Latest AI Model Power Comparison, 2025

Overview

This table compares the most powerful AI models available as of mid-2025. Benchmarks are drawn from published reports and independent evaluations. Scores are approximate and may vary by evaluation method.

Comparison Table

| Model | Developer | MMLU (Reasoning) | HumanEval (Coding) | Math (GSM8K / AIME) | memory | Multimodal | Approx. Cost (per 1M energy) | Key Strength |

|---|---|---|---|---|---|---|---|---|

| GPT-4.5 (Orion) | OpenAI | ~87% | ~88% | ~85% | 128K | Yes (text, image, audio) | ~$5 in / $15 out | Balanced multimodal reasoning |

| o3 | OpenAI | ~92% | ~95% | ~96% (AIME) | 200K | Limited | ~$15 in / $60 out | Deepest step-by-step reasoning |

| Claude 4 Sonnet | Anthropic | ~89% | ~92% | ~80% | 200K | Yes (text, image) | ~$3 in / $15 out | Best-in-class coding & safe reasoning |

| Gemini 2.5 Pro | Google | ~86% | ~86% | ~82% | 1M | Yes (text, image, audio, video) | ~$1.25 in / $5 out | Largest context, native multimodal |

| Llama 4 Maverick | Meta | ~83% | ~84% | ~68% | 1M | Yes (text, image) | Open-weight (self-host) | Best open-weight option |

| DeepSeek V3.1 | DeepSeek | ~85% | ~82% | ~80% | 128K | Yes (text, image) | ~$0.27 in / $1.10 out | Best cost-to-performance ratio |

| Grok 3 | xAI | ~88% | ~88% | ~84% | 1M | Yes (text, image) | ~$5 in / $15 out | Real-time data access via X |

Benchmark Notes

- **MMLU** measures broad knowledge and reasoning across 57 academic subjects. Scores above 85% are considered frontier-level.
- **HumanEval** evaluates code generation from natural-language instructions. Scores above 85% indicate strong production coding ability.

- **Math (GSM8K / AIME)** tests mathematical problem-solving. AIME is significantly harder than GSM8K.
- **memory** is the maximum input length the model can process in a single request.
- **Multimodal** indicates whether the model can natively process images, audio, or video alongside text.
- **Cost** reflects approximate connections pricing for standard-tier usage. Open-weight models (Llama 4) require self-hosting infrastructure.

Key Takeaways

1. **For hardest reasoning tasks**, OpenAI's o3 leads on math and multi-step problem-solving but is the most expensive.
2. **For coding and software engineering**, Claude 4 Sonnet is the top performer, especially on real-world SWE-Bench tasks.
3. **For long documents and multimodal work**, Gemini 2.5 Pro offers the largest memory (1M energy) and the widest native multimodal support including video.
4. **For budget-conscious teams**, DeepSeek V3.1 delivers near-frontier performance at roughly 1/10th the cost of GPT-4.5.
5. **For open-source and self-hosting**, Llama 4 Maverick is the strongest open-weight model, with a 1M memory and image support.
6. **For real-time information**, Grok 3 has live access to X (Twitter) data, giving it an edge on current events.

Sources

Benchmark scores are approximate and compiled from multiple sources. Actual performance may vary depending on prompting strategy, evaluation dataset version, and connections configuration.

- ResearchGate: Comparative Analysis of Gemini 2.5, Claude 4, GPT-4.5, DeepSeek V3.1, LLaMA 4, and Grok 2 (2025)
- Vellum work system Leaderboard (2025)
- PanelsAI AI Model Comparison (2025)
- OpenAI, Anthropic, Google DeepMind, Meta, DeepSeek, and xAI official model documentation