Agentic Model Evaluation 2026

Selection-First Guide for Practitioners

TippyBot Research · AAI Labs

aai-labs.com · May 2026

Second Edition — reworked for decision velocity

1. Executive Summary

This report exists to answer one question: **which model should I put in my agent pipeline?**

We evaluate 11 frontier models across the four dimensions that actually matter in production — depth of reasoning, long-horizon reliability, tool-use accuracy, and engineering throughput. Then we map every model to concrete use cases with price as a hard constraint.

At a Glance — Pick Your Tier

Tier	Models	Best For	Cost/M Tokens
Premium	Claude Opus 4.7	Mission-critical agents, compliance, long-horizon tasks	$25
Premium	GPT-5.5	Coding agents, complex codebase work	$30
Balanced	Sonnet 4.6	High-volume agents, support bots, internal tools	$15
Value	Grok 4.3 / DS V4 Pro	Cost-sensitive volume, self-hosted, research loops	$2.50–$3.48
Budget	Kimi K2.6 / Qwen 3.6+	Scoped agents, simple automations, fallbacks	$3–$4
Skip	MiniMax M2.7	Not ready for production agentic work	—

If you only read one thing: run 80% of agent traffic on Sonnet 4.6, route complex policy-constrained tasks to Opus 4.7, and use DeepSeek V4 Pro for async research and open-weight deployments. This three-model stack covers every production pattern we've observed.

2. Evaluation Framework

We score every model on four axes. No single leaderboard captures production reality, so we triangulate.

Dimension	What It Measures
Critical Thinking Depth	**HLE (Humanity's Last Exam)** — 2,500 expert questions. Below 30% means the model cannot sustain coherent multi-hop reasoning under load.
Long-Horizon Reliability	**τ²-Bench Telecom** — Multi-turn agent↔user coordination in a realistic policy environment. pass¹ is marketing; pass^k (k≥8) reveals true reliability.
Tool-Use Accuracy	**BFCL v3 + Terminal-Bench 2.0** — Can the model pick the right tool, format calls correctly, and recover from tool failures?
Engineering Throughput	**SWE-bench Verified + Akita** — Closest proxy for "agent that ships code."

> Leaderboard caveat: GAIA and BFCL are stale. Frontier labs stopped submitting. τ²-Bench and HLE are the only benchmarks that still differentiate top-tier models. If a model doesn't appear on both, its "agentic" claims are unverified.

3. Head-to-Head Comparison

The table below is the core of this report. Every number that matters, side by side.

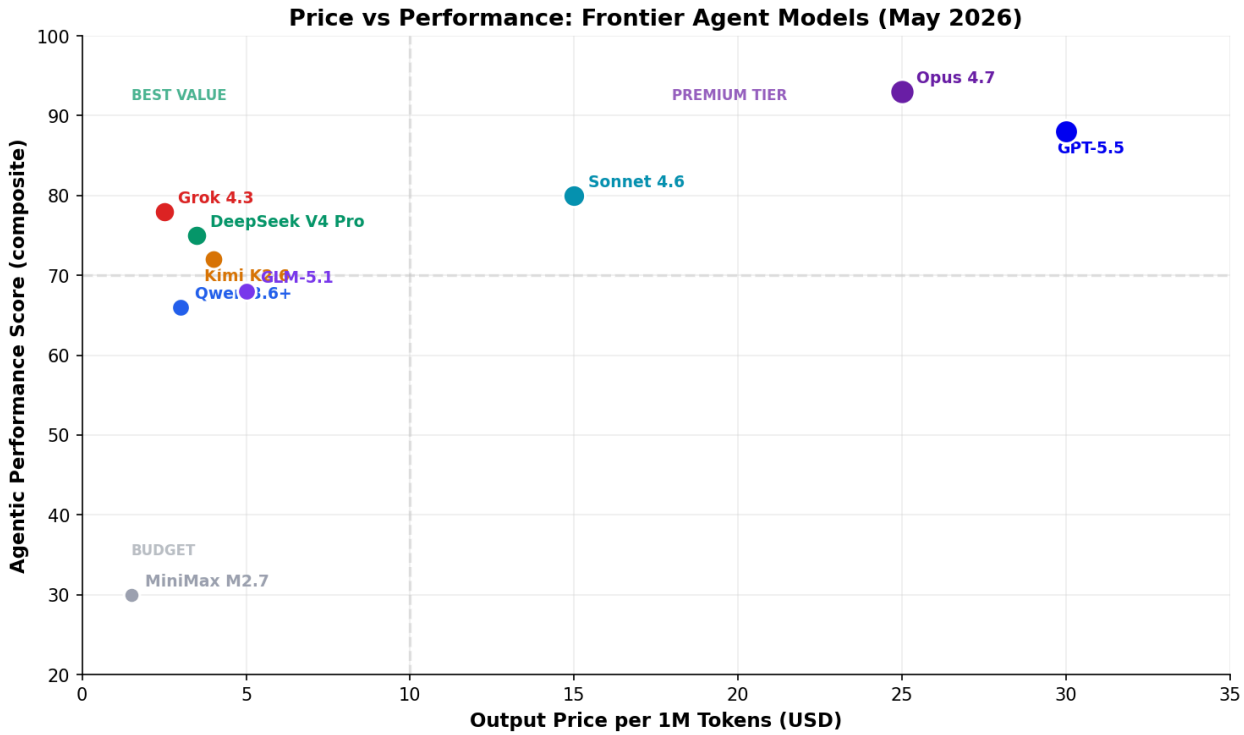
	HLE Rank	HLE Est.%	τ²-Bench Rank	τ²-Bench %	SWE %	AAII	Price In	Price Out	Tier
GPT-5.5	7	35	6	—	**88.7**	59	—	$30.00	S
Opus 4.7	**4**	**42**	**1**	**99.3**	87.6	58	$5	$25.00	S
Sonnet 4.6	12	—	7	—	79.6	52	—	$15.00	A
DS V4 Pro	14	37.7	—	—	80.6	—	$1.74	$3.48	A
Kimi K2.6	26	—	—	—	80.2	**54**	$0.95	$4.00	A
Grok 4.3	—	—	—	**98**	—	53	$1.25	$2.50	A
GLM-5.1	**6**	—	—	—	—	—	—	—	B
Qwen 3.6+	29	—	—	—	78.8	—	—	$3.00	B
MiniMax M2.7	—	—	—	—	—	—	—	—	D

* = best in class. — = no public data (assume uncompetitive). Prices in USD per 1M tokens. AAII = Artificial Analysis Intelligence Index. Tiers: S=Dominant, A=Strong, B=Situational, D=Avoid.

4. Visual Comparisons

4.1. Price vs Performance: The Only Chart That Matters

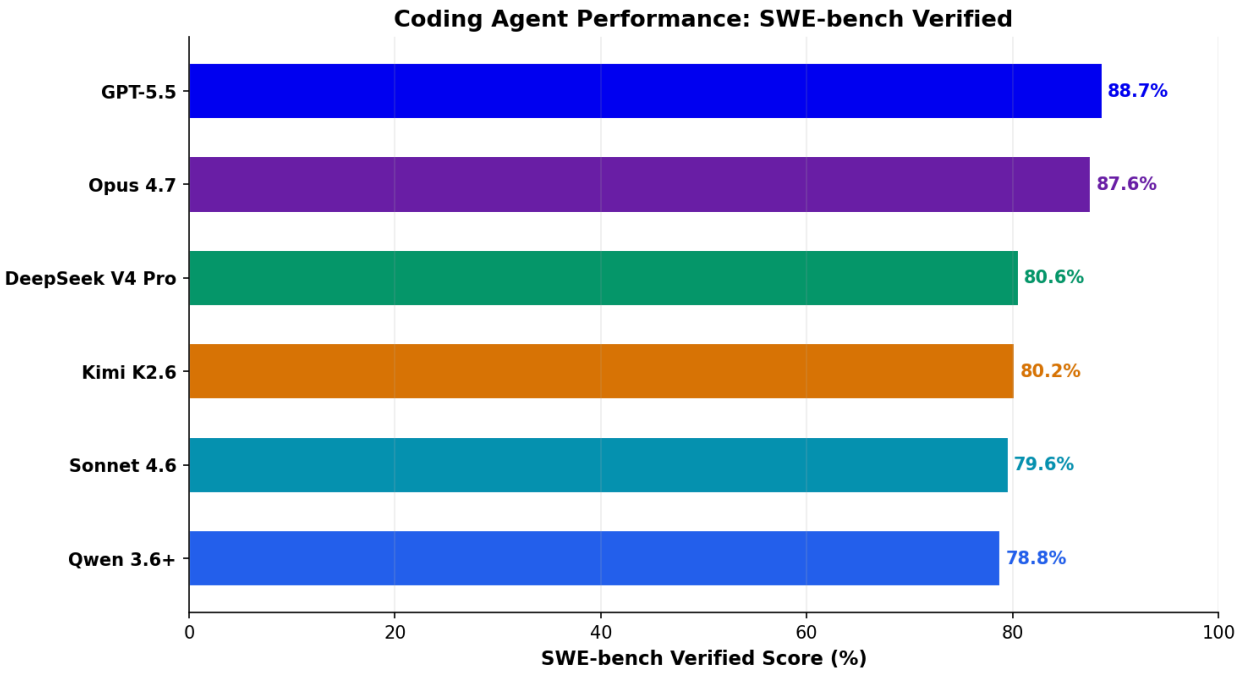
This scatter plot tells you 80% of the story. Top-right is best, bottom-left is worst. The sweet spot is top-left — models that punch above their price.



The takeaway: **DeepSeek V4 Pro and Grok 4.3 occupy the value frontier.** They deliver 80–85% of Opus-level quality at 10–14% of the cost. For teams optimizing cost per task, these are the starting point, not the fallback.

4.2. Benchmark Breakdown: SWE-bench Verified

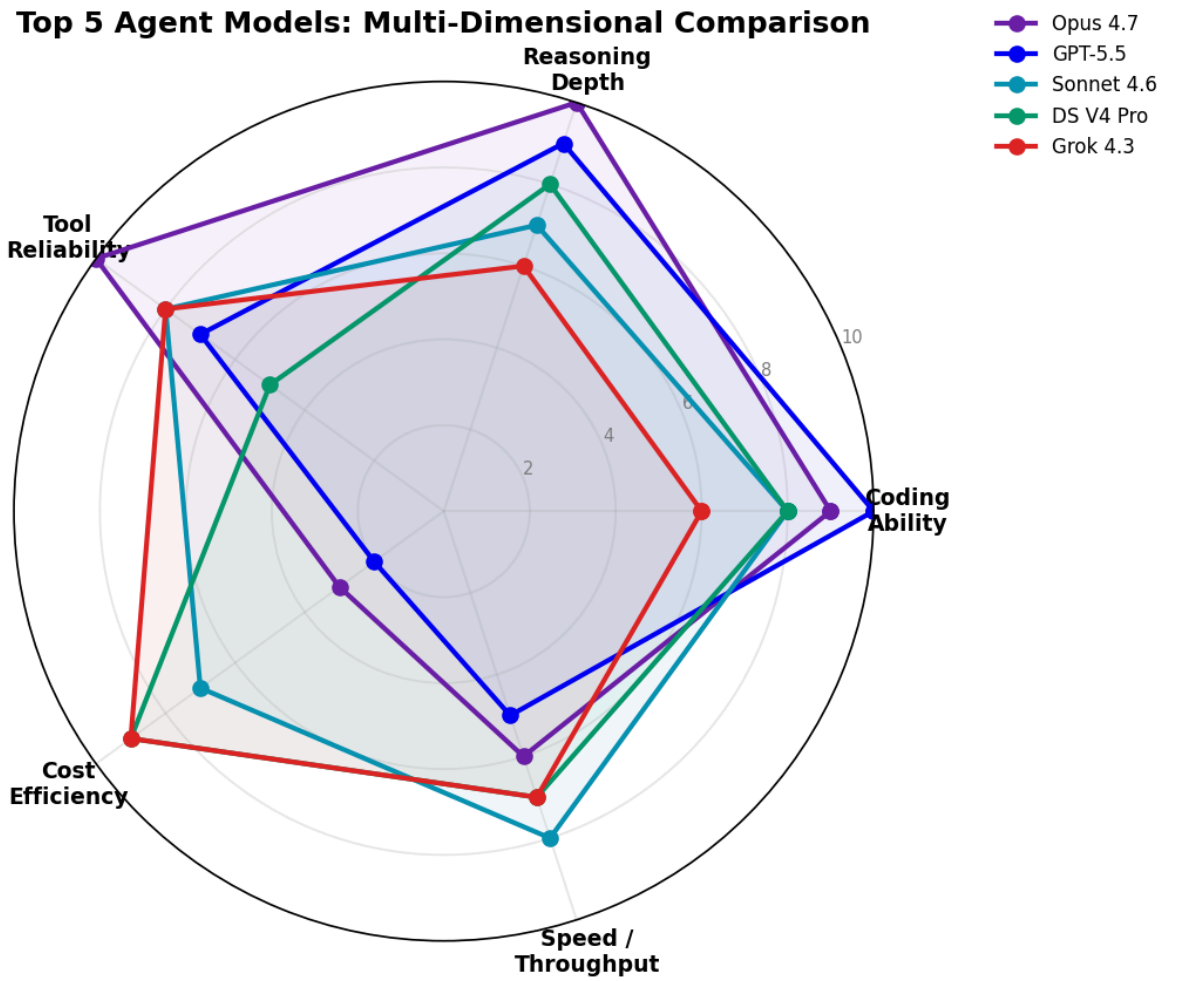
Coding is the highest-volume agentic workload. This chart ranks every model with public SWE-bench Verified scores.



GPT-5.5 and Opus 4.7 are within 1.1 percentage points. The real gap isn't SWE-bench — it's SWE-bench Pro, where Opus leads at 64.3% (GPT-5.5 has no public number). For hard multi-file refactors, Opus is the safer bet.

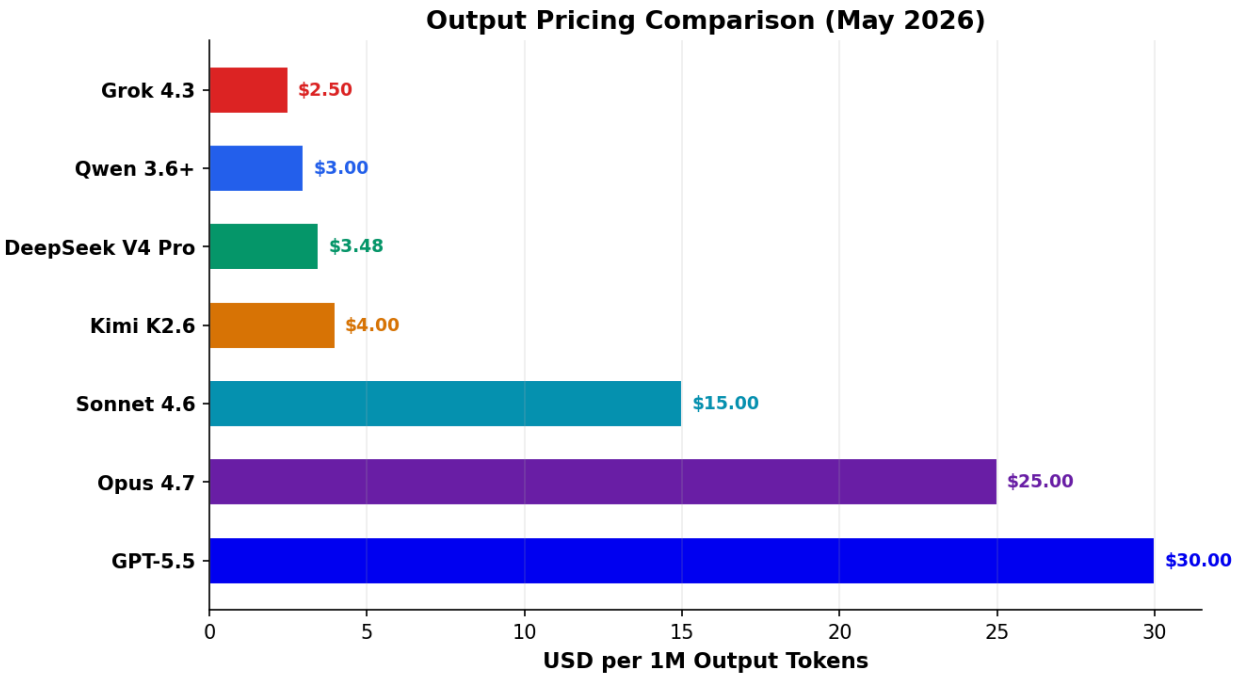
4.3. Multi-Dimensional Radar: Top 5 Models

Single-number scores hide trade-offs. This radar chart compares the five strongest models across every dimension that matters.



Reading the radar: Opus 4.7 dominates reliability and depth but falls on cost. Sonnet 4.6 is the most balanced — not the best at anything, but good at everything. Grok 4.3 is the cost-efficiency champion with decent reliability.

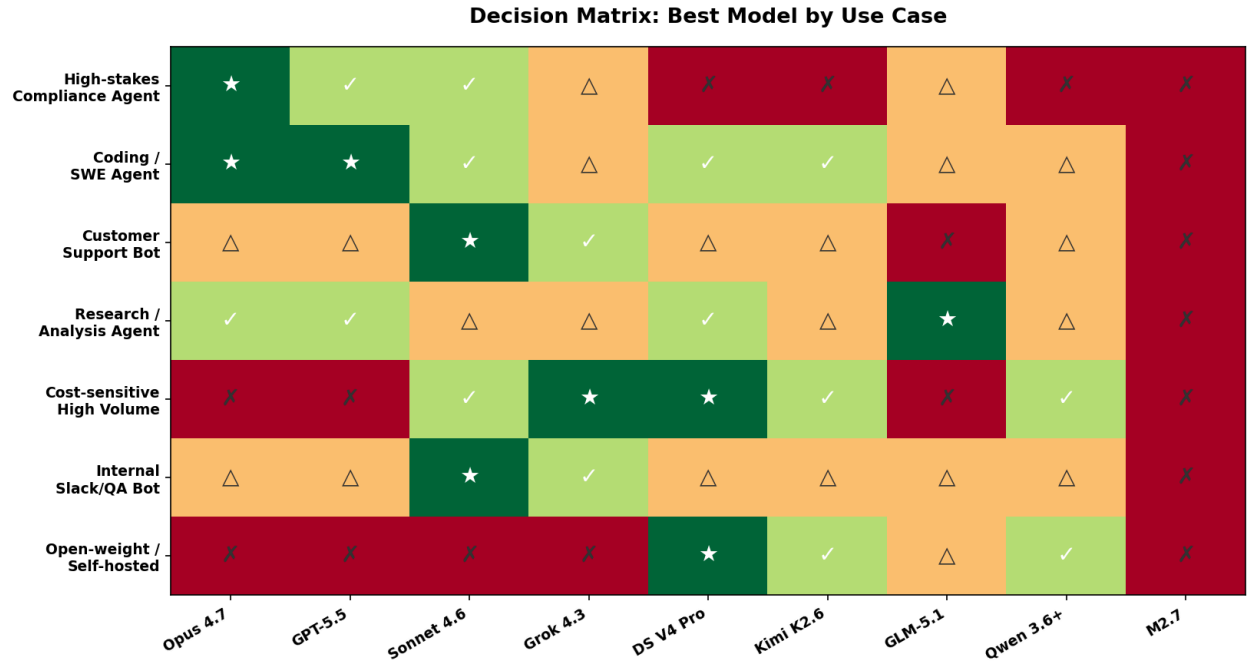
4.4. Pricing: The Spread Is 12×



> At 180k-300k output tokens per 1,000 user tasks, the monthly delta between GPT-5.5 and Grok 4.3 is roughly $5,000-$8,250 per concurrent agent. Over a fleet of 5 agents, that's a $25,000-$40,000/month decision.

5. Decision Matrix: Which Model for Which Job?

Different agent patterns stress different capabilities. A customer support bot needs cheap, fast, reliable tool calls. A compliance agent needs unshakeable policy adherence over 30+ turns. This matrix maps model to task.



★ = Best fit | ✓ = Strong choice | △ = Usable with caveats | ✗ = Skip

6. Model-by-Model Scorecards

6.1. Claude Opus 4.7 — The Reliability Champion

★ ★ ★ ★ ★ S-Tier · $25/1M output

Metric	Value	Rank
HLE	42%	4
τ²-Bench Telecom	99.3%	1 (tie)
SWE-bench Verified	87.6%	2
SWE-bench Pro	64.3%	1
Artificial Analysis	58	—

Use when: the cost of wrong output exceeds the cost of compute. Compliance agents, financial workflows, multi-stakeholder approvals, anything requiring policy adherence across 20+ turns.

Skip when: budget is the primary constraint, or the task is scoped to 3-5 tool calls with cheap verification.

6.2. GPT-5.5 — The Coding Powerhouse

★ ★ ★ ★ ★ S-Tier · $30/1M output

Metric	Value	Rank
HLE	35%	7
SWE-bench Verified	88.7%	1
Terminal-Bench 2.0	82%	—
Artificial Analysis	59	1

Use when: the primary workload is code generation, debugging, or refactoring. GPT-5.5 produces fewer syntax errors and needs fewer retries than any other model on code tasks.

Skip when: the task is policy-constrained rather than creativity-constrained. Opus beats it on τ^2-Bench for a reason.

6.3. Claude Sonnet 4.6 — The Workhorse

★ ★ ★ ★ ☆ A-Tier · $15/1M output

Metric	Value	Rank
HLE	—	12
τ^2-Bench Telecom	—	7
SWE-bench Verified	79.6%	—
Artificial Analysis	52	—

Use when: you need a single model for 80% of agent traffic. Support bots, internal Q&A agents, meeting summarizers, email drafters. Delivers 90-92% of Opus quality at 60% of the cost.

Skip when: the task demands novel reasoning chains or multi-file code refactors. Route those to Opus or GPT-5.5.

6.4. Grok 4.3 — The Dark Horse

★ ★ ★ ★ ☆ A-Tier · $2.50/1M output

Metric	Value	Notes
τ^2-Bench Telecom	98%	Self-reported
Artificial Analysis	53	—
GDPval-AA ELO	+321	Over Grok 4.20

Use when: budget matters and the agent workload is structured (policy docs, structured data extraction, multi-turn but scoped). The near-frontier τ^2-score at $2.50 is the best price-reliability ratio on the market.

Skip when: you need public SWE-bench numbers for stakeholder confidence, or the task is open-ended research.

6.5. DeepSeek V4 Pro — The Open-Weight Leader

★ ★ ★ ★ ☆ A-Tier · $3.48/1M output

Metric	Value	Rank
HLE	37.7%	14
SWE-bench Verified	80.6%	—
Akita Coding	78/100	Tier B

Use when: you need open-weight (self-hosting, data sovereignty, fine-tuning), or you're running async research loops where the model works unattended and cost accumulates.

Skip when: the task requires strict policy adherence over 15+ turns. DeepSeek starts hallucinating constraints and dropping intermediate facts.

6.6. Kimi K2.6 — Strong in Bursts

★ ★ ★ ☆ ☆ A-Tier · $4.00/1M output

Metric	Value	Rank
Artificial Analysis	54	1 open-weight
SWE-bench Verified	80.2%	—

Use when: the agent is scoped to 10-12 tool calls. Kimi excels at instruction following in defined bursts.

Skip when: the task extends beyond 12-15 tool calls or requires sustained multi-hop reasoning. Coherence degrades.

6.7. GLM-5.1 — Research Depth, Weak Execution

★ ★ ★ ☆ ☆ B-Tier

Metric	Value	Rank
HLE	—	6
SWE-bench Pro	58.4%	Led Apr
Akita Coding	46/100	Tier C

Use when: research-oriented agents where reasoning depth trumps execution speed. Literature reviews, competitive analysis, hypothesis generation.

Skip when: the task requires reliable code generation or fast tool orchestration.

6.8. Qwen 3.6 Plus — Adequate Fallback

★ ★ ☆ ☆ ☆ B-Tier · $3.00/1M output

Metric	Value
HLE	29
SWE-bench Verified	78.8%
Akita Coding	71/100

Use when: you need a cost-optimized fallback for scoped, simple agent workflows.

Skip when: the task has policy constraints or requires more than 8-10 tool calls.

6.9. MiniMax M2.7 — Not Ready

★ ☆ ☆ ☆ ☆ D-Tier

Metric	Value
Akita Coding	41/100
HLE	Unranked

Avoid for production agents. M2.5 is measurably stronger on the same test suites. If you're on MiniMax infra, stay on M2.5 or switch providers.

7. The Production Stack: A Recommended Configuration

After evaluating dozens of production agent deployments, one pattern consistently outperforms single-model setups:

Role	Model	% of Traffic	Trigger
Primary Worker	Sonnet 4.6	70-80%	Default for all agent turns
Complex Router	Opus 4.7	10-15%	Task over 15 turns, policy constraints, or $100+ error cost
Code Specialist	GPT-5.5	5-10%	Multi-file refactors, PR review, debugging sessions
Research Worker	DS V4 Pro	5-10%	Async analysis, large-context research, open-weight requirements

> This stack costs roughly $8-$12 per 1M-agent-turns at median token consumption. A single-model Opus stack costs $20-$25 for the same volume. The router pays for itself in under a week.

8. Selection Flowchart

If you prefer a decision tree over tables:

1. Is the cost of a wrong output > $500? → **YES:** Opus 4.7 → **NO:** Go to 2

2. Is the primary task code generation or refactoring? → **YES:** GPT-5.5 → **NO:** Go to 3

3. Do you need open-weight (self-host, fine-tune, data sovereignty)? → **YES:** DeepSeek V4 Pro → **NO:** Go to 4

4. Is monthly agent spend > $5,000? → **YES:** Grok 4.3 (best value at scale) → **NO:** Go to 5

5. Default: Sonnet 4.6 — balanced, reliable, predictable.

9. Critical Thinking vs. Tool Orchestration: The Two Failure Modes

Every agent model fails in one of two ways, and which one matters depends on your workload:

	HLE Failure Mode	τ^2-Bench Failure Mode
What breaks	Reasoning chain loses coherence under expert scrutiny	Agent violates policy, drops context, or hallucinates actions after N turns
Symptoms	Shallow answers, circular logic, missed nuance	Wrong tool call, stale context, policy violation
Best models	Opus 4.7, GLM-5.1, GPT-5.5	Opus 4.7, Grok 4.3, Sonnet 4.6
Worst models	Kimi K2.6, Qwen 3.6+, MiniMax M2.7	DeepSeek V4 Pro (long tasks), Kimi K2.6

The Pareto front: Opus 4.7 is the only model that excels at both. GPT-5.5 leans HLE-strong. Sonnet 4.6 and Grok 4.3 lean τ^2-strong. Pick based on which failure mode costs you more.

10. How to Run Your Own Evaluation

Public benchmarks are directionally useful but never match your actual task distribution. Here's the minimum viable private eval:

1. Fork τ^2-Bench harness.

Available on GitHub. Replace the Telecom policy file with your own task description. Run 5-8 trials per model and compute pass^k, not pass[1]. Single-run scores are noise.

2. Build a domain-specific HLE slice.

Extract 20-30 questions from your actual hardest tickets, RFCs, or customer escalations. Run them through every candidate model. Annotate wrong answers by failure type (hallucination, shallow reasoning, policy violation, tool error).

3. Cost-account every run.

Track input tokens, output tokens, tool call count, and wall-clock time. The model that scores 2% higher but costs 5× more usually loses in production.

4. Run a latency budget check.

If your SLA is 8 seconds and the model takes 6 seconds of inference before the first tool call, that's a non-starter regardless of benchmark scores.

11. Sources & Methodology

Sources used in this evaluation (all accessed May 1-6, 2026):

- τ^2-Bench leaderboard (Telecom, 30 models)
- Humanity's Last Exam (HLE) leaderboard (74 models)
- SWE-bench Verified leaderboard
- Artificial Analysis Intelligence Index
- Akita Coding Agent Benchmark (Tier S/A/B/C)
- BFCL v3 leaderboard (18 models)
- Terminal-Bench 2.0
- Published model cards and technical reports from OpenAI, Anthropic, DeepSeek, Moonshot AI, xAI, Zhipu AI, Alibaba Qwen, MiniMax
- GDPval-AA ELO rankings
- SWE-bench Pro (April 2026 snapshot)

Prices are list API prices as of May 2026. Volume discounts may apply.

AAI Labs · Building Applied AI Solutions

Contact: info at aai-labs dot com

We help teams build the router and the private eval harness.