Retrospective Benchmarks for Machine Intelligence

Evaluating Current AI Against Historical Specifications

Chapter 5: The Critique Benchmark (2019)

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 2019 case presents a unique challenge: François Chollet's definition was explicitly designed as a *critique* of task-performance benchmarking—the very methodology this series employs. His framework argues that measuring skill at any task, however impressive, falls short of measuring intelligence. Intelligence, he claims, is about how efficiently you acquire skills, not what skills you have.

This creates a productive tension. We are using a task-performance-based methodology to evaluate a definition that rejects task-performance as the measure. Where possible, we operationalize Chollet's framework on its own terms. Where we cannot avoid task-performance metrics, we acknowledge the circularity.

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 Keras Creator's Counterargument

In November 2019, François Chollet published a 62-page paper that would become one of the most influential critiques of the AI benchmarking paradigm.[2] The timing was significant. GPT-2 had been released earlier that year, demonstrating that language models could generate coherent text at unprecedented scales. The AI community was beginning to coalesce around a hypothesis: scale up the models, scale up the data, and intelligence would emerge.

Chollet disagreed. Profoundly.

He was not a newcomer to deep learning. In 2015, he had created Keras, an open-source neural network library that would become one of the most widely used frameworks in the field, with over 2.5 million developers.[3] He worked at Google, surrounded by researchers pushing the boundaries of what deep learning could do. He understood the technology from the inside. And he thought the field was confusing skill with intelligence.

"Skill is heavily modulated by prior knowledge and experience," Chollet wrote. "Unlimited priors or unlimited training data allow experimenters to 'buy' arbitrary levels of skills for a system, in a way that masks the system's own generalization power."[4] A chess engine trained on millions of games might beat any human, but it cannot write a poem. A language model trained on the entire internet might generate fluent text, but show it a novel puzzle unlike anything in its training data, and it flounders.

The distinction Chollet drew was between *crystallized intelligence*—accumulated skills and knowledge—and *fluid intelligence*—the ability to adapt to genuinely new situations. Modern AI systems, he argued, were achieving impressive levels of crystallized intelligence while lacking fluid intelligence almost entirely.

To demonstrate this, he created a benchmark: the Abstraction and Reasoning Corpus (ARC). Each task in ARC is a visual puzzle: a grid of colored squares showing a few input-output examples, and a test input for which the system must produce the correct output. The puzzles are designed to be trivially easy for humans—most can be solved in seconds—but require genuine abstraction and reasoning rather than pattern matching against memorized solutions.

When Chollet released ARC, GPT-3 scored 0%. GPT-4 scored near 0%. GPT-4o, in 2024, reached 5%. Despite a roughly 50,000× scale-up in model parameters, performance on this test of fluid intelligence barely budged from zero.[5]

Then, in December 2024, something changed. OpenAI's o3 model—using extended reasoning and test-time computation—scored 87.5% on ARC-AGI-1 at high compute levels.[6] The benchmark that had seemed impervious to scaling suddenly fell.

Chollet's response was measured. He acknowledged the breakthrough but noted that o3 used massive computational resources—roughly $1,000 per task at the high-compute setting.[7] Humans solve the same tasks in seconds, for pennies worth of metabolic energy. He released ARC-AGI-2, a harder version where pure LLMs score 0% and even reasoning systems achieve only single-digit percentages, while humans still solve every task.[8]

The question remains: by Chollet's own definition, does current AI exhibit intelligence?

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

[3] Wikipedia, "Keras." https://en.wikipedia.org/wiki/Keras

[4] Chollet 2019, p. 2.

[5] Chollet, François. Talk at AI Startup School, San Francisco, June 2025. Transcript at https://singjupost.com/francois-chollet-how-we-get-to-agi-transcript/

[6] ARC Prize Foundation. "OpenAI o3 Breakthrough High Score on ARC-AGI-Pub." December 2024. https://arcprize.org/blog/oai-o3-pub-breakthrough

[7] Ibid.

[8] ARC Prize Foundation. "Announcing ARC-AGI-2 and ARC Prize 2025." March 2025. https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025

2 The Original Definition

From "On the Measure of Intelligence," published November 2019:[9]

> *The intelligence of a system is a measure of its skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty.*

The informal version: "Intelligence is the rate at which a learner turns its experience and priors into new skills at valuable tasks that involve uncertainty and adaptation."[10]

The formal measure, expressed in terms of algorithmic information theory:[11]

$$I^{\Theta}_{IS,\text{scope}} = \mathbb{E}_{T\sim\text{scope}}\left[\omega_T \cdot \frac{GD^{\Theta}_{T,C} \cdot \text{Skill}^{\Theta}_{IS,T,C}}{P^{\Theta}_{IS} + E^{\Theta}_{IS,T,C}}\right]$$

Where:

- I is the intelligence measure
- IS is the intelligent system being evaluated
- scope is the set of tasks considered
- T is a specific task
- ω_T is the value weight of task T
- $GD^{\Theta}_{T,C}$ is the generalization difficulty of T under curriculum C
- Skill is the skill level achieved
- P^{Θ}_{IS} measures the priors the system brings
- $E^{\Theta}_{IS,T,C}$ measures the experience (training data) used

2.1 Context

Chollet's paper synthesized two traditions: psychometrics (the measurement of human intelligence) and algorithmic information theory (the mathematical study of complexity and compression). From psychometrics, he borrowed the distinction between abilities and skills, and the emphasis on measuring broad cognitive capacities rather than narrow task performance. From algorithmic information theory, he borrowed tools for formalizing concepts like "simplicity" and "generalization difficulty."

The key insight is in the denominator: intelligence is *inversely proportional* to priors and experience. A system that achieves high skill using vast amounts of training data and elaborate built-in knowledge is *less* intelligent, by this measure, than a system that achieves the same skill with fewer resources. The numerator includes generalization difficulty: solving a genuinely novel problem demonstrates more intelligence than solving a problem similar to training examples.

This inverts the usual AI benchmarking logic. Most benchmarks ask: "How well does the system perform?" Chollet asks: "How efficiently did the system acquire that performance?"

2.2 Core Knowledge Priors

Chollet grounds his definition in developmental psychology's concept of "core knowledge"—the innate cognitive capacities that human infants bring to learning.[12] These include:

- **Object permanence and cohesion** — Objects persist and move as wholes

[9] Chollet 2019, pp. 27–28.

[10] ARC Prize Foundation. "What is ARC-AGI?" https://arcprize.org/arc-agi

[11] Chollet 2019, p. 28.

[12] Spelke, Elizabeth S., and Katherine D. Kinzler. "Core knowledge." *Developmental Science* 10, no. 1 (2007): 89–96. Cited in Chollet 2019, pp. 37–39.

- **Numerosity** — Basic counting and quantity comparison
- **Elementary geometry** — Spatial relationships and transformations
- **Agenthood** — Some objects act with goals and intentions

ARC tasks are designed to require only these core knowledge priors, making them fair tests for comparing human and machine intelligence. A system with vastly more prior knowledge than a human—such as an LLM trained on the entire internet—should, by Chollet's framework, have its performance discounted accordingly.

2.3 Operationalization

From Chollet's definition and its operationalization in ARC, we extract five criteria:

1. **Skill-acquisition efficiency** — High skill achieved with minimal experience
2. **Generalization to novel tasks** — Performance on tasks unlike training data
3. **Prior-efficiency** — Achieving results with minimal built-in knowledge
4. **Sample efficiency** — Learning from few examples
5. **Resource efficiency** — Achieving results with reasonable computational cost

The last criterion—resource efficiency—was added to ARC-AGI-2, reflecting Chollet's view that intelligence must be measured relative to the resources consumed. Brute-force search can eventually solve any computable problem; intelligence is about finding solutions efficiently.

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

3 Criterion 1: Skill-Acquisition Efficiency

3.1 What Chollet Meant

The core of Chollet's definition: intelligence is not about what skills you have, but how efficiently you acquire them. "If intelligence lies in the process of acquiring skills, then there is no task X such that skill at X demonstrates intelligence, unless X is actually a meta-task involving skill-acquisition across a broad range of tasks."[13]

A system that achieves 90% accuracy on a benchmark after training on millions of examples is less intelligent than one that achieves 80% accuracy after training on hundreds—if both start from comparable priors.

3.2 Performance vs. Examples Required

Measure: Skill level achieved per unit of task-specific experience.

Reference values:

- Humans on ARC-AGI-1: 73–85% accuracy with typically 3 training examples per task[14]
- Humans on ARC-AGI-2: 60% average accuracy; 100% of tasks solved by at least 2 humans in under 2 attempts[15]
- LLMs (zero-shot): Near 0% on ARC-AGI-1; 0% on ARC-AGI-2
- LLMs (few-shot): Marginal improvement; GPT-4.5 reaches ∼10% on ARC-AGI-1[16]
- Reasoning systems (o3 high-compute): 87.5% on ARC-AGI-1[17]

Threshold: Human-comparable skill from human-comparable experience (3–5 examples per task).

Assessment: Pure LLMs achieve near-zero skill despite seeing 3–5 examples—far below human skill-acquisition efficiency. Reasoning systems achieve high skill but require extensive computation (see Criterion 5). The skill/experience ratio for current AI is orders of magnitude below human levels on tasks designed to test this specifically.

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

Caveats: This assessment applies to ARC-type tasks specifically. On tasks within training distribution, LLMs show impressive few-shot learning. Chollet would argue this reflects memorization of similar patterns rather than genuine skill acquisition.

3.3 In-Context Learning Efficiency

Measure: Performance improvement from in-context examples on novel task types.

Reference values:

- GPT-3 (2020): Demonstrated significant few-shot learning on tasks similar to training[18]
- Frontier LLMs (2025): Strong few-shot performance on in-distribution tasks
- Novel task types (ARC): Minimal improvement from examples

[13] Chollet 2019, p. 22.

[14] Johnson, Aaditya, et al. "Testing ARC on Humans: A Large-Scale Assessment." NYU, 2024. https://lab42.global/arc-agi-benchmark-human-study/

[15] ARC Prize Foundation. "ARC-AGI-2." https://arcprize.org/arc-agi/2/

[16] Chollet 2025 talk, op. cit.

[17] ARC Prize o3 announcement, op. cit.

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

Threshold: Measurable improvement from examples on tasks outside training distribution.

Assessment: In-context learning works well for tasks similar to training data. For genuinely novel tasks, examples provide minimal benefit. This is exactly what Chollet's framework predicts: LLMs retrieve and apply memorized programs rather than synthesizing new ones.

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

3.4 Transfer Learning Efficiency

Measure: Ability to apply knowledge from one domain to novel problems in another.

Reference values:

- Webb et al. (2023): LLMs show some analogical reasoning comparable to humans[19]
- Chollet's critique: Such transfer often reflects surface similarity to training data, not genuine abstraction[20]

Threshold: Consistent transfer to tasks with high generalization difficulty (developer-aware novelty).

Assessment: Transfer exists but is inconsistent and difficult to disentangle from training data coverage. Chollet argues that apparent transfer often reflects pattern matching to similar training examples rather than genuine abstraction.

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

[19] 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

[20] Chollet, Dwarkesh Podcast interview, June 2024. https://www.dwarkesh.com/p/francois-chollet

4 Criterion 2: Generalization to Novel Tasks

4.1 What Chollet Meant

Chollet distinguishes between “system-centric” and “developer-aware” generalization difficulty. System-centric novelty asks whether the task differs from what the system has seen. Developer-aware novelty asks whether the task differs from what the system *or its developers* anticipated during design and training.[21]

The latter is crucial. If developers can anticipate a class of tasks and include similar examples in training, then success on those tasks demonstrates skill, not intelligence. True generalization requires handling tasks that neither the system nor its creators prepared for.

4.2 ARC-AGI-1 Performance

Measure: Performance on the benchmark explicitly designed to test developer-aware generalization.

Reference values:

- Humans: 73–85% (varying by study and time limits)
- GPT-3 (2020): 0%
- GPT-4 (2023): ∼0%
- GPT-4o (2024): ∼5%
- Best AI, late 2024 (Kaggle competition): ∼55% on private set
- o3 (high compute): 87.5% on semi-private set
- o3 (low compute): 75.7% on semi-private set

Threshold: ≥75% (human average level).

Assessment: o3 has crossed the human-average threshold on ARC-AGI-1, though at significant computational cost. The 75.7% low-compute score meets the threshold. However, ARC-AGI-1 is now considered saturating.

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

Caveats: o3 was trained on 75% of the ARC-AGI-1 public training set. The degree to which its performance reflects genuine generalization versus sophisticated pattern matching remains debated. Chollet has described o3 as achieving “task adaptation ability never seen before” but has not declared the problem solved.[22]

4.3 ARC-AGI-2 Performance

Measure: Performance on the harder benchmark designed to remain challenging for reasoning systems.

Reference values:

- Humans: 60% average; 100% of tasks solved by at least 2 humans in under 2 attempts
- Pure LLMs: 0%
- AI reasoning systems (early 2025): Single-digit percentages
- Claude Opus 4.5 (Thinking, 64k): 37.6% at $2.20/task[23]

[21] Chollet 2019, pp. 24–26.

[22] ARC Prize o3 announcement, op. cit.

[23] ARC Prize Foundation. “ARC Prize 2025 Results and Analysis.” December 2025. https://arcprize.org/blog/arc-prize-2025-results-analysis

- Best commercial system (Gemini 3 Deep Think): 45% at $77/task
- Best refinement solution (Poetiq + Gemini 3 Pro): 54% at $30/task[24]
- NVARC (Kaggle 1st place): 24% on private set at $0.20/task[25]

Threshold: ≥60% (human average level).

Assessment: No system has reached human-average performance on ARC-AGI-2. The best result (54%) approaches but does not meet the threshold, and requires $30 per task versus human costs of $2–17 per task.

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

4.4 Robustness to Problem Reformulation

Measure: Does performance degrade when problems are rephrased or presented differently?

Reference values:

- Documented brittleness to surface changes in LLMs[26]
- ARC-AGI-2 specifically designed to resist "gaming" through surface-level heuristics
- Pure LLMs scoring 0% suggests complete brittleness to this task format

Threshold: Less than 20% performance degradation under surface reformulation.

Assessment: The gap between LLM performance on standard benchmarks (>85% MMLU) and ARC (near 0%) suggests extreme brittleness when tasks differ from training patterns.

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

[24] Poetiq. "Poetiq Shatters ARC-AGI-2 State of the Art." December 2025. https://poetiq.ai/posts/arcagi_verified/

[25] NVIDIA. "NVIDIA Kaggle Grandmasters Win Artificial General Intelligence Competition." December 2025. https://developer.nvidia.com/blog/nvidia-kaggle-grandmasters-win-artificial-general-intelligence-competition/

[26] McCoy, Tom, et al. "Right for the Wrong Reasons." ACL 2019. https://aclanthology.org/P19-1334/

5 Criterion 3: Prior-Efficiency

5.1 What Chollet Meant

Intelligence is inversely proportional to priors. A system with extensive built-in knowledge that solves a problem is less intelligent than one with minimal priors that solves the same problem. This is why Chollet grounds ARC in "core knowledge" priors that humans and AI can share—basic concepts of objects, space, numbers, and agency.

"If an AI system has access to extensive, task-specific prior knowledge that is not available to a human, its performance on that task becomes a measure of the developer's cleverness in encoding that knowledge, not the AI's inherent intelligence."[27]

5.2 Training Data Scale

Measure: Size of training corpus relative to task-relevant human experience.

Reference values:

- Human lifetime language exposure: ∼1 billion words[28]
- GPT-4 training data: Estimated trillions of tokens
- LLM training includes: Vast amounts of reasoning examples, math problems, code, academic papers
- ARC training set: 400 tasks with 3–5 examples each

Threshold: Performance achieved with human-comparable prior exposure.

Assessment: LLMs have ingested orders of magnitude more task-relevant experience than any human. By Chollet's framework, their performance must be heavily discounted. The denominator of his intelligence equation ($P + E$) is enormous for LLMs.

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

5.3 Architectural Priors

Measure: Built-in structure that encodes task-relevant knowledge.

Reference values:

- Human brain: Innate core knowledge (object permanence, numerosity, etc.)
- LLM architecture: Attention, positional encoding—general-purpose, not task-specific
- Reasoning systems: Chain-of-thought, search algorithms—encode meta-level reasoning strategies

Threshold: Architectural priors comparable to human core knowledge (no task-specific encoding).

Assessment: LLM base architectures are relatively prior-free. However, training on vast data effectively encodes extensive priors in the weights. Reasoning systems add meta-level priors for search and deliberation. The overall prior load far exceeds human core knowledge.

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

[27] ARC Prize Foundation. "What is ARC-AGI?" op. cit.

[28] Hart, Betty, and Todd R. Risley. "The Early Catastrophe: The 30 Million Word Gap by Age 3." *American Educator* 27, no. 1 (2003): 4–9.

6 Criterion 4: Sample Efficiency

6.1 What Chollet Meant

Sample efficiency is closely related to skill-acquisition efficiency but focuses specifically on the number of examples required. A truly intelligent system should learn from minimal examples—ideally one or two demonstrations, as humans often do.

"Humans can often learn genuinely new skills from 1–3 examples plus explanation."[29]

6.2 Examples Required for Novel Tasks

Measure: Number of examples needed to learn a genuinely new task type.

Reference values (ARC):

- Humans: Typically 3–5 examples per task (provided in the task itself)
- LLMs: Fail to learn from provided examples (0% on ARC-AGI-2)
- Reasoning systems: Some success but require extensive computation

Reference values (other domains):

- LLMs on in-distribution tasks: Effective few-shot learning
- LLMs on novel formats: May require dozens of examples or fine-tuning
- Humans on novel concepts: Often 1–3 examples with explanation

Threshold: Learn novel task types from ≤ 5 examples.

Assessment: On tasks designed to test this specifically (ARC), LLMs fail entirely. On in-distribution tasks, few-shot learning works. The distinction is precisely what Chollet's framework predicts: efficiency on novel tasks, not familiar ones, measures intelligence.

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

6.3 One-Shot vs. Many-Shot Improvement

Measure: Performance gain from each additional example.

Reference values:

- LLMs on familiar tasks: Substantial zero-to-few-shot gains
- LLMs on novel tasks (ARC): Minimal gain from examples
- Humans on ARC: Can typically solve from given examples

Threshold: Measurable improvement from each example on novel tasks.

Assessment: On ARC-type tasks, LLMs show no measurable improvement from the provided examples. This is the crux of Chollet's critique: the examples are useless because the system cannot synthesize new programs from them.

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

[29] Chollet, Dwarkesh interview, op. cit.

7 Criterion 5: Resource Efficiency

7.1 What Chollet Meant

ARC-AGI-2 introduced efficiency as an explicit metric. "Intelligence is not solely defined by the ability to solve problems or achieve high scores. The efficiency with which those capabilities are acquired and deployed is a crucial, defining component."[30]

Brute-force search can eventually solve any computable problem. That is not intelligence. Intelligence is finding solutions efficiently—with minimal compute, time, and energy.

7.2 Computational Cost per Task

Measure: Cost (compute, time, money) to solve tasks.

Reference values (ARC-AGI-2):

- Humans: $2–17 per task (study conditions; theoretical limit perhaps $2–5)[31]
- Claude Opus 4.5 (Thinking): 37.6% at $2.20/task
- Gemini 3 Deep Think: 45% at $77/task
- Poetiq + Gemini 3 Pro: 54% at $30/task
- NVARC (Kaggle winner): 24% at $0.20/task

Reference values (ARC-AGI-1):

- o3 (high compute): 87.5% at estimated $1,000+ per task
- o3 (low compute): 75.7% at <$10,000 total

Threshold: Human-comparable cost ($2–20 per task) at human-comparable accuracy.

Assessment: On ARC-AGI-1, o3 meets the accuracy threshold but at costs orders of magnitude above human levels. On ARC-AGI-2, Claude Opus 4.5 achieves reasonable cost ($2.20/task) but at 37.6% accuracy, well below human average (60%). The Pareto frontier has not reached human-level cost-performance.

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

7.3 Scaling Behavior

Measure: Does performance improve efficiently with additional resources?

Reference values:

- o3 on ARC-AGI-1: 75.7% → 87.5% with 172× more compute[32]
- AI systems on ARC-AGI-2: Log-linear scaling insufficient to beat benchmark[33]
- Human scaling: Minimal—humans solve tasks quickly or not at all

Threshold: Sublinear compute scaling (diminishing returns at human-level performance).

Assessment: Current systems show that performance can be "bought" with compute, but at steep marginal costs. This is exactly what Chollet predicted: compute can substitute for intelligence, but inefficiently.

[30] ARC Prize Foundation. "ARC-AGI-2," op. cit.
[31] ARC Prize announcement, op. cit.
[32] ARC Prize o3 announcement, op. cit.
[33] ARC Prize announcement, op. cit.

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

8 Summary: The Critique Benchmark

Criterion	**Subcriterion**	**Score**
1. Skill-Acquisition Efficiency	1.1 Performance vs. examples required	0%
	1.2 In-context learning efficiency	50%
	1.3 Transfer learning efficiency	50%
	Criterion average	**33%**
2. Generalization to Novel Tasks	2.1 ARC-AGI-1 performance	100%
	2.2 ARC-AGI-2 performance	50%
	2.3 Robustness to reformulation	0%
	Criterion average	**50%**
3. Prior-Efficiency	3.1 Training data scale	0%
	3.2 Architectural priors	50%
	Criterion average	**25%**
4. Sample Efficiency	4.1 Examples for novel tasks	0%
	4.2 One-shot vs. many-shot improvement	0%
	Criterion average	**0%**
5. Resource Efficiency	5.1 Computational cost per task	50%
	5.2 Scaling behavior	50%
	Criterion average	**50%**
Overall Critique Benchmark Score		**32%**

9 Interpretation

9.1 What Frontier AI Clearly Achieves (100%)

- Human-level performance on ARC-AGI-1 (the original benchmark)

Only one subcriterion scores 100%. This is by far the lowest proportion of any benchmark in this series. Chollet's definition was explicitly designed to capture what current AI lacks.

9.2 What Remains Contested (50%)

- In-context learning on tasks with some training distribution similarity
- Transfer learning (present but inconsistent)
- Architectural prior-efficiency (base architectures are general-purpose)
- ARC-AGI-2 performance (approaching but not reaching human average)
- Computational cost (some efficient solutions, but accuracy tradeoffs)
- Scaling behavior (compute can buy performance, but inefficiently)

9.3 What Is Clearly Not Achieved (0%)

- Skill-acquisition efficiency on genuinely novel tasks (ARC)
- Training data efficiency (LLMs require orders of magnitude more data than humans)
- Sample efficiency on novel task types
- Improvement from examples on novel tasks
- Robustness to problem reformulation

The concentration of 0% scores is striking. This is not an artifact of harsh grading—it reflects Chollet's framework accurately. He designed a definition that would expose the limitations of skill-based AI systems, and current systems exhibit exactly those limitations.

10 The Verdict (Provisional)

Chollet's definition describes intelligence as skill-acquisition efficiency: the rate at which a system converts priors and experience into new skills on genuinely novel tasks. At 32%, current frontier AI falls far short of this definition.

10.1 The Fundamental Gap

Chollet's critique has proven remarkably prescient. The gap between LLM performance on standard benchmarks (>85% MMLU, >80% SWE-Bench) and performance on tasks designed to test genuine skill acquisition (∼0% ARC-AGI-2 for pure LLMs) is exactly what his framework predicts. Standard benchmarks measure crystallized intelligence—accumulated skills and knowledge. ARC measures fluid intelligence—the ability to acquire new skills on the fly. LLMs have the former in abundance; the latter remains elusive.

10.2 The Reasoning Revolution

Something has changed since 2019. The emergence of "AI reasoning systems"—models that perform extended computation at test time, searching through solution spaces and refining their answers—has shifted the landscape. o3's performance on ARC-AGI-1 would have seemed impossible to Chollet (and most observers) in 2019.

Chollet's interpretation: these systems achieve "test-time adaptation"—the ability to modify behavior dynamically based on specific data encountered during inference. This is different from static pattern matching and represents genuine progress. But it still requires massive compute, and ARC-AGI-2 remains largely unsolved.[34]

10.3 What Chollet Would Say

Unlike the other figures in this series, Chollet is not merely alive but vocally active in commenting on exactly these questions. His position is clear:

"Automation is not the same as intelligence... You can automate more and more things. Yes, this is economically valuable. Yes, potentially there are many jobs you could automate away like this. That would be economically valuable. You're still not going to have intelligence."[35]

"LLMs are a dead end to AGI."[36]

"OpenAI basically set back progress to AGI by five to 10 years."[37]

By his own definition, current AI is not intelligent. It can exhibit high skill on tasks similar to its training data. It can even, with extended reasoning, solve some tasks requiring abstraction. But it cannot efficiently acquire new skills on genuinely novel tasks—the core of what Chollet means by intelligence.

We do not claim Chollet is correct. His definition is one among many, and others in this series yield different verdicts. But on his own terms, applied fairly, current AI falls clearly short.

[34] Chollet 2025 talk, op. cit.

[35] Chollet, Dwarkesh interview, op. cit.

[36] Chollet, quoted in Big Think, August 2024. https://bigthink.com/the-future/arc-prize-agi/

[37] Ibid.

10.4 Comparison with Earlier Benchmarks

Benchmark	Year	Score
Gubrud	1997	66%
Reinvention (Legg/Goertzel/Voss)	2002	80%
Formalization (Legg & Hutter)	2007	67%
Corporatization (OpenAI Charter)	2018	52%
Critique (Chollet)	2019	32%

The Critique benchmark yields the lowest score of the five evaluated so far—by a substantial margin. This is not surprising: Chollet explicitly designed his framework to expose the limitations of current approaches. Earlier definitions asked whether AI could match human capabilities; Chollet asks whether AI can match human *efficiency* at acquiring capabilities. The answer, so far, is no.

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 Chollet's priorities that we cannot make with confidence. His formula treats priors and experience symmetrically in the denominator; we could argue skill-acquisition efficiency is the "core" criterion. But imposing weights would mean substituting our judgment for his.

The circularity problem. Chollet's definition rejects task-performance benchmarking. Yet we are using task-performance data (ARC scores) to evaluate systems against his definition. This is defensible because Chollet himself created ARC as an operationalization of his framework. But readers should note the tension: we are using the methodology Chollet critiques to evaluate his critique of that methodology.

The adversarial design. Unlike earlier definitions in this series, Chollet's was explicitly designed to show that current AI approaches are not intelligent. This is not a neutral specification that happens to be unmet; it is a critique that was constructed to demonstrate specific limitations. Our evaluation confirms those limitations—but this is less surprising given the adversarial framing.

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 comparison of LLM training data scale to human language exposure
- Rigorous quantification of "generalization difficulty" for various benchmarks
- Independent verification of ARC-AGI human baselines with larger samples
- Compute cost breakdowns for reasoning systems on ARC-AGI tasks
- Systematic study of few-shot learning on out-of-distribution task types
- Formal analysis of whether o3's ARC performance reflects program synthesis or sophisticated pattern matching
- Chollet's own scoring of current systems against his definition (if he has published one)

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 Chollet's 2019 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. Skill-Acquisition Efficiency	1.1 Performance vs. examples Human-comparable skill from 3–5 examples	□	□	□
	1.2 In-context learning efficiency Improvement from examples on novel tasks	□	□	□
	1.3 Transfer learning efficiency Consistent cross-domain transfer	□	□	□
2. Generalization to Novel Tasks	2.1 ARC-AGI-1 performance $\geq$75% (human average)	□	□	□
	2.2 ARC-AGI-2 performance $\geq$60% (human average)	□	□	□
	2.3 Robustness to reformulation $<$20% degradation on surface changes	□	□	□
3. Prior-Efficiency	3.1 Training data scale Human-comparable data exposure	□	□	□
	3.2 Architectural priors Core knowledge only, no task-specific	□	□	□
4. Sample Efficiency	4.1 Examples for novel tasks Learn from $\leq$5 examples	□	□	□
	4.2 One-shot improvement Measurable gain per example	□	□	□
5. Resource Efficiency	5.1 Cost per task Human-comparable ($2–20/task) at accuracy	□	□	□
	5.2 Scaling behavior Sublinear compute requirements	□	□	□

Criterion Averages:

1. Skill-Acquisition Efficiency: ______________________________
2. Generalization to Novel Tasks: ______________________________
3. Prior-Efficiency: ______________________________
4. Sample Efficiency: ______________________________
5. Resource Efficiency: ______________________________

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: