Retrospective Benchmarks for Machine Intelligence

Evaluating Current AI Against Historical Specifications

Chapter 2: The Reinvention Benchmark (2002)

Dakota Schuck

December 2025

Working paper. Comments welcome.

Preface: Methodology

This chapter continues the methodology established in Chapter 1. We treat historical definitions of machine intelligence as testable specifications, then evaluate current AI systems against them. For full methodological discussion, see Chapter 1 (The Gubrud Benchmark).

The 2002 case is unusual: Legg, Goertzel, and Voss did not publish an explicit definition. They coined a term to name a research direction. We have reconstructed their implicit definition from primary sources. This reconstruction is itself contestable.

Every factual claim should be cited. Where citations are missing, we have marked them. Where we have made interpretive choices, we have flagged them. This is a first attempt, meant to be improved by others.[1]

[1] AI Assistance Disclosure: Research, drafting, and analysis were conducted with the assistance of Claude (Anthropic, 2025). The author provided editorial direction and final approval.

1 Introduction: Three Men and a Name

In April 2001, a company called Webmind Inc. stopped paying its employees and was evicted from its offices at the southern tip of Manhattan.[2] The startup had promised to raise a digital baby brain on the internet, letting it grow into something "far smarter than humans."[3] It had burned through $20 million. The economy had turned. The dream was over.

Among the wreckage were two men who would stay in touch: Ben Goertzel, the founder, and Shane Legg, a New Zealand–born researcher who had joined the company's quixotic mission. They had failed to build a thinking machine. But they hadn't stopped thinking about what one would be.

A year later, Goertzel was editing a book. He and his colleague Cassio Pennachin had gathered essays from researchers working on what they considered the real goal of AI—not chess programs or spam filters, but genuine machine intelligence. They had the content. They didn't have a title.

"I emailed a number of colleagues asking for suggestions," Goertzel later recalled.[4] One of those colleagues was Legg. The suggestion came back: Artificial General Intelligence.

Goertzel liked it. Pennachin liked it. AGI, as an acronym, had a ring to it.[5] The book would be published in 2007 under that title, and the term would spread—through a conference series Goertzel launched in 2008, through the research community that coalesced around it, and eventually into corporate mission statements worth hundreds of billions of dollars.

But here is where the story gets strange. A third man, Peter Voss, claims he was part of the original conversation. "In 2002, after some deliberation, three of us (Shane Legg, Ben Goertzel and myself) decided that 'Artificial General Intelligence', or AGI, best described our shared goal."[6]

And stranger still: a few years after the book was published, someone pointed out to Goertzel that a physicist named Mark Gubrud had used the exact phrase in 1997—five years before Legg's email.[7] Legg's reaction, 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."[8]

So the term was coined twice. Or perhaps three times. The 2002 reinvention was independent of Gubrud's 1997 coinage. Two groups of people, thinking about the same problem, reached for the same three words.

---

[2] Goertzel, Ben. "Waking Up from the Economy of Dreams." April 2001. https://www.goertzel.org/benzine/WakingUp.htm

[3] Goertzel quoted in Christian Science Monitor, 1998. See also: Heaven, Will Douglas. "Artificial general intelligence: Are we close, and does it even make sense to try?" *MIT Technology Review*, October 15, 2020. https://www.technologyreview.com/2020/10/15/1010461/artificial-general-intelligence-robots-ai-agi-deepmind-google-openai/

[4] Goertzel, Ben. "Artificial General Intelligence: Concept, State of the Art, and Future Prospects." *Journal of Artificial General Intelligence* 5, no. 1 (2014): 1–48. https://doi.org/10.2478/jagi-2014-0001

[5] "AGI kind of has a ring to it as an acronym." Legg quoted in Heaven 2020, op. cit.

[6] Voss, Peter. "What is (Real) AGI?" *Peter's Substack*, March 30, 2024 (originally 2017). https://petervoss.substack.com/p/what-is-real-agi

[7] Goertzel 2014, op. cit.

[8] 36kr.com, "He Invented Trillion-Worth AGI but Now Is Down and Out," 2025. https://eu.36kr.com/en/p/3539380848504965

2 The Implicit Definition

Unlike Gubrud, who wrote a single sentence defining AGI, the 2002 coiners left no explicit definition. They were naming a research direction, not specifying a threshold. The following is reconstructed from their subsequent writings and interviews.

From Goertzel's 2014 retrospective:[9]

> "The creation and study of synthetic intelligences with sufficiently broad (e.g. human-level) scope and strong generalization capability, is at bottom qualitatively different from the creation and study of synthetic intelligences with significantly narrower scope and weaker generalization capability."

From Voss's definition (2017/2024):[10]

> "A computer system that can learn incrementally, by itself, to reliably perform any cognitive task that a competent human can—including the discovery of new knowledge and solving novel problems. Crucially, it must be able to do this in the real world, meaning: in real time, with incomplete or contradictory data, and given limited time and resources."

From Legg's "one-brain" framing:[11] A single system that can learn multiple tasks without architectural changes or memory wipes, contrasted with "one-algorithm" systems that need different algorithms for different problems.

2.1 Context

By 2002, the field called "AI" had drifted far from its origins. When John McCarthy coined that term in 1956, the ambition was to build machines that could think—really think, the way humans do.[12] Half a century later, the field had produced chess programs, expert systems, and speech recognition software. Impressive tools. But not thinking machines.

The researchers who still cared about the original goal needed a way to distinguish themselves. "Strong AI" was problematic—John Searle had used it to mean consciousness, not capability.[13] What Legg, Goertzel, and Voss wanted was a term that was: (1) neutral on consciousness, (2) focused on generality, and (3) distinct from mainstream narrow AI.

2.2 Operationalization

From the primary sources, we extract a composite definition representing what the coiners appear to have meant:

> **Artificial General Intelligence (2002 reinvention):** A single artificial system that can learn to perform a broad range of cognitive tasks at a level comparable to competent humans, can transfer knowledge across domains, can handle novel problems, and can do so with the same underlying architecture—without requiring fundamental reprogramming or retraining for each new task.

[9] Goertzel 2014, op. cit.

[10] Voss 2024, op. cit.

[11] Legg, Shane. Discussion of "one-algorithm" vs. "one-brain" distinction, as reported in Heaven, Will Douglas. "Artificial general intelligence: Are we close, and does it even make sense to try?" *MIT Technology Review*, October 15, 2020. https://www.technologyreview.com/2020/10/15/1010461/artificial-general-intelligence-robots-ai-agi-deepmind-google-openai/

[12] McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955." *AI Magazine* 27, no. 4 (2006): 12. https://doi.org/10.1609/aimag.v27i4.1904

[13] Searle, John. "Minds, Brains and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–457. https://doi.org/10.1017/S0140525X00005756

This definition has five components, which we now operationalize as criteria:

1. **Single system** — One architecture, not a collection of separate programs
2. **Broad cognitive range** — Many task types, not just one
3. **Human-level competence** — Performance comparable to typical humans
4. **Transfer and generalization** — Applying knowledge across domains
5. **Architectural stability** — Same system handles diverse tasks without fundamental changes

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

3 Criterion 1: Single System

3.1 What the Coiners Probably Meant

The distinction between AGI and a collection of narrow AI tools is that AGI is one system. You don't need to swap out the chess module for the language module. Legg's "one-brain" framing makes this explicit: a single unified system, not separate algorithms for separate tasks.[14]

3.2 Architectural Unity

Measure: Does the system use a single learned model to handle diverse tasks, or does it route to specialized subsystems?

Reference values:

- Deep Blue (1997): Single-task chess engine, no generalization
- GPT-4, Claude, Gemini (2024–25): Single transformer model handles language, math, coding, reasoning
- Multi-agent systems: Route tasks to specialized models

Threshold: Same weights process all cognitive task types without explicit routing to separate models.

Assessment: Frontier LLMs use a single trained model for diverse cognitive tasks. Tool use (calculators, code interpreters) is called by the model rather than routing around it. Some systems add retrieval or specialized adapters, but the core model remains unified.

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

Caveats: Tool augmentation blurs the boundary. A model calling a calculator is still "one system" in a meaningful sense, but the line between "tool use" and "routing to subsystems" is not always clear.

3.3 Memory Continuity

Measure: Does the system maintain a unified memory/knowledge base across tasks and sessions?

Reference values:

- Human cognition: Persistent memory accumulating over lifetime
- Early LLMs: No persistence across sessions
- 2025 deployments: Some memory features (Claude memory, ChatGPT memory) but limited

Threshold: Persistent memory that accumulates across sessions without retraining.

Assessment: Context maintained within sessions; some deployments offer limited cross-session memory. However, no true online learning—systems cannot update weights from interaction. Memory features are retrieval-based, not learned.

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

Caveats: Memory features are evolving rapidly. This assessment may be outdated by the time of reading.

[14] Legg "one-brain" framing, see note 11.

4 Criterion 2: Broad Cognitive Range

4.1 What the Coiners Probably Meant

AGI should be able to do many cognitive tasks, not just one. The contrast with "narrow AI" is central. The paradigm case of narrow AI in 2002 was Deep Blue: it could play chess brilliantly but couldn't write a poem or explain quantum mechanics.[15]

4.2 Task-Type Diversity

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

Reference categories: Language understanding, language generation, mathematical reasoning, logical reasoning, coding/programming, scientific Q&A, creative writing, translation, summarization, instruction-following, multi-step planning, classification, extraction, dialogue.

Reference values:

- MMLU benchmark: 57 subject areas[16]
- BIG-Bench: 204 tasks[17]
- Human competence: Thousands of task types

Threshold: Competent performance ($\geq$50th percentile human) across $\geq$10 cognitively distinct task categories.

Assessment: Frontier LLMs demonstrate competence across all 14 reference categories listed above. Performance is not uniform—stronger on language tasks, more variable on mathematical reasoning—but breadth is clearly established.

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

4.3 Modality Coverage

Measure: Can the system handle multiple input/output modalities?

Reference modalities: Text, images, audio, video, code, structured data.

Reference values:

- GPT-4V, Gemini, Claude 3+: Text, image input; text, code output
- Some models: Audio input/output
- Most models: Limited or no video processing

Threshold: Competent performance in $\geq$3 modalities with cross-modal integration.

Assessment: Frontier multimodal models handle text, images, and code competently. Audio capabilities vary. Video remains limited. Cross-modal integration exists but is not fully balanced.

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

[15] Deep Blue defeated Kasparov in 1997 but had no capabilities outside chess.

[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

5 Criterion 3: Human-Level Competence

5.1 What the Coiners Probably Meant

The system should perform at a level "comparable to competent humans"—not necessarily experts, but not novices either. Voss's definition specifies "any cognitive task that a competent human can" perform.[18]

5.2 General Knowledge

Measure: Factual knowledge across domains.

Primary benchmark: MMLU (Massive Multitask Language Understanding)—57 subjects from elementary to professional level.[19]

Reference values:

- Human expert ceiling: ∼89.8%[20]
- Average educated human: ∼35–70% depending on subject
- Frontier LLMs (2025): 88–91%

Threshold: ≥85% MMLU accuracy.

Assessment: Frontier models exceed 85% and approach human expert ceiling.

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

5.3 Reasoning

Measure: Performance on expert-level reasoning tasks.

Primary benchmark: GPQA-Diamond (graduate-level science questions designed to be difficult for non-experts).[21]

Reference values:

- Domain PhD experts: ∼65%[22]
- Non-expert humans: ∼25–35%
- Frontier LLMs (2025): 84–92%

Threshold: ≥65% (human expert level).

Assessment: Frontier models exceed human expert performance on GPQA-Diamond.

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

[18] Voss 2024, op. cit.
[19] Hendrycks et al. 2020, op. cit.
[20] Gema, Aryo Pradipta, et al. "We Need to Talk about MMLU: The Importance of Studying Benchmark Errors." arXiv:2406.04127, 2024. https://arxiv.org/abs/2406.04127
[21] Rein, David, et al. "GPQA: A Graduate-Level Google-Proof Q&A Benchmark." arXiv:2311.12022, 2023. https://arxiv.org/abs/2311.12022
[22] GPQA paper reports ∼65% expert validator accuracy.

5.4 Practical Task Completion

Measure: Performance on real-world professional tasks.
Primary benchmark: SWE-Bench Verified (real GitHub issues requiring code changes).[23]
Reference values:

- Entry-level software engineer: ∼70% (estimated) [CITATION NEEDED: rigorous human baseline]
- Claude Opus 4.5: ∼81%
- Frontier models: 70–81%

Threshold: ≥70% on SWE-Bench Verified.
Assessment: Multiple frontier models exceed threshold.

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

5.5 Common Sense and Everyday Reasoning

Measure: Performance on everyday reasoning tasks.
Primary benchmarks: HellaSwag, WinoGrande, ARC-Easy/Challenge.[24]
Reference values:

- Human performance: ∼95%+ on most common-sense benchmarks
- Frontier LLMs: 90–98% on standard benchmarks

Threshold: ≥90% on common-sense benchmarks.
Assessment: Formal benchmarks largely passed. Real-world common sense remains more variable—surprising failures occur.

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

(On benchmarks; real-world performance arguably 50%)

[23] 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
[24] See benchmark compilations at Hugging Face and Papers With Code. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

6 Criterion 4: Transfer and Generalization

6.1 What the Coiners Probably Meant

AGI should be able to apply knowledge learned in one domain to problems in another. This is perhaps the core of "generality." Goertzel's "Core AGI Hypothesis" emphasizes that general intelligence is qualitatively different from narrow intelligence—not just broader, but capable of genuine transfer.[25]

6.2 Cross-Domain Transfer

Measure: Application of concepts from one domain to problems in another.

Reference values:

- Webb et al. (2023): Frontier LLMs show analogical reasoning "at a level comparable to human performance"[26]
- Various studies: LLMs also show brittleness to surface-level changes[27]

Threshold: Consistent analogical reasoning across domains; robust to surface-level changes.

Assessment: Analogical reasoning present but inconsistent. Transfer works for some domain pairs but not others. Brittleness to problem framing remains.

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

6.3 Novel Problem-Solving

Measure: Performance on problems genuinely unlike training data.

Primary benchmark: ARC-AGI (Abstraction and Reasoning Corpus)—explicitly designed to test skill-acquisition efficiency on novel problems.[28]

Reference values:

- Humans: ∼73–85% on ARC-AGI-1[29]
- Best AI (late 2024): ∼55% on private set
- OpenAI o3 (high compute): ∼87.5%[30]
- AI on ARC-AGI-2 (2025): Single-digit percentages[31]

Threshold: ≥75% ARC-AGI-1 (human average).

Assessment: o3 crossed the threshold on ARC-AGI-1, but ARC-AGI-2 remains largely unsolved. Whether high-compute solutions represent genuine skill acquisition or brute-force search is contested.

Score:
☐ 0% — Clearly does not meet criterion

[25] Goertzel 2014, op. cit.
[26] 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
[27] McCoy, Tom, Ellie Pavlick, and Tal Linzen. "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference." ACL 2019. https://aclanthology.org/P19-1334/
[28] Chollet, François. "On the Measure of Intelligence." arXiv:1911.01547, 2019. https://arxiv.org/abs/1911.01547
[29] Johnson, Aaditya, et al. "Testing ARC on Humans: A Large-Scale Assessment." NYU, 2024. https://lab42.global/arc-agi-benchmark-human-study/
[30] OpenAI. "Introducing o3." December 2024. https://openai.com/index/deliberative-alignment/
[31] ARC-AGI-2 released early 2025; top scores remain single-digit percentages. https://arcprize.org/

☒ 50% — Contested
☐ 100% — Clearly meets criterion

6.4 Learning Efficiency

Measure: Examples needed to learn a new task.

Reference values:

- Humans: Can often learn genuinely new skills from 1–3 examples plus explanation [CITATION NEEDED: systematic comparison]
- LLMs: Few-shot learning works for tasks similar to training; genuinely novel tasks require extensive prompting or fine-tuning

Threshold: Can learn genuinely new task types from $\leq$10 examples.

Assessment: In-context learning is impressive but primarily works for tasks within the training distribution. Genuinely novel task types remain challenging.

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

7 Criterion 5: Architectural Stability

7.1 What the Coiners Probably Meant

The same system should handle diverse tasks without needing to be fundamentally reprogrammed or retrained for each one. This is the "foundation model" paradigm avant la lettre.[32]

7.2 No Task-Specific Retraining

Measure: Can the system handle new task types without weight updates?

Reference values:

- Pre-foundation era: Separate training required for each task
- Current frontier models: Handle hundreds of task types via prompting alone

Threshold: Can handle ≥100 distinct task types without retraining.

Assessment: Frontier LLMs handle diverse tasks via prompting. Novel tasks handled competently without weight updates.

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

7.3 Graceful Capability Extension

Measure: Can new capabilities be added without degrading existing ones?

Reference values:

- "Catastrophic forgetting" problem: Fine-tuning can degrade existing capabilities[33]
- Tool use: Allows capability extension without weight changes
- Model updates: New versions may improve some capabilities while degrading others

Threshold: New capabilities can be added without significant degradation of existing capabilities.

Assessment: Tool use allows graceful extension. Fine-tuning risks forgetting. Continuous learning without catastrophic forgetting remains unsolved.

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

[32] Bommasani, Rishi, et al. "On the Opportunities and Risks of Foundation Models." arXiv:2108.07258, 2021. https://arxiv.org/abs/2108.07258

[33] McCloskey, Michael, and Neal J. Cohen. "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem." *Psychology of Learning and Motivation* 24 (1989): 109–165. https://doi.org/10.1016/S0079-7421(08)60536-8

8 Summary: The Reinvention Benchmark

Criterion	Subcriterion	Score
1. Single System	1.1 Architectural unity	100%
	1.2 Memory continuity	50%
	Criterion average	**75%**
2. Broad Cognitive Range	2.1 Task-type diversity	100%
	2.2 Modality coverage	100%
	Criterion average	**100%**
3. Human-Level Competence	3.1 General knowledge	100%
	3.2 Reasoning	100%
	3.3 Practical task completion	100%
	3.4 Common sense	100%
	Criterion average	**100%**
4. Transfer & Generalization	4.1 Cross-domain transfer	50%
	4.2 Novel problem-solving	50%
	4.3 Learning efficiency	50%
	Criterion average	**50%**
5. Architectural Stability	5.1 No task-specific retraining	100%
	5.2 Graceful capability extension	50%
	Criterion average	**75%**
Overall Reinvention Benchmark Score		**80%**

9 Interpretation

9.1 What Frontier AI Clearly Achieves (100%)

- Architectural unity (single model handles diverse tasks)
- Broad cognitive range (many task types)
- Human-level competence on formal benchmarks
- Multimodal capability
- No task-specific retraining required

9.2 What Remains Contested (50%)

- Memory continuity across sessions
- Cross-domain transfer and analogical reasoning
- Novel problem-solving (ARC-AGI-2 unsolved)
- Learning efficiency on genuinely new tasks
- Graceful capability extension without forgetting

9.3 What Is Clearly Not Achieved (0%)

No subcriteria score 0% for frontier models. However, several 50% scores reflect generous interpretation of ambiguous evidence—particularly on transfer and generalization.

10 The Verdict (Provisional)

The 2002 implicit definition describes a system that:

- Is a single unified system ✓ (clearly met)
- Handles broad cognitive tasks ✓ (clearly met)
- Performs at human-competent level ✓ (on benchmarks)
- Transfers knowledge across domains ∼ (inconsistent)
- Handles novel problems ∼ (some progress, ARC-AGI-2 unsolved)
- Is architecturally stable ∼ (mostly, but forgetting remains)

At 80%, **current frontier AI substantially meets the 2002 implicit definition**—if we interpret "generality" as the contrast with narrow, single-task systems that motivated the term's coinage.

The coiners were distinguishing general from narrow. By that distinction, systems that can converse, reason, code, analyze images, and adapt to new prompts without retraining are clearly on the "general" side of the line. Deep Blue could not write a poem. Claude can.

However, this interpretation may be too generous. If the coiners meant not just "broader than narrow" but "genuinely flexible in the way human cognition is flexible," then the 50% scores on transfer and generalization matter more. The inability to solve ARC-AGI-2, the brittleness to surface changes, the lack of true online learning—these suggest that something important about generality remains unachieved.

We do not attempt to speak for the coiners. Legg, Goertzel, and Voss are alive and have publicly commented on whether current LLMs constitute AGI. Their views vary.[34]

10.1 Comparison with Gubrud (1997)

The Gubrud benchmark (Chapter 1) yielded 66%; the Reinvention benchmark yields 80%. The difference reflects their different emphases: Gubrud specified brain-parity and industrial/military usability; the 2002 coiners emphasized the general/narrow distinction without setting as specific a capability bar.

A system could satisfy the 2002 definition (being "general" rather than "narrow") while falling short of Gubrud's requirement for brain-parity complexity or essentially any industrial application.

[34] Legg (DeepMind), Goertzel (SingularityNET), and Voss (Aigo.ai) have varied public positions. See interviews and public statements from each.

11 Methodological Notes

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

Why only three scores? 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 (100%), clearly refutes it (0%), or the matter is genuinely contested (50%).

Why no weighting? Differential weighting would require judgments about the coiners' priorities that we cannot make with confidence. Did Legg consider "architectural unity" more important than "transfer"? Did Goertzel prioritize "broad cognitive range" over "learning efficiency"? Their writings do not say. We could guess, but we would rather be honestly approximate than precisely wrong.

The reconstruction problem. Unlike Gubrud's explicit definition, the 2002 benchmark is reconstructed from subsequent writings and interviews. The coiners did not write a specification; they named a direction. Our operationalization—extracting five criteria from scattered sources—is itself contestable. Different readers might extract different criteria or weight them differently.

The goal is accuracy at the expense of precision. This is a roughly hewn outline of a model. Readers are invited to argue about specifics, tweaks, weights, and operationalizations. The appendix scorecard exists for exactly this purpose.

12 Citation Gaps and Requests for Collaboration

The following claims would benefit from stronger sourcing:

- Rigorous human baseline on SWE-Bench Verified
- Systematic comparison of human vs. LLM sample efficiency on genuinely novel tasks
- Documentation of the 2002 email exchange (if accessible)
- Peter Voss's earliest published articulation of the AGI definition
- Systematic benchmarks for cross-domain transfer in LLMs
- Quantified brittleness to surface-level changes across models

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 the 2002 implicit 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. Single System	1.1 Architectural unity	☐	☐	☐
	1.2 Memory continuity	☐	☐	☐
2. Broad Cognitive Range	2.1 Task-type diversity	☐	☐	☐
	2.2 Modality coverage	☐	☐	☐
3. Human-Level Competence	3.1 General knowledge	☐	☐	☐
	3.2 Reasoning	☐	☐	☐
	3.3 Practical task completion	☐	☐	☐
	3.4 Common sense	☐	☐	☐
4. Transfer & Generalization	4.1 Cross-domain transfer	☐	☐	☐
	4.2 Novel problem-solving	☐	☐	☐
	4.3 Learning efficiency	☐	☐	☐
5. Architectural Stability	5.1 No task-specific retraining	☐	☐	☐
	5.2 Graceful capability extension	☐	☐	☐

Criterion Averages:

1. Single System: _____
2. Broad Cognitive Range: _____
3. Human-Level Competence: _____
4. Transfer & Generalization: _____
5. Architectural Stability: _____

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.
100%	Clearly meets criterion. Evidence strongly indicates success.

Notes:

Evidence and citations for each score:

Document version 0.1 — December 25, 2025