Benchmarking Top Agentic AI Deep Research Models

The Jagged Topographic Frontier

A Topology of Frontier Machine Capability, May 2026

Raymond Uzwyshyn, Ph.D., MBA, MLIS
Acting AUL for Research and Technology
Director of Research Services
University of California, Riverside

UCR Library AI Literacy Series
May 2026

Abstract

Between February and late April 2026, more than fifteen frontier large language models entered general availability within an eight-week window—among them Claude Opus 4.7 (April 16), Kimi K2.6 (April 20), GPT-5.5 (April 23), and DeepSeek V4 Pro (April 24). This report maps the resulting landscape across five independent benchmarks: GPQA Diamond for graduate scientific reasoning, Humanity's Last Exam for cross-disciplinary expert breadth, SWE-bench Verified for real software engineering, ARC-AGI-2 for fluid intelligence and novel generalization, and the Artificial Analysis Intelligence Index v4.0 for composite capability. The principal finding is structural rather than incremental: machine intelligence at the frontier is plural rather than scalar. No single model dominates across all dimensions, and the model leading composite capability is simultaneously the model most likely to fabricate when uncertain. Three structural gaps—abstraction, reliability, and agency—separate frontier machine capability from human expert performance, and have widened rather than closed across the May 2026 generation. A Chinese open-weights system ties the leading proprietary model on the hardest cross-disciplinary academic benchmark at approximately one-fifth the per-token cost. The report proposes a tiered orchestration architecture in which model selection is treated as a first-order research skill rather than a procurement decision.

Keywords: large language models; capability benchmarking; GPQA; HLE; ARC-AGI; SWE-bench; AA-Omniscience; research libraries; AI literacy; orchestration architecture.

1. The Terrain

The frontier of machine intelligence in May 2026 resists the leaderboard's flatness. It has dimension, depth, direction. Where one model towers in graduate scientific reasoning, another plunges in novel generalization. Where a third leads in raw factual recall, it falters most steeply on fidelity to fact. The map is not a ranking. It is a terrain.

This finding—first articulated in the present series' February 2026 essay on the same subject—has not weakened across the intervening generation of releases. It has deepened. In the eight weeks between late March and late April, more than fifteen serious frontier models arrived. Each carries a distinct cognitive signature. Each excels in some dimensions and recedes in others. Reading those contours has become the most actionable form of information literacy now available to researchers, educators, and students working at a research university.

Three architectural shifts produced this terrain. The first was the introduction of post-training reasoning, beginning with OpenAI's o-series and DeepSeek's R1 in mid-2025, which gave models a visible chain of thought and the ability to allocate compute to inference rather than training. The second was the maturation of multimodal capability from research demonstration to standard production feature: images, documents, audio, and structured data became default inputs rather than exceptions. The third was the emergence of genuine agentic systems—models capable of multi-step task execution with tool use, web access, and sustained operation across extended thinking windows, producing what are now routinely called deep research reports.

By early 2026 the field had outgrown its first generation of benchmarks. MMLU, the standard 57-discipline test, had reached saturation. Multiple models exceeded human expert performance. The community required harder evaluations and developed them: GPQA Diamond for graduate-level scientific reasoning that resists web search, Humanity's Last Exam for the cross-disciplinary edges of expert knowledge, ARC-AGI-2 for genuine fluid intelligence, and the Artificial Analysis Intelligence Index v4.0 to aggregate ten such measurements into a defensible composite. These are the instruments through which the present analysis proceeds.

What follows is a topographic survey. Section 2 describes the five benchmark instruments and their distinct cognitive valences. Section 3 profiles the seven frontier models that meaningfully define the May 2026 landscape. Section 4 identifies three structural gaps that

persist between frontier machine capability and human expert performance, and that have widened, not closed, across the latest generation. Section 5 proposes a tiered orchestration architecture for academic research workflows. Section 6 reflects on what this terrain implies for research practice. Appendix A documents the source-verification methodology applied throughout; Appendix B provides the underlying benchmark data.

2. Five Instruments, Five Cognitive Capacities

Five instruments measure five capacities. They are not interchangeable. To average their scores is to commit the analytical error of treating a blood-pressure cuff and an electrocardiogram as equivalent readings of cardiac health: both are cardiac; neither is a substitute for the other. Where multiple measurements of the same benchmark exist, this report prefers, in order: the model vendor's own model card; independent re-evaluation by Artificial Analysis under standardized conditions on first-party APIs; and aggregator leaderboards as cross-reference. Where vendor and Artificial Analysis figures diverge, both are noted.

GPQA Diamond: graduate scientific reasoning.

One hundred ninety-eight four-way multiple-choice questions at the frontier of graduate-level physics, chemistry, and biology, designed by domain experts to resist web search—to be, in the benchmark's official phrase, "Google-proof." Answers cannot be retrieved; they must be reasoned toward through multi-step inference. Human PhD experts score sixty-five to seventy-four percent within their fields; non-experts with thirty minutes of search access score approximately thirty-four percent. In May 2026, the top four frontier models—Gemini 3.1 Pro Preview at 94.3 percent, Claude Opus 4.7 at 94.2 percent, GPT-5.5 at 93.5 percent, and Kimi K2.6 at 90.5 percent—all exceed the human expert ceiling by twenty to thirty percentage points. Structured scientific reasoning, in any practical sense, is solved. Differentiation among models in this dimension has collapsed.

Humanity's Last Exam: cross-disciplinary expert breadth.

Approximately 2,500 expert-level questions, crowdsourced from nearly one thousand PhD-level contributors at more than five hundred institutions across fifty countries. The set spans more than one hundred disciplines—from advanced mathematics and theoretical physics to medieval manuscript analysis and sesamoid bone anatomy. Roughly fourteen percent of questions require multimodal interpretation. Reported scores depend critically on whether tools are permitted: the same model can vary by thirty percentage points across "with tools" and "without tools" configurations, a distinction this report flags where it materially affects

interpretation. When HLE launched in January 2025, frontier models scored in the single digits. In May 2026 Claude Opus 4.7 leads at 54.7 percent with tools, matched by Kimi K2.6 at 54.0 percent. Human experts still score approximately ninety percent within their own domains. The gap remains substantial. The trajectory is no longer in doubt.

SWE-bench Verified: real software engineering.

Five hundred real GitHub issues from active Python open-source repositories. Each model must read a natural-language bug description and produce a working patch that passes the project's test suite—approximating the daily diagnostic work of a mid-level software engineer. Claude Opus 4.7 leads at 87.6 percent. One methodological caveat applies: OpenAI's early-2026 internal audit found that every frontier model showed memorization on this benchmark—verbatim recall of gold patches—and the contamination-resistant SWE-bench Pro returns scores twenty to twenty-five percentage points lower for the same systems. Verified scores in this report should be interpreted as directional indicators rather than precise capability measurements.

ARC-AGI-2: fluid intelligence and novel generalization.

Novel grid-based visual puzzles requiring rule inference from two or three examples. No language. No domain knowledge. No possibility of memorization. The benchmark was designed by François Chollet and the ARC Prize Foundation expressly to resist scale and brute-force search—to measure genuine generalization rather than sophisticated pattern completion. Average human solve time is 2.3 minutes. GPT-5.5 leads frontier AI at 85 percent, a real architectural breakthrough relative to earlier generations. Gemini 3.1 Pro Preview follows at 77.1 percent. Claude Opus 4.7 trails near 63 percent. Several frontier systems do not report ARC-AGI-2 scores at all—an absence which, as later sections will argue, is itself diagnostic.

Artificial Analysis Intelligence Index v4.0: composite capability.

A rigorous independent aggregate of ten standardized evaluations run on first-party APIs: GDPval-AA, τ^2-Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, and CritPt. GPT-5.5 leads at 60 points. Claude Opus 4.7, Gemini 3.1 Pro Preview, and GPT-5.4 are tied at 57. Kimi K2.6 leads the open-weights field at 54. The Index is the most defensible single-number summary in current operation—and, as the seven profiles below make clear, single-number summaries miss most of what matters at the frontier.

3. Seven Profiles

3.1 GPT-5.5 — The Ambitious Formalist

OpenAI's April flagship occupies an unusual position in the topology. It knows more than any model yet released, and it fabricates more confidently than any model yet released. On the AA-Omniscience evaluation it records the highest factual accuracy in the public record at 57 percent, and simultaneously the highest hallucination rate at the frontier at 86 percent. On ARC-AGI-2 it reaches 85 percent—a genuine architectural breakthrough relative to the previous generation. On the composite Artificial Analysis Intelligence Index it leads at 60, three points clear of the next tier.

The pattern is not contradictory but diagnostic. GPT-5.5 has been trained to attempt answers where other models abstain; it succeeds more often when the answer is recoverable, and fails more often when it is not. For computational research, agentic workflow design, terminal automation, and abstraction-heavy reasoning, the model's capabilities are unmatched at the May 2026 frontier. For research in social sciences, history, medicine, or law—domains where a fabricated citation carries real consequences—its hallucination profile constitutes a structural reliability risk requiring systematic verification protocols.

A higher-tier variant, GPT-5.5 Pro, scores 57.2 percent on HLE but is priced at thirty dollars input and one hundred eighty dollars output per million tokens—six times the base model. The two should not be conflated in aggregator data, as they frequently are.

Best suited to: *computational research, agentic workflows, terminal automation, code generation, abstract reasoning tasks.*

3.2 Claude Opus 4.7 — The Reliable Polymath

Anthropic's April flagship was built for a different equilibrium. Where GPT-5.5 reaches for answers, Opus 4.7 reaches for honest abstention. It leads or co-leads three frontier evaluations: Humanity's Last Exam with tools at 54.7 percent, SWE-bench Verified at 87.6 percent, and the AA-Omniscience hallucination rate at 36 percent—the lowest at the frontier and a steep descent from Opus 4.6's 61 percent in a single version. On GPQA Diamond it scores 94.2 percent, within rounding of the leader. On the composite AA-Omniscience Index it places second at 26, behind Gemini 3.1 Pro Preview. On the Artificial Analysis Intelligence Index it ties for second at 57.

Anthropic's own documentation states that the model "correctly reports when data is missing instead of providing plausible-but-incorrect fallbacks." For citation-dependent scholarship—literature review, legal research, historical analysis, the careful labor of the humanities—this property has direct operational implications. The model produces fewer answers under genuine uncertainty, and the answers it does produce carry the lowest fabrication probability available at the frontier.

One operational caveat: the new Opus 4.7 tokenizer produces approximately 1.0 to 1.35 times the token count of Opus 4.6 for identical input text. At unchanged sticker pricing—five dollars input, twenty-five dollars output per million tokens—effective per-request cost can rise materially. High-volume institutional deployments should budget accordingly.

Best suited to: *humanities and social sciences, qualitative research, legal and historical analysis, literature synthesis, citation-critical work, any task where factual honesty is non-negotiable.*

3.3 Gemini 3.1 Pro Preview — The Analytical Breadth Champion

Google DeepMind's flagship, released February 19, 2026, occupies the strongest balanced position in the May 2026 landscape. On GPQA Diamond it leads independently re-evaluated scores at 94.3 percent. On MMLU-Pro it tops the field at 91 percent. On ARC-AGI-2 it posts 77.1 percent—the highest verified score among third-party-evaluated models and more

than double the predecessor Gemini 3 Pro. On the composite AA-Omniscience Index it leads at 33, the strongest reliability composite at the frontier.

The pricing is the structural achievement. At two dollars input and twelve dollars output per million tokens for prompts up to 200,000 context, Gemini 3.1 Pro Preview is approximately one-fifth the per-token cost of Claude Opus 4.7 at comparable performance on most STEM evaluations. The one-million-token context window—at the same price tier—makes it particularly suited to document-heavy synthesis tasks where context length matters more than premium-tier reasoning depth.

Best suited to: *STEM research, multimodal analysis, large-document synthesis, science education, cost-conscious academic workflows requiring balanced reliability.*

3.4 Kimi K2.6 — The Open Frontier Explorer

Moonshot AI's April release achieves something genuinely remarkable. On Humanity's Last Exam with tools—the hardest cross-disciplinary academic benchmark yet constructed—Kimi K2.6 scores 54.0 percent, statistically indistinguishable from Claude Opus 4.7's leading 54.7 percent. A model from a Beijing laboratory, released as open weights under a Modified MIT license, ties the world's most capable proprietary system on humanity's deepest test of cross-domain expert knowledge.

The system architecture: one trillion total parameters in a Mixture-of-Experts configuration, with thirty-two billion active per token. Agent swarms scale to three hundred parallel sub-agents executing across four thousand coordinated steps. First-party API pricing from Moonshot is approximately sixty cents input and four dollars output per million tokens; blended provider pricing ranges from roughly $1.15 to $2.15 per million tokens depending on host.

Kimi K2.6 does not publicly report ARC-AGI-2 scores. As with other knowledge-heavy systems from the same architectural lineage, this absence likely reflects vast coverage paired with limited novel generalization—a structural feature of current training trade-offs rather than a deficit, but a feature researchers should understand before relying on the model for tasks demanding inference from minimal data.

Best suited to: *specialized interdisciplinary domain research, literature synthesis at the edges of human knowledge, agentic long-horizon research tasks, self-hosted or privacy-constrained institutional deployments.*

3.5 DeepSeek V4 Pro — The Economic Disruptor

DeepSeek's April release represents one of the most aggressive cost-performance offerings currently available. On GPQA Diamond it scores 90.1 percent; on Humanity's Last Exam, 37.7 percent; on SWE-bench Verified, 80.6 percent; on LiveCodeBench, 93.5 percent—leading all frontier models for competitive programming. Open weights under MIT license, one-million-token context window. Standard pricing of $1.74 input and $3.48 output per million tokens (a 75 percent launch promotion at $0.435 and $0.87 expired on May 5, 2026). The independent NIST CAISI evaluation places V4 Pro approximately eight months behind leading United States frontier models on aggregate capability—narrowing from the twelve-month gap estimated for its predecessor.

The reliability profile carries a meaningful caveat. On the composite AA-Omniscience Index that jointly rewards correct answers and penalizes confident error, V4 Pro scores negative ten—the only frontier model in this comparison registering below zero, meaning the system produces more confidently incorrect answers than correct ones across the Omniscience question set. For high-volume document processing or computational research where downstream verification is structural, this is manageable. For citation-critical work, it is not.

Best suited to: *high-volume document processing, budget-constrained research infrastructure, computational research, competitive programming, self-hosted institutional deployments.*

3.6 Qwen 3.6-Max-Preview — The Alibaba Contender

Alibaba released three Qwen 3.6 variants during March and April 2026: Qwen3.6-Plus on March 30 as a hosted preview, Qwen3.6-Max-Preview on April 20 with closed weights, and Qwen3.6-27B on April 22 under Apache 2.0 open weights. The flagship Max-Preview posts approximately 88.8 percent on GPQA Diamond and an Artificial Analysis Intelligence Index of 52. The shift of the flagship variant to closed weights signals commercial maturation;

open-weights variants remain available for academic fine-tuning. The system is most operationally useful in Chinese-language research contexts and Alibaba Cloud enterprise environments.

Best suited to: *Chinese-language research, Alibaba Cloud enterprise integrations, STEM research at near-frontier performance.*

3.7 Grok 4.20 — The Vast Real-Time Archive

xAI's model entered public beta on February 17, 2026, with API access following on March 10. The system offers a two-million-token context window—the largest at the frontier—and unique integration with real-time data from the X social media platform. Pricing for the reasoning variant is two dollars input and six dollars output per million tokens.

The ARC-AGI-2 score of approximately 30 percent stands as the most striking internal disparity in the May 2026 landscape. Grok 4.20 reads the most and generalizes least flexibly. The distance between encyclopedic knowledge and adaptive reasoning is the model's defining cognitive trait—a useful diagnostic for any researcher evaluating whether breadth of training data is what their actual task requires.

Best suited to: *social media research, real-time discourse analysis, political science, competitive intelligence, contexts where access to the most current information is the primary requirement.*

4. Three Gaps

The May 2026 measurement set is more precise than any predecessor, but its principal diagnostic finding has not changed: machine intelligence is structurally plural rather than scalar. Three gaps separate frontier model capability from human expert performance. None has closed materially across the latest generation, and one—the reliability gap—has become the most consequential single finding of the period.

4.1 The Abstraction Gap

ARC-AGI-2 measures fluid intelligence: the capacity to infer rules from minimal examples and generalize them to novel cases. GPT-5.5 at 85 percent represents the strongest result yet observed and is a genuine architectural improvement over earlier generations. The interpretation, however, requires care. Adult humans solve these puzzles in 2.3 minutes on average. Children aged six to eight outperform frontier AI on child-friendly subcorpora of the same benchmark. Kimi K2.6 and DeepSeek V4 Pro—both among the most capable frontier systems on knowledge-heavy evaluations—do not report ARC-AGI-2 scores at all. The pattern is consistent: vast knowledge coverage paired with limited novel generalization. ARC-AGI-2 was specifically designed to resist scale. The resistance is holding. The cognitive engine beneath scientific discovery, artistic invention, and entrepreneurial pattern recognition remains, in any measurable sense, distinctly human.

4.2 The Hallucination-Reliability Gap

Accuracy and reliability are distinct properties, and the May 2026 benchmarks are the first to measure them as separable dimensions rather than a single composite. The AA-Omniscience evaluation distinguishes three quantities: how often a model is correct when it answers (accuracy), how often it fabricates when it does not know (hallucination rate), and a composite Omniscience Index that rewards correct answers, penalizes confident error, and applies no penalty for honest abstention. The three measurements produce three different rankings.

GPT-5.5's 86 percent hallucination rate—recorded simultaneously with the highest composite intelligence score at the frontier—crystallizes the central reliability finding of the May 2026 generation. The most capable model is simultaneously the least trustworthy narrator. The picture grows more nuanced when accuracy and abstention are both factored in: on the composite AA-Omniscience Index, Gemini 3.1 Pro Preview leads at 33, Claude Opus

4.7 follows at 26, GPT-5.5 trails at 20, and DeepSeek V4 Pro registers negative ten. The hierarchy holds practical force.

The operational translation is direct. A researcher posing GPT-5.5 a hundred factual questions will receive more correct answers than from any other frontier model—and confidently fabricated answers to most of the questions where the model lacks adequate training data. A researcher posing Claude Opus 4.7 the same hundred questions will receive a higher rate of "this information is not available" responses, and the answers given will carry the lowest fabrication probability at the frontier. Gemini 3.1 Pro Preview sits between the two and, on the composite measure that punishes confident error and rewards honest uncertainty, holds the top position.

For research domains where fabricated citations or invented statistics carry real consequences—medicine, law, qualitative social science, historical analysis, citation-dependent humanities scholarship—the conclusion is structural rather than stylistic. Systematic verification is no longer optional. It is a required component of the workflow.

4.3 The Agentic-Judgment Gap

No frontier model achieves even 85 percent on Terminal-Bench Hard, which evaluates real-world multi-step computing tasks in sandboxed environments. Terminal-Bench Hard is, moreover, among the simpler forms of real-world agency currently measurable: it does not test sustained project leadership over weeks or months, adaptive research design under genuine uncertainty, or ethical navigation of ambiguous stakeholder requirements. The capacities required for directing an actual research program—orchestrating a multi-month inquiry, navigating institutional politics, pivoting in response to a failed experiment—remain firmly outside the frontier's reach. This is the gap least amenable to architectural improvement on current trajectories, and the gap that most directly defines where human researchers add irreducible value in a research workflow.

5. An Orchestration Architecture for Academic Research

The May 2026 landscape supports a four-tier orchestration architecture for academic research workflows. The structuring principle is correspondence: match each cognitive task to the model whose specific spike in the capability matrix best fits the task's actual demands.

For frontier-difficulty questions requiring factual precision—literature review at the edge of known scholarship, cross-disciplinary synthesis, citation-critical analysis where fabrication carries real cost—Claude Opus 4.7 serves as primary for reliability, paired with Kimi K2.6 for depth of domain coverage at the cross-disciplinary edges, with human expert verification at the synthesis stage.

For STEM analysis, multimodal interpretation, and large-document synthesis, Gemini 3.1 Pro Preview offers the strongest combination of independently verified GPQA performance, one-million-token context window, and price-performance ratio among premium proprietary models.

For computational research, code generation, and agentic workflow design where abstraction depth matters more than factual fidelity, GPT-5.5 provides unmatched terminal capability, paired with structural human oversight for any factual claims that touch citation or quotation.

For budget-constrained high-volume processing—classification, entity extraction, initial screening of large corpora—DeepSeek V4 Flash at fourteen cents input and twenty-eight cents output per million tokens handles bulk work, escalating to V4 Pro or Kimi K2.6 for the frontier-difficulty subset.

The human researcher at the orchestration layer—selecting which model to consult for which task, deciding when to verify, synthesizing across complementary cognitive profiles—is not peripheral to this architecture. The human researcher is the architecture.

6. Closing: The Cartography of Cognition

The map of machine intelligence in May 2026 grows more detailed across each successive generation of releases. The map, however, is not a leaderboard. Three months of rapid development have confirmed what the February 2026 analysis first proposed: the frontier is not converging toward a single peak. It is differentiating into a topography of cognitive specializations.

Models that dominate graduate scientific reasoning do not dominate cross-disciplinary breadth. The model that leads the composite intelligence index fabricates most confidently when pressed beyond its knowledge. A Chinese open-source release from a laboratory most academics could not have named eighteen months ago now ties the world's most capable proprietary model on humanity's hardest academic benchmark. The model with the deepest encyclopedic knowledge generalizes least flexibly from minimal examples.

The practical implication for researchers at a research university is direct. The decision of which model to consult for which task—once a procurement question—is now a research skill in its own right, requiring the same careful attention paid to the choice of archive, the design of a query, the evaluation of a source. Knowing that Claude Opus 4.7's 36 percent hallucination rate matters more than its composite rank for a humanities literature review; that Gemini 3.1 Pro Preview's 94.3 percent GPQA Diamond performance at two dollars per million tokens is the rational choice for STEM synthesis; that Kimi K2.6's 54 percent HLE score opens cross-disciplinary depth no physical library can match—this is the new information literacy.

The frontier is not a problem to be solved. It is the terrain on which research is now built. Reading its contours—understanding which peaks measure what, which valleys reveal where, and where the human capacities for novelty, judgment, and meaning-making remain irreplaceable—confers decisive advantage in every academic discipline. The machine's map grows more detailed. The reader's must keep pace.

Appendix A: Source Verification Methodology

Numerical claims in this report were verified against a three-tier source hierarchy. First-tier sources are vendor model cards and official release announcements (anthropic.com, openai.com, deepmind.google, api-docs.deepseek.com, huggingface.co/moonshotai, x.ai). Second-tier sources are independent re-evaluations conducted by Artificial Analysis under standardized conditions on first-party APIs. Third-tier sources are aggregator leaderboards (LLM-Stats, BenchLM, Vellum, Epoch AI), used for cross-reference rather than primary attribution because of documented attribution errors in aggregator data flows.

Three corrections to widely circulated values were applied during preparation. Gemini 3.1 Pro Preview launched on February 19, 2026, per Google's official announcement and the DeepMind model card; the "March 19" date appearing in several aggregator entries reflects a mid-March consumer-application rollout, not the model's release. Grok 4.20 entered public beta on February 17, 2026, with API access following March 10, 2026; the commonly cited "March 3" date refers to a Beta 2 patch, not the public release. The HLE score of 57.2 percent attributed to "GPT-5.5" in several aggregator pages belongs to the higher-tier GPT-5.5 Pro variant—priced six times the base model—and should not be conflated with base GPT-5.5 performance.

Two values are reported with explicit lower confidence. Qwen 3.6 GPQA and MMLU-Pro figures could not be cleanly mapped to a single Alibaba SKU during the verification window; the approximate values reflect the Qwen3.6-Max-Preview tier per Alibaba Cloud documentation and Artificial Analysis tracking. SWE-bench Verified scores carry the contamination caveat documented in Section 2: OpenAI's early-2026 audit found verbatim memorization across every frontier model, and the contamination-resistant SWE-bench Pro is the more conservative measure (scores typically twenty to twenty-five percentage points lower for the same systems).

The DeepSeek V4 Pro pricing reflects the standard post-promotional rate; a 75 percent launch promotion at $0.435 input and $0.87 output per million tokens expired May 5, 2026, and has not been renewed.

Appendix B: Consolidated Capability Matrix

Costs are reported in United States dollars per one million tokens (input / output). The Hallucination column reports the AA-Omniscience hallucination rate (the percentage of cases in which the model produces a confident incorrect answer rather than abstaining); where the raw rate is not reported but the composite AA-Omniscience Index is, the Index is shown with the suffix "Idx" (range −100 to +100, higher is more reliable). "n.r." indicates not reported by vendor or independent evaluator at the time of writing. DeepSeek V4 Pro pricing reflects the standard post-promotional rate; a 75 percent launch promotion at $0.435 input and $0.87 output per million tokens expired on May 5, 2026. Kimi K2.6 pricing is the Moonshot first-party API rate; provider blended rates range from approximately $1.15 to $2.15. SWE-bench Verified scores should be read with the contamination caveat noted in Section 2 and Appendix A.

Model	GPQA Diamond	HLE (tools)	SWE-bench V.	ARC-AGI-2	AA Index	Hallucination	Cost $ in/out
GPT-5.5	93.5%	—	—	85%	60	86%	5 / 30
Claude Opus 4.7	94.2%	54.7%	87.6%	~63%	57	36%	5 / 25
Gemini 3.1 Pro	94.3%	44.4%	80.6%	77.1%	57	50%	2 / 12
Kimi K2.6	90.5%	54.0%	—	n.r.	54	n.r.	0.60 / 4
DeepSeek V4 Pro	90.1%	37.7%	80.6%	n.r.	52	−10 Idx	1.74 / 3.48
Qwen 3.6 Max	~88.8%	—	—	—	52	n.r.	enterprise
Grok 4.20	—	—	—	~30%	49	n.r.	2 / 6

Sources: *Vendor model cards (anthropic.com, openai.com, deepmind.google, api-docs.deepseek.com, huggingface.co/moonshotai); Artificial Analysis Intelligence Index v4.0 and AA-Omniscience evaluation leaderboards, May 2026; NIST CAISI evaluation, May 2026.*

References

Anthropic. (2026, April 16). Introducing Claude Opus 4.7. https://www.anthropic.com/news/claude-opus-4-7

Anthropic. (2026). Claude Opus 4.7 Model Card and System Notes. https://www.anthropic.com/research

Artificial Analysis. (2026, May). Intelligence Index v4.0: Model Leaderboard. https://artificialanalysis.ai/leaderboards/models

Artificial Analysis. (2026). GPQA Diamond Evaluation. https://artificialanalysis.ai/evaluations/gpqa-diamond

Artificial Analysis. (2026). Humanity's Last Exam Evaluation. https://artificialanalysis.ai/evaluations/humanitys-last-exam

Artificial Analysis. (2026). AA-Omniscience: Knowledge and Hallucination Benchmark. https://artificialanalysis.ai/evaluations/omniscience

Artificial Analysis. (2026, April). OpenAI's GPT-5.5 is the new leading AI model. https://artificialanalysis.ai/articles/openai-gpt5-5-is-the-new-leading-AI-model

Artificial Analysis. (2026, April). Opus 4.7: Everything you need to know. https://artificialanalysis.ai/articles/opus-4-7-everything-you-need-to-know

Chollet, F. (2024). On the measure of intelligence (updated edition). ARC Prize Foundation. https://arcprize.org/

Chollet, F., et al. (2025). ARC-AGI-2: A benchmark for fluid intelligence. ARC Prize Foundation. https://arcprize.org/arc-agi/2

DeepSeek. (2026, April 24). DeepSeek V4 Pro — Release notes. https://api-docs.deepseek.com/news/news260424

DeepSeek. (2026). DeepSeek V4 Pro Model Card. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro

Epoch AI. (2026). GPQA Diamond Benchmark Tracking. https://epoch.ai/benchmarks/gpqa-diamond

Google DeepMind. (2026, February 19). Gemini 3.1 Pro Preview announcement. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/

Google DeepMind. (2026). Gemini 3.1 Pro Model Card. https://deepmind.google/models/model-cards/gemini-3-1-pro/

Jackson, D., Keating, W., Cameron, G., & Hill-Smith, M. (2025). AA-Omniscience: Evaluating cross-domain knowledge reliability in large language models. arXiv:2511.13029. https://arxiv.org/abs/2511.13029

Jimenez, C., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? International Conference on Learning Representations. https://www.swebench.com/

Moonshot AI. (2026, April 20). Kimi K2.6 Model Card. https://huggingface.co/moonshotai/Kimi-K2.6

NIST Center for AI Standards and Innovation (CAISI). (2026, May). Evaluation of DeepSeek V4 Pro. https://www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro

OpenAI. (2026, April 23). Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/

OpenAI. (2026). GPT-5.5 System Card. https://openai.com/index/gpt-5-5-system-card/

OpenAI. (2026). SWE-bench Verified: Notes on memorization contamination and the transition to SWE-bench Pro. https://openai.com/research/swe-bench-verified-update

Phan, L., Gatti, A., et al. (2025). Humanity's Last Exam. arXiv:2501.14249. https://arxiv.org/abs/2501.14249

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). GPQA: A graduate-level Google-proof Q&A benchmark. arXiv:2311.12022. https://arxiv.org/abs/2311.12022

Scale AI. (2026). Humanity's Last Exam Leaderboard. https://labs.scale.com/leaderboard/humanitys_last_exam

Uzwyshyn, R. (2026, February). The jagged frontier: Mapping AI intelligence and human agency. UCR Library AI Tools Series.

xAI. (2026). Grok 4.20 Release Notes. https://x.ai/blog/grok-4-20