Retrospective Benchmarks for Machine Intelligence

Evaluating Current AI Against Historical Specifications

Chapter 3: The Formalization Benchmark (2007)

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 2007 case presents a unique challenge: Legg and Hutter produced the most rigorous formalization of machine intelligence to date, but their measure is technically incomputable. It relies on Kolmogorov complexity, which cannot be calculated for arbitrary strings. This chapter therefore assesses whether current systems exhibit the *properties* the definition points to, rather than computing the measure itself.

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: The Mathematician's Answer

By 2007, Shane Legg had been thinking about intelligence for a decade. His master's thesis at the University of Auckland had been on Solomonoff induction—the mathematical theory of optimal prediction.[2] He had worked at Webmind, watched it collapse, coined the term "AGI" with Ben Goertzel, and landed at IDSIA in Switzerland to work with Marcus Hutter, one of the world's leading theorists of algorithmic information.[3]

The two men shared a frustration. "A fundamental problem in artificial intelligence," they wrote, "is that nobody really knows what intelligence is."[4] Psychologists had their IQ tests, but those were designed for humans and normalized to human populations. Computer scientists had benchmark after benchmark, but each measured something narrow. What was lacking was a formal definition—one grounded in mathematics rather than intuition, applicable to any system rather than just humans, and precise enough to admit no ambiguity.

So they built one.

Their approach was systematic. First, they surveyed the literature, collecting over 70 informal definitions of intelligence from psychologists, AI researchers, and philosophers.[5] From this survey, they extracted common themes: learning, adaptation, goal-achievement, dealing with novel situations, performing well across diverse environments. They distilled these into a single informal definition:

> *Intelligence measures an agent's ability to achieve goals in a wide range of environments.*

Then they did what most researchers had not: they formalized it. Using tools from algorithmic information theory—Kolmogorov complexity, Solomonoff induction, the reinforcement learning framework—they converted the informal definition into a precise mathematical equation. The result was what they called *universal intelligence*: a single number, in principle, that could be computed for any agent, biological or artificial, measuring its intelligence in the broadest reasonable sense.

The equation is elegant:

$$\Upsilon(\pi) = \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi$$

Where π is the agent being evaluated, E is the set of all computable environments, $K(\mu)$ is the Kolmogorov complexity of environment μ, and V_μ^π is the expected reward the agent achieves in that environment. The agent's intelligence is the weighted sum of its performance across all possible environments, with simpler environments counting more (via the $2^{-K(\mu)}$ term, which embodies Occam's razor).

There was just one problem. Kolmogorov complexity is not computable. No algorithm can calculate $K(\mu)$ for arbitrary μ. The definition was mathematically precise but practically unmeasurable—a Platonic ideal of intelligence that could never be directly tested.

Legg knew this. In his 2008 PhD thesis, *Machine Super Intelligence*, he acknowledged: "The main drawback, however, is that the Kolmogorov complexity function K is not computable and

[2] Legg, Shane. "Solomonoff Induction." MSc thesis, University of Auckland, 1996. https://researchspace.auckland.ac.nz/handle/2292/3087

[3] For biographical details, see 36kr.com, "He Invented Trillion-Worth AGI but Now Is Down and Out," 2025. https://eu.36kr.com/en/p/3539380848504965; Wikipedia, "Shane Legg." https://en.wikipedia.org/wiki/Shane_Legg

[4] Legg, Shane, and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence." *Minds and Machines* 17, no. 4 (2007): 391–444. https://arxiv.org/abs/0712.3329

[5] Legg, Shane, and Marcus Hutter. "A Collection of Definitions of Intelligence." In *Advances in Artificial General Intelligence*, edited by Ben Goertzel and Pei Wang, 17–24. IOS Press, 2007. https://arxiv.org/abs/0706.3639

can only be approximated."[6] The definition was meant to capture the concept perfectly, even if measurement required approximation.

Today, Legg is Chief AGI Scientist at Google DeepMind. In 2023, he co-authored a new paper attempting to operationalize AGI progress with practical "Levels of AGI"—a more empirical approach that sidesteps the incomputability problem.[7] But the 2007 formalization remains influential as the most rigorous attempt to define machine intelligence from first principles.

The question we must ask: even if we cannot compute universal intelligence directly, can we assess whether current AI systems exhibit the properties the definition specifies?

[6] Legg, Shane. *Machine Super Intelligence.* PhD thesis, University of Lugano, 2008. p. 24. http://www.vetta.org/documents/Machine_Super_Intelligence.pdf

[7] Morris, Meredith Ringel, et al. "Levels of AGI: Operationalizing Progress on the Path to AGI." arXiv:2311.02462, 2023. Legg is a co-author. https://arxiv.org/abs/2311.02462

2 The Original Definition

From "Universal Intelligence: A Definition of Machine Intelligence," published in *Minds and Machines*, December 2007:[8]

> *Intelligence measures an agent's ability to achieve goals in a wide range of environments.*

The formal measure:

$$\Upsilon(\pi) = \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi$$

Where:

- π is the agent
- E is the space of all computable reward-summable environments
- $K(\mu)$ is the Kolmogorov complexity of environment μ
- V_μ^π is the expected value (sum of discounted rewards) the agent achieves in μ

2.1 Context

Legg and Hutter were working within the tradition of algorithmic information theory, building on Solomonoff's theory of universal prediction and Hutter's AIXI agent—a theoretical model of the optimally intelligent agent.[9] AIXI was provably optimal in a specific sense: no computable agent could outperform it across all environments. But AIXI itself was incomputable.

Universal intelligence was derived from AIXI's "intelligence order relation"—a way of ranking agents by their expected performance across environments.[10] The formalization converted this ranking into a scalar measure.

The definition has several notable properties:

Non-anthropocentric. Unlike IQ tests, which are normalized to human populations, universal intelligence applies equally to humans, animals, and machines. A bee could have its universal intelligence measured (in principle) on the same scale as a human or a supercomputer.

Performance-focused. The definition measures what an agent *does*, not how it does it. Internal mechanisms are irrelevant; only goal-achievement counts.

Occam-weighted. Simple environments count more than complex ones. An agent that excels only at complex, contrived tasks but fails at simple ones will score lower than one that handles simple tasks well. This embodies the intuition that intelligence involves recognizing patterns, and simpler patterns are more fundamental.

Dynamic. The agent-environment framework is interactive. The agent takes actions, receives observations and rewards, and must learn and adapt over time. This is not a static test of knowledge but a dynamic measure of learning ability.

2.2 Operationalization

The formal definition cannot be computed, but we can extract five testable properties that an intelligent agent, according to this definition, should exhibit:

1. **Goal-achievement.** The agent should be able to achieve goals (maximize rewards) in its environment.

[8] Legg and Hutter 2007, op. cit.

[9] Hutter, Marcus. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability.* Springer, 2005. https://doi.org/10.1007/b138233

[10] See Definition 5.14 in Hutter 2005.

2. **Wide environmental range.** The agent should succeed across many different types of environments, not just one.

3. **Learning and adaptation.** The agent should improve its performance as it gains experience in an environment.

4. **Simplicity handling.** The agent should perform well on simple, structured problems (which count most in the measure).

5. **Generality over specialization.** A generalist that performs adequately across many environments should score higher than a specialist that excels at one but fails at others.

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

3 Criterion 1: Goal-Achievement

3.1 What Legg and Hutter Meant

The agent-environment framework places goal-achievement at the center. The agent receives reward signals from the environment, and its objective is to maximize cumulative reward. "The agent's goal is then simply to maximise the amount of reward it receives."[11]

This is not about having preferences or desires in any philosophical sense. It is purely operational: given a reward signal, can the agent act to increase it?

3.2 Reward Maximization in Training

Measure: Do frontier AI systems learn to maximize objective functions during training?

Reference values:

- Reinforcement learning agents (AlphaGo, AlphaZero): Explicitly trained to maximize game-winning reward
- LLMs (GPT-4, Claude, Gemini): Trained via next-token prediction and RLHF to maximize reward models
- All modern deep learning: Gradient descent on loss functions (equivalent to reward maximization)

Threshold: System is trained via objective maximization.

Assessment: All frontier AI systems are trained to maximize some objective function. This is definitionally true of modern machine learning.

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

Caveats: Training-time optimization differs from deployment-time goal pursuit. The question of whether trained models "have goals" at inference time is philosophically contested.

3.3 Instruction-Following as Goal-Achievement

Measure: Can the system achieve user-specified goals via natural language instruction?

Reference values:

- IFEval (instruction-following evaluation): Frontier models achieve 80–90%+ on strict instruction compliance[12]
- SWE-Bench Verified: 70–81% on real software engineering tasks[13]
- GAIA benchmark: Variable performance on real-world assistant tasks[14]

Threshold: ≥75% on instruction-following benchmarks.

Assessment: Frontier models reliably follow instructions and achieve stated goals across many task types.

Score:
☐ 0% — Clearly does not meet criterion

[11] Legg and Hutter 2007, p. 13.

[12] Zhou et al. "Instruction-Following Evaluation for Large Language Models." arXiv:2311.07911, 2023. https://arxiv.org/abs/2311.07911

[13] As cited in Chapters 1–2.

[14] Mialon et al. "GAIA: A Benchmark for General AI Assistants." arXiv:2311.12983, 2023. https://arxiv.org/abs/2311.12983

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

3.4 Autonomous Goal Pursuit

Measure: Can the system pursue goals over extended interactions without step-by-step human guidance?

Reference values:

- Agentic coding tools (Claude Code, Cursor, Devin): Can complete multi-step software tasks[15]
- Research agents: Can conduct extended investigations with web search and tool use
- Current limitations: Require human oversight; struggle with very long-horizon tasks

Threshold: Can autonomously complete multi-step tasks over 10+ actions without human intervention.

Assessment: Frontier systems demonstrate meaningful autonomous goal pursuit in constrained domains. Extended autonomy remains limited.

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

[15] Various product announcements and benchmarks, 2024–2025. See Anthropic, “Introducing Claude Code,” 2025. https://www.anthropic.com/claude-code

4 Criterion 2: Wide Environmental Range

4.1 What Legg and Hutter Meant

The summation over “all computable environments” is central to the definition. An agent that performs well in one environment but poorly in others is not intelligent by this measure. “Intelligence is not simply the ability to perform well at a narrowly defined task; it is much broader. An intelligent agent is able to adapt and learn to deal with many different situations, kinds of problems and types of environments.”[16]

The contrast case they cite is IBM’s Deep Blue: “While Gary Kasparov would still be a formidable player if we were to change the rules of chess, IBM’s Deep Blue chess super computer would be rendered useless without significant human intervention.”[17]

4.2 Task-Type Diversity

Measure: Number of cognitively distinct task types handled competently.

Reference values:

- Deep Blue (1997): 1 task type (chess)
- GPT-2 (2019): Dozens of language tasks
- Frontier LLMs (2025): Hundreds of task categories across language, math, coding, reasoning, creative writing, translation, summarization, extraction, etc.

Threshold: Competent performance across ≥100 cognitively distinct task categories.

Assessment: Frontier models demonstrably handle hundreds of distinct task types. This is the defining feature of “foundation models” and stands in stark contrast to narrow AI of the Deep Blue era.

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

4.3 Environmental Diversity

Measure: Can the system adapt to genuinely different types of environments (not just different tasks within one paradigm)?

Reference values:

- Text-only environments: Strong performance
- Multimodal environments (text + image): Good performance
- Interactive environments (tool use, web browsing): Reasonable performance
- Embodied/robotic environments: Limited (requires specialized systems)
- Real-time physical control: Minimal

Threshold: Competent performance in ≥3 fundamentally different environmental types.

Assessment: Frontier multimodal models operate in text, image, and interactive tool-use environments. Physical embodiment remains a gap.

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

[16] Legg and Hutter 2007, p. 17.

[17] Ibid.

4.4 Robustness to Distribution Shift

Measure: Does performance degrade gracefully when environments differ from training distribution?

Reference values:

- In-distribution performance: Strong
- Moderate distribution shift: Generally robust
- Significant distribution shift: Performance varies; some brittleness documented[18]
- Adversarial environments: Vulnerable

Threshold: Less than 20% performance degradation under moderate distribution shift.

Assessment: Robustness is improving but inconsistent. Some tasks show strong generalization; others reveal brittleness.

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

[18] McCoy et al. "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference." ACL 2019. https://aclanthology.org/P19-1334/

5 Criterion 3: Learning and Adaptation

5.1 What Legg and Hutter Meant

The emphasis on "dynamic tests" over "static tests" is explicit in their paper. Static tests measure current knowledge; dynamic tests measure the ability to learn. "In a dynamic test the test subject interacts over a period of time with the tester, who now becomes a kind of teacher. The tester's task is to present the individual with a series of problems. After each attempt at solving a problem, the tester provides feedback to the individual who then has to adapt their behaviour accordingly."[19]

The AIXI agent, which maximizes universal intelligence, is explicitly a *learning* agent—one that updates its beliefs based on experience.

5.2 In-Context Learning

Measure: Can the system improve performance on a task from examples provided within a single interaction?

Reference values:

- Zero-shot: Reasonable performance on many tasks
- Few-shot (3–5 examples): Consistent improvement across most task types[20]
- Many-shot (50+ examples): Further improvement, especially on novel formats

Threshold: Measurable improvement from zero-shot to few-shot across diverse task types.

Assessment: In-context learning is a defining capability of modern LLMs and represents genuine within-session adaptation.

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

5.3 Skill Acquisition Efficiency

Measure: Examples required to learn a genuinely new task type.

Reference benchmark: ARC-AGI, explicitly designed to test skill acquisition on novel problems.[21]

Reference values:

- Humans: 73–85% on ARC-AGI-1 with typically 3–5 training examples per task[22]
- Best AI (late 2024): ∼55% on ARC-AGI-1 private set
- OpenAI o3 (high compute): ∼87.5% on ARC-AGI-1[23]
- AI on ARC-AGI-2 (2025): Single-digit percentages[24]

Threshold: ≥75% on ARC-AGI-1 with human-comparable example counts.

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

[19] Legg and Hutter 2007, p. 7.

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

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

[22] Johnson et al. "Testing ARC on Humans." NYU, 2024. https://lab42.global/arc-agi-benchmark-human-study/

[23] OpenAI. "Introducing o3." December 2024. https://openai.com/index/deliberative-alignment/

[24] https://arcprize.org, 2025.

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

5.4 Cross-Session Learning

Measure: Can the system improve across multiple sessions via persistent memory or weight updates?

Reference values:

- Humans: Continuous learning across lifetime
- LLMs: No weight updates from interaction (frozen after training)
- Retrieval-augmented memory: Can store and retrieve facts, but not true learning
- Fine-tuning: Possible but typically done by developers, not users

Threshold: True online learning—improving weights from user interactions.

Assessment: Current LLMs do not learn in the sense of updating their weights from deployment-time interactions. Memory features provide continuity but not learning.

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

6 Criterion 4: Simplicity Handling (Occam's Razor)

6.1 What Legg and Hutter Meant

The $2^{-K(\mu)}$ weighting is crucial. Simple environments—those with short algorithmic descriptions—count more than complex ones. "It is important then that the agent is able to quickly learn and adapt so as to perform as well as possible over a wide range of environments, situations, tasks and problems."[25]

They make this concrete with an example: In IQ tests, when asked to continue the sequence 2, 4, 6, 8, the "correct" answer is 10—the simplest pattern. A polynomial that also fits the data but predicts 58 is rejected. "Why then, even if we are aware of the larger polynomial, do we consider the first answer to be the most likely one? It is because we apply, perhaps unconsciously, the principle of Occam's razor."[26]

An intelligent agent should recognize simple patterns and prefer simple hypotheses.

6.2 Simple Pattern Recognition

Measure: Performance on simple, well-structured reasoning tasks.

Reference benchmarks: Elementary math, basic logic, simple analogies.

Reference values:

- GSM8K (grade school math): Frontier models achieve 90%+[27]
- Simple reasoning chains: Near-perfect performance
- Basic pattern completion: Strong performance

Threshold: ≥90% on elementary reasoning benchmarks.

Assessment: Frontier models excel at simple, structured problems.

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

6.3 Preference for Simple Hypotheses

Measure: When multiple explanations fit the data, does the model prefer simpler ones?

Assessment method: Qualitative evaluation of model outputs on ambiguous problems.

Observations:

- LLMs generally prefer simple explanations when prompted for reasoning
- Chain-of-thought prompting encourages step-by-step simple reasoning
- Occasional failures: models sometimes overcomplicate or miss simple patterns
- No formal guarantee of Occam-like behavior

Threshold: Consistent preference for simpler hypotheses in ambiguous cases.

Assessment: Generally exhibits simplicity preference but not reliably. Difficult to test systematically.

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

[25] Legg and Hutter 2007, p. 9.
[26] Ibid., p. 18.
[27] Various benchmark reports, 2024–2025.

6.4 Resistance to Overfitting

Measure: Does the model generalize rather than memorize?

Reference values:

- Generalization on novel phrasings of trained tasks: Generally good
- Generalization to novel task types: Mixed
- Evidence of memorization: Some training data can be extracted[28]

Threshold: Generalizes to novel formulations of trained concepts.

Assessment: Models generalize well in many cases but evidence of memorization exists.

Score:

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

[28] Carlini et al. "Extracting Training Data from Large Language Models." USENIX Security, 2021. https://arxiv.org/abs/2012.07805

7 Criterion 5: Generality Over Specialization

7.1 What Legg and Hutter Meant

The contrast with Deep Blue is instructive. Deep Blue had extremely high performance in one environment (chess) but zero performance in virtually all others. By the universal intelligence measure, this yields a low score: the chess environment is complex (high $K(\mu)$, low $2^{-K(\mu)}$), and Deep Blue scores zero everywhere else.

A generalist agent that performs moderately well across many environments will score higher than a specialist. "We are interested in common themes and general perspectives on intelligence that could be applicable to many kinds of systems."[29]

7.2 Generalist vs. Specialist Architecture

Measure: Is the system designed as a generalist or specialist?

Reference values:

- Deep Blue: Pure specialist (chess only)
- AlphaGo/AlphaZero: Specialist with some generalization (board games)
- Frontier LLMs: Generalist (single model handles diverse tasks)
- Narrow ML models: Specialists (image classifiers, speech recognition)

Threshold: Single model architecture handles diverse task types without task-specific retraining.

Assessment: Frontier LLMs are explicitly designed as generalists. This is the foundation model paradigm.

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

7.3 Performance Breadth vs. Depth

Measure: How does generalist performance compare to specialist performance in specific domains?

Reference values:

- Generalist LLMs vs. specialized chess engines: Specialists vastly superior
- Generalist LLMs vs. specialized translation systems: Roughly competitive[30]
- Generalist LLMs vs. specialized code models: Generalists competitive or superior on many coding tasks

Threshold: Generalist within 20% of specialist performance across most tested domains.

Assessment: Varies by domain. Generalists competitive in language tasks; specialists still dominate in narrow domains like chess.

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

[29] Legg and Hutter 2007, p. 3.

[30] Various comparisons show frontier LLMs competitive with specialized NMT in many language pairs.

7.4 No Catastrophic Forgetting on New Tasks

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

Reference values:

- Within training: Modern training techniques manage multi-task learning
- Post-training fine-tuning: Risks catastrophic forgetting[31]
- Tool use: Allows capability extension without weight changes

Threshold: New capabilities can be added without significant degradation.

Assessment: Tool use provides graceful extension. Fine-tuning remains risky. True continual learning without forgetting is unsolved.

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

[31] 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 Formalization Benchmark

Criterion	Subcriterion	Score
1. Goal-Achievement	1.1 Reward maximization in training	100%
	1.2 Instruction-following	100%
	1.3 Autonomous goal pursuit	50%
	Criterion average	**83%**
2. Wide Environmental Range	2.1 Task-type diversity	100%
	2.2 Environmental diversity	50%
	2.3 Robustness to distribution shift	50%
	Criterion average	**67%**
3. Learning and Adaptation	3.1 In-context learning	100%
	3.2 Skill acquisition efficiency	50%
	3.3 Cross-session learning	0%
	Criterion average	**50%**
4. Simplicity Handling	4.1 Simple pattern recognition	100%
	4.2 Preference for simple hypotheses	50%
	4.3 Resistance to overfitting	50%
	Criterion average	**67%**
5. Generality Over Specialization	5.1 Generalist architecture	100%
	5.2 Performance breadth vs. depth	50%
	5.3 No catastrophic forgetting	50%
	Criterion average	**67%**
Overall Formalization Benchmark Score		**67%**

9 Interpretation

9.1 What Frontier AI Clearly Achieves (100%)

- Goal-achievement through training (reward/loss optimization)
- Instruction-following and task completion
- Broad task-type diversity (hundreds of cognitive task types)
- In-context learning (few-shot adaptation)
- Simple pattern recognition (elementary reasoning)
- Generalist architecture (single model, many tasks)

9.2 What Remains Contested (50%)

- Autonomous extended goal pursuit (multi-step agentic tasks)
- Environmental diversity beyond text/image/tool-use
- Robustness to significant distribution shift
- Skill acquisition on genuinely novel task types (ARC-AGI-2)
- Consistent Occam-like hypothesis preference
- Resistance to overfitting/memorization
- Breadth vs. depth trade-offs
- Capability extension without forgetting

9.3 What Is Clearly Not Achieved (0%)

- True cross-session learning (online weight updates from interaction)

This single 0% score is notable. Legg and Hutter's framework emphasizes that intelligence is about *learning*—adapting from experience over time. Current LLMs are frozen after training; they cannot update their weights from deployment-time interactions. Memory features provide continuity but not learning in the sense the definition requires.

10 The Verdict (Provisional)

The Legg-Hutter definition describes intelligence as goal-achievement across environments, with learning, simplicity-preference, and generality as key properties. At 67%, current frontier AI exhibits many of these properties—but with significant gaps.

The strongest match is architectural: frontier LLMs are genuine generalists that handle diverse tasks via a single model. This stands in stark contrast to the narrow AI of the Deep Blue era that motivated the definition.

The weakest match is learning: the definition's emphasis on dynamic adaptation and learning from experience is only partially met. LLMs learn within context but not across sessions. They cannot update weights from user interaction. In the agent-environment framework Legg and Hutter specify, this is a fundamental limitation.

10.1 Comparison with Earlier Benchmarks

Benchmark	**Year**	**Score**
Gubrud	1997	66%
Reinvention (Legg/Goertzel/Voss)	2002	80%
Formalization (Legg & Hutter)	2007	67%

The Formalization benchmark yields a similar score to Gubrud (67% vs. 66%) despite being more rigorous. The lower score compared to the 2002 Reinvention benchmark (80%) reflects the Formalization's stricter requirements for learning and adaptation.

10.2 The Incomputability Caveat

We have assessed current AI against the *properties* the Legg-Hutter definition implies, not against the formal measure itself. The formal measure $\Upsilon(\pi)$ cannot be computed. Whether our operationalization captures the definition's intent is itself contestable.

Legg and Hutter acknowledged this limitation: "In order to use universal intelligence more generally we will need to construct a workable test that approximates an agent's Υ value."[32] That test was never built. Our operationalization is one attempt; others might weight the criteria differently.

We do not speak for the authors. Shane Legg is alive and actively working on AGI at DeepMind. His 2023 "Levels of AGI" paper with colleagues represents his current thinking on operationalizing progress—and suggests he too has moved toward more pragmatic, less formally pure approaches to measurement.[33]

[32] Legg and Hutter 2007, p. 27.

[33] Morris et al. 2023, op. cit. https://arxiv.org/abs/2311.02462

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 Legg and Hutter's priorities that we cannot make with confidence. Did they consider "goal-achievement" more important than "learning"? Did they prioritize "generality" over "simplicity handling"? Their text emphasizes all of these properties without ranking them. We could guess at weights, but we would rather be honestly approximate than precisely wrong.

The operationalization problem. The Legg-Hutter definition is mathematically precise but incomputable. We have extracted five properties that the definition implies an intelligent agent should exhibit. This extraction involves interpretation. Different readers might identify different properties, or operationalize the same properties differently. Why five criteria rather than four or six? Why these subcriteria rather than others? These choices are defensible but not uniquely correct.

The formalization gap. There is an irony in assessing a formal definition via informal operationalization. The whole point of Legg and Hutter's project was to move beyond informal definitions. Our assessment necessarily steps back from that rigor. We are asking whether current systems exhibit the *spirit* of the definition, knowing we cannot test the *letter*.

The goal is accuracy at the expense of precision. This is a roughly hewn outline of a model. 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.

12 Citation Gaps and Requests for Collaboration

The following claims would benefit from stronger sourcing:

- Systematic benchmarks for Occam-like hypothesis preference in LLMs
- Quantified distribution shift degradation across frontier models
- Formal comparison of generalist vs. specialist performance across domains
- Systematic study of catastrophic forgetting in LLM fine-tuning
- Human baseline data on skill acquisition efficiency (examples needed for novel tasks)
- Rigorous agentic task completion benchmarks with standardized scoring

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 Legg-Hutter 2007 formalization. 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. Goal-Achievement	1.1 Reward maximization	□	□	□
	1.2 Instruction-following	□	□	□
	1.3 Autonomous goal pursuit	□	□	□
2. Wide Env. Range	2.1 Task-type diversity	□	□	□
	2.2 Environmental diversity	□	□	□
	2.3 Robustness to dist. shift	□	□	□
3. Learning & Adapt.	3.1 In-context learning	□	□	□
	3.2 Skill acquisition efficiency	□	□	□
	3.3 Cross-session learning	□	□	□
4. Simplicity Handling	4.1 Simple pattern recognition	□	□	□
	4.2 Preference for simple hyp.	□	□	□
	4.3 Resistance to overfitting	□	□	□
5. Generality	5.1 Generalist architecture	□	□	□
	5.2 Breadth vs. depth	□	□	□
	5.3 No catastrophic forgetting	□	□	□

Criterion Averages:

1. Goal-Achievement: ____________________
2. Wide Environmental Range: ____________________
3. Learning and Adaptation: ____________________
4. Simplicity Handling: ____________________
5. Generality Over Specialization: ____________________

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: