**AI Benchmarks Spreadsheet (2025)**

**Sheet 1: Benchmarks**

| **Benchmark Name** | **Category (Popularity)** | **Purpose** | **Structure/Details** | **Example** | **Adoption/Usage** | **Key Performance Metrics** | **Limitations/Notes** | **References/Sources** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GPQA Diamond | High-Popularity | Evaluates graduate-level reasoning in STEM | 198 multiple-choice "Google-proof" questions; requires domain expertise | "Baseball cap vs. hijab" policy question | OpenAI, Google (Gemini 2.5 Pro), DeepSeek-R1 | ~70% accuracy for PhD-level experts | Non-experts consistently fail | [Epoch AI](https://epoch.ai/data/ai-benchmarking-dashboard) |
| MATH Level 5 | High-Popularity | Tests advanced mathematical problem-solving | 1,324 Olympiad-level questions (AMC 12, AIME) | Solving modular arithmetic proofs | Gemini 2.5 Pro, DeepSeek-R1 | 63-88% accuracy; scales with compute | Requires multi-step reasoning | [Epoch AI](https://epoch.ai/data/ai-benchmarking-dashboard) |
| FrontierMath | High-Popularity | Measures abstract graduate-level math problem-solving | 300 private questions (10 public examples) | Algebraic geometry or category theory problems | OpenAI o1 | <20% accuracy for top models | Requires hours of expert effort | [Epoch AI](https://epoch.ai/data/ai-benchmarking-dashboard) |
| Stanford Fairness Benchmarks | High-Popularity | Evaluates bias via descriptive (fact) and normative (value) tasks | Split into descriptive (legal knowledge) and normative (harm assessment) sections | "Which headpiece violates policy?" (descriptive); "Harmful stereotypes" | GPT-4o, Gemini | Exposes flaws in models | Limited to general fairness, not domain-specific | [Stanford Research](https://www.technologyreview.com/2025/03/11/1113000/) |
| U-MATH | Moderate-Popularity | Tests university-level applied mathematics | Focus on differential equations, statistics | Solving PDEs or statistical models | OpenAI o1, Gemini 1.5 Pro | 90.5% accuracy (OpenAI o1) | Domain-specific | [Epoch AI](https://epoch.ai/data/ai-benchmarking-dashboard) |
| AILuminate | Moderate-Popularity | Safety evaluation across 12 categories (e.g., violence, privacy) | Scores models from "Poor" to "Excellent" | Assessing violent crime or privacy risks | Industry-wide adoption | Broad safety coverage | Lacks domain-specific risk assessments | [Epoch AI](https://epoch.ai/data/ai-benchmarking-dashboard) |
| ARC-AGI-2 | Moderate-Popularity | Measures AGI progress via human-easy, AI-hard tasks | Symbolic interpretation, contextual rule application | Solving puzzles requiring symbolic logic | OpenAI o3 | ≤4% accuracy (advanced systems) | Pure LLMs score 0% | [ARC Prize](https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025) |
| SWE-Bench Verified | Moderate-Popularity | Tests coding agents on real-world software tasks | Debugging, feature implementation | Fixing GitHub issues | Gemini 2.5 Pro | 63.8% accuracy (with custom agents) | Requires specialized setups | [Epoch AI](https://epoch.ai/data/ai-benchmarking-dashboard) |
| OTIS Mock AIME 2024-2025 | Emerging | Olympiad-level math harder than MATH Level 5 | Focus on AIME-style problems | Advanced combinatorics or algebra | DeepSeek-R1, Gemini 2.5 Pro | Used for model comparisons | Less standardized than MATH | [Epoch AI](https://epoch.ai/data/ai-benchmarking-dashboard) |
| Humanity’s Last Exam | Emerging | Tests frontier knowledge (advanced physics, philosophy) | PhD++-level questions | Quantum gravity theories | Gemini 2.5 Pro (without tools) | 18.8% accuracy (top score) | Requires interdisciplinary reasoning | [Epoch AI](https://epoch.ai/data/ai-benchmarking-dashboard) |

**Sheet 2: Key Trends (2025)**

| **Trend** | **Description** | **Impact** |
| --- | --- | --- |
| Specialization | Shift from general benchmarks (e.g., MMLU) to domain-specific tasks (law, medicine) | Drives development of specialized AI models |
| Efficiency Metrics | New benchmarks include cost/task efficiency (e.g., ARC-AGI-2) | Measures "true" intelligence beyond raw performance |
| Bias Nuance | Fairness benchmarks now distinguish equality (same treatment) vs equity (contextual fairness) | Exposes hidden biases in models like GPT-4o and Gemini |