

Defining AGI: Mathematical and Scientific Benchmarks

Artificial General Intelligence (AGI) is often treated as a subjective milestone, but in the fields of computer science and cognitive mathematics, it is defined by specific, measurable criteria. To determine with absolute precision when AGI has been reached, we must look at the convergence of three distinct frameworks: Universal Intelligence, Capability Levels, and Skill-Acquisition Efficiency.

1. The Mathematical Foundation: Universal Intelligence

The most rigorous mathematical definition of AGI is derived from the **Legg-Hutter Universal Intelligence** measure. This framework moves away from specific tasks (like chess or coding) and focuses on an agent's performance across a space of all possible computable environments.

The Universal Intelligence Formula

The intelligence Υ (Upsilon) of an agent π is defined as the weighted sum of its performance across all environments:

Components of the Equation:

- **E (The Environment Space):** The set of all computable "reward-summable" environments.
- **V_{μ}^{π} (Value/Performance):** The expected total reward the agent π achieves in environment μ .
- **$K(\mu)$ (Kolmogorov Complexity):** The length of the shortest program that can describe environment μ .
- **$2^{-K(\mu)}$ (The Weighting):** This is based on **Occam's Razor**. It dictates that simpler, more fundamental logical environments are more significant indicators of intelligence than complex, "noisy" ones.

2. The Scientific Framework: Levels of AGI

In collaboration with the research community, organizations like Google DeepMind have established a taxonomy of AGI to provide a roadmap for "absolute precision" in tracking progress.

Level	Name	Performance Benchmark	Real-World Equivalent
Level 1	Emerging	Equal to or slightly better than an unskilled human.	GPT-4, Gemini 1.5, Claude 3.5.
Level 2	Competent	At least the 50th	The AGI Threshold. A

Level	Name	Performance Benchmark	Real-World Equivalent
		percentile of skilled adults on a broad range of tasks.	system capable of autonomous office work.
Level 3	Expert	At least the 90th percentile of skilled adults.	Outperforms most humans in law, medicine, and engineering.
Level 4	Virtuoso	At least the 99th percentile of skilled adults.	Nobel Prize level or Grandmaster performance.
Level 5	Superhuman	Outperforms 100% of humans .	ASI (Artificial Superintelligence).

3. Level 2: The "Economically Valuable Task" Benchmark

To move from "Emerging" (Level 1) to "Competent" (Level 2), a system must prove it can handle **Economically Valuable Tasks (EVTs)**. As of 2026, two primary benchmarks define this threshold with absolute precision:

A. The GDPval Benchmark

Released in late 2025, **GDPval** measures AI performance across 44 occupations that contribute significantly to global GDP. To pass Level 2, the AI must achieve a **>50% win rate** against skilled human professionals in the following domains:

- **Legal Research:** Drafting a 15-page legal memorandum that accurately cites case law and identifies procedural risks.
- **Financial Analysis:** Processing three years of raw SEC filings and bank statements to produce a 10-year valuation model for a mid-cap company.
- **Market Strategy:** Identifying 5 specific product-market fits for a new SaaS tool based on real-time web scraping and historical competitor data.
- **Software Deployment:** Not just writing code, but setting up the CI/CD pipeline, deploying to a cloud provider, and debugging a live latency issue.

B. The OfficeQA Benchmark

While typical LLMs answer questions, **OfficeQA** tests "Grounded Reasoning." A Level 2 system must be able to:

1. **Retrieve:** Search through 1,000+ unorganized company PDFs, spreadsheets, and emails.
2. **Aggregate:** Combine data from multiple sources (e.g., "What was our total expenditure on travel in Q3 vs Q2, accounting for the currency conversion in the London branch?").
3. **Produce:** Generate the final deliverable (spreadsheet or report) in a format ready for a Board of Directors.

4. The Efficiency Metric: ARC-AGI

AGI is defined scientifically by **Skill-Acquisition Efficiency**. If a system solves a problem because it saw the answer in its training data, it is not "intelligent"—it is merely "retrieving."

- **Human Baseline on ARC:** ~85%
- **Current SOTA (2026):** ~65%
- **AGI Benchmark:** When a model hits **>85%** on hidden ARC puzzles without being fine-tuned on the specific logic of those puzzles, it has achieved "Generalization."

5. Summary: When do we "know for sure"?

We have reached AGI when a single, non-specialized model can:

1. **Score >85% on ARC-AGI** (Proving reasoning, not just memory).
2. **Pass the GDPval Benchmark** at the 50th percentile (Proving economic utility).
3. **Demonstrate Long-Horizon Agency** (Performing 30+ steps of a project over several days without human error-correction).

As of March 2026, we are currently in the **"Jagged Transition"**—performing at Level 3 in code/math, but still Level 1 in long-term planning and high-stakes reasoning.