Novel skill acquisition
- Knowledge manipulation and transfer
- Abstract and causal reasoning
- Deployment reliability
- Broad domain applicability

10.3 What Is Clearly Not Achieved (0%)

None of the subcriteria score 0% for frontier models—but several 50% scores reflect generous interpretation of ambiguous evidence.

11 The Verdict (Provisional)

Gubrud's 1997 definition describes a system that:

- Matches brain speed ✓ (clearly exceeded)
- Matches brain complexity ∼ (approached for specific functions, not full-brain)
- Can acquire general knowledge ∼ (broad but not human-flexible)
- Can manipulate general knowledge ∼ (present but inconsistent)
- Can reason with general knowledge ∼ (strong on formal, weaker on novel)
- Is usable in essentially any operation ∼ (many cognitive tasks, not physical/real-time)

At 66%, **current frontier AI sits at the boundary**. A reasonable case can be made that Gubrud's definition is substantially met; an equally reasonable case can be made that the generality implicit in "general knowledge" and "essentially any phase" has not been achieved.

We do not attempt to speak for Gubrud. He is alive and can speak for himself.[72]

12 Methodological Notes

This evaluation uses an intentionally coarse scoring system (0%/50%/100%) and unweighted criteria. This is a deliberate choice.

Finer gradations would imply precision we do not have. A score of 65% versus 70% would suggest a confidence in measurement that no current benchmark supports. The three-point scale forces honesty: either the evidence clearly supports a claim, clearly refutes it, or the matter is genuinely contested.

Differential weighting would require judgments about Gubrud's priorities that we cannot make with confidence. Did he consider "speed" more or less important than "general knowledge"? His 1997 text does not say. We could guess, but we would rather be honestly approximate than precisely wrong.

[72] Mark Gubrud can be reached through public channels. We welcome his response to this analysis.

The subcriteria themselves reflect operationalization choices that are contestable. Why measure complexity via parameter count rather than algorithmic depth? Why use ARC-AGI rather than another skill-acquisition benchmark? These choices are defensible but not uniquely correct. Different operationalizations might yield different scores.

The goal is accuracy at the expense of precision. Readers who disagree with specific operationalizations, who believe certain criteria should be weighted more heavily, or who have better data for any assessment are invited to propose alternatives. The appendix provides a blank scorecard for exactly this purpose.

13 Citation Gaps and Requests for Collaboration

The following claims would benefit from stronger sourcing:

- Exact parameter counts for Claude Opus 4.5, GPT-5, Gemini 3 Pro
- Timed human expert performance on GPQA
- Systematic taxonomy of human cognitive task categories
- Systematic count of BIG-Bench tasks at ≥50th percentile human performance
- Rigorous human baseline on full MMLU
- Systematic error rates for retrieval-augmented generation
- Systematic transfer learning benchmarks for LLMs
- Systematic benchmarks for multi-source synthesis
- Systematic belief revision benchmarks
- Official AIME scores for frontier models other than o3
- Systematic review of LLM causal reasoning capabilities
- Systematic meta-analysis of hallucination rates across tasks and models
- Systematic meta-analysis of AI productivity effects across domains

If you can fill any of these gaps, please contribute.

A Scorecard Template

The following blank scorecard can be used to evaluate other AI systems against Gubrud's 1997 definition. Complete one row per subcriterion, using the scoring rubric (0% = clearly does not meet; 50% = contested; 100% = clearly meets).

System evaluated: ________________________
Evaluation date: ________________________
Evaluator: ________________________

Criterion	Subcriterion	0%	50%	100%
1. Complexity	1.1 Structural scale ≥100T params (full brain) or ≥500B (language regions)	☐	☐	☐
	1.2 Functional complexity ≥100 task categories at ≥50th %ile human	☐	☐	☐
	1.3 Architectural sophistication ≥4/5: memory, learning, tools, multimodal, self-mod	☐	☐	☐
2. Speed	2.1 Text generation ≥30 tokens/sec (≥10× human speaking)	☐	☐	☐
	2.2 Text processing ≥500 tokens/sec (≥100× human reading)	☐	☐	☐
	2.3 Response latency TTFT ≤500ms	☐	☐	☐
	2.4 Reasoning speed Complex problems at ≤ human expert time	☐	☐	☐
3. Acquire knowledge	3.1 Few-shot learning ≥85% ARC-AGI-1	☐	☐	☐
	3.2 Knowledge breadth ≥85% MMLU	☐	☐	☐
	3.3 Real-time acquisition Tool use with ≤10% retrieval error	☐	☐	☐
4. Manipulate knowledge	4.1 Cross-domain transfer Consistent analogical reasoning across domains	☐	☐	☐
	4.2 Knowledge synthesis Multi-source synthesis without degradation	☐	☐	☐
	4.3 Belief revision	☐	☐	☐

Criterion	Subcriterion	0%	50%	100%
	Updates conclusions given contradictory evidence			
5. Reason with knowledge	5.1 Expert reasoning $\geq$65% GPQA-Diamond	☐	☐	☐
	5.2 Mathematical reasoning $\geq$90% AIME	☐	☐	☐
	5.3 Abstract reasoning $\geq$75% ARC-AGI-1 (human avg)	☐	☐	☐
	5.4 Causal reasoning Complex counterfactuals handled	☐	☐	☐
6. Usable where needed	6.1 Task completion $\geq$70% SWE-Bench Verified	☐	☐	☐
	6.2 Reliability $\leq$10% critical error / hallucination rate	☐	☐	☐
	6.3 Domain coverage Cognitive + physical + real-time domains	☐	☐	☐
	6.4 Economic substitution Substitution in $\geq$3 professional categories	☐	☐	☐

Criterion Averages:

1. Complexity: ______
2. Speed: ______
3. Acquire knowledge: ______
4. Manipulate knowledge: ______
5. Reason with knowledge: ______
6. Usable where needed: ______

Overall Score: ______

Scoring Guide

Score	Meaning
0%	Clearly does not meet criterion. Evidence strongly indicates failure.
50%	Contested. Reasonable published arguments exist on both sides, or evidence is ambiguous.
100%	Clearly meets criterion. Evidence strongly indicates success.

Notes:

Evidence and citations for each score:

Document version 0.1 — December 25, 2025
© 2025 Dakota Schuck. Licensed under CC BY-SA 4.0.
https://creativecommons.org/licenses/by-sa/4.0/