PRACTICAL BUYER'S GUIDE

Choose the right AI model for the job.

A concise framework for comparing language models by task quality, reliability, latency, cost, security, and deployment fit - not by leaderboard position alone.

01	**Start with the workflow** Define the user, decision, source data, expected output, and cost of a wrong answer.
02	**Use real examples** Test representative tasks from your business, including difficult, ambiguous, and unsafe cases.
03	**Measure the system** Evaluate prompts, retrieval, tools, permissions, and human review - not only the base model.
04	**Route by difficulty** Use a smaller model for routine work and escalate only the tasks that need frontier reasoning.
05	**Re-test before switching** Model versions change quickly. Keep a stable evaluation set and compare before every upgrade.

The best model is the least expensive system that reliably meets your acceptance criteria.

A practical selection map

Shortlist by workload first. A single business solution may route different tasks to different models.

Workload	Start with	Evaluate most closely	Watch for
Extraction and classification	Fast, lower-cost model with structured outputs	Field accuracy, schema validity, throughput, cost per document	Edge cases, OCR quality, silent omissions
Knowledge assistant / RAG	Balanced model with citations and retrieval support	Groundedness, answer completeness, retrieval recall, permission handling	Plausible answers unsupported by sources
Customer service	Fast model with strong instruction following	Resolution rate, escalation accuracy, tone, policy compliance	Over-promising, unsafe actions, poor handoffs
Complex analysis	Frontier reasoning model	Rubric score, evidence quality, calculation accuracy, reviewer preference	High latency, token cost, unsupported claims
Agentic workflow	Model with dependable tool use and state handling	Task completion, tool-call accuracy, recovery, steps and total cost	Loops, excess permissions, irreversible actions
Coding and product development	Strong coding model in the intended developer harness	Passing tests, review quality, maintainability, security, time to merge	Benchmark contamination, environment mismatch
Sensitive or regulated work	Approved deployment and controls before model capability	Privacy, retention, residency, auditability, bias, human oversight	Using public benchmarks as a risk assessment

Architecture questions that can outweigh model scores

- Where can prompts and outputs be processed and retained?
- Can the model return valid structured data and call the tools you need?
- Does the platform support identity, permissions, audit logs, and private networking?
- Can you pin a stable version, monitor changes, and move workloads if necessary?
- What is the complete cost per successful task, including retries, retrieval, tools, and review?

Build an evaluation that reflects your business

Public benchmarks help create a shortlist. A decision requires evidence from your own workflow.

1. Define acceptance criteria	Write down what a correct answer must contain, what it must avoid, and when a person must approve the result.
2. Assemble 50-100 representative cases	Include ordinary work, hard cases, incomplete inputs, prompt injection attempts, sensitive data, and cases that should be refused or escalated.
3. Freeze the test setup	Use the same prompts, retrieval sources, tools, permissions, reasoning effort, and output format for every candidate.
4. Run repeated trials	Non-deterministic systems can vary. Repeat important cases and track both average performance and worst-case failures.
5. Review errors, not only averages	A 95% score may still be unacceptable if the remaining 5% includes fabricated citations, privacy failures, or unauthorized actions.
6. Pilot with real users	Measure time saved, completion rate, review burden, adoption, and business impact before expanding.

A simple weighted scorecard

Dimension	Suggested weight	Example measure
Task quality	30%	Rubric score or exact-match accuracy
Reliability	20%	Pass rate across repeated runs and difficult cases
Groundedness and safety	20%	Supported claims, policy compliance, correct refusal and escalation
Latency	10%	P50 and P95 end-to-end time
Cost	10%	Total cost per successful task
Security and operations	10%	Deployment fit, observability, versioning, support

Adjust the weights before testing. For high-risk workflows, make safety and governance hard gates rather than tradeable points.

Current model snapshot

Selected public information as of September 2026. Use it to create a shortlist, then verify current availability, pricing, and performance with the linked official sources.

Model	AA Index v4.1.1*	Selected reported performance	Published API price / 1M tokens
OpenAI GPT-6 Astra	61.2	AutomationBench 41.4%; BrowseComp 91.5%; OSWorld 2.0 partial score 72.6%	$10 input / $50 output
Anthropic Claude Fable 5.1	65.7	AutomationBench 31.4% in OpenAI's comparison; Anthropic reports its strongest Claude performance on coding and knowledge work	$10 input / $50 output
Anthropic Claude Opus 5	63.1	AutomationBench 26.9%; BrowseComp 90.8%; OSWorld 2.0 partial score 70.2%	Check current Anthropic pricing
Google Gemini 3.8 Flash	58.7	Google positions it for long-horizon engineering, autonomous agents, multimodal inputs, and enterprise workflows	$0.75 input / $3.75 output through Dec 31, 2026

* Artificial Analysis Intelligence Index v4.1.1 values and selected comparison results were published by OpenAI on September 3, 2026. They were run at maximum reported effort in a research environment or API. Benchmark versions, harnesses, safeguards, and model settings can materially change results. A dash or missing score is not evidence of poor performance.

Official sources

OpenAI model overview: platform.openai.com/docs/models

OpenAI GPT-6 Astra release and comparison: openai.com/index/gpt-6-astra

Anthropic Claude Fable 5.1: anthropic.com/claude/fable

Google Gemini model catalog and pricing: ai.google.dev/gemini-api/docs/models and /pricing

AWS model evaluation: docs.aws.amazon.com/bedrock/latest/userguide/evaluation.html

NIST AI Risk Management Framework: nist.gov/itl/ai-risk-management-framework

Need help selecting and validating a model?

Ai Applied can turn your business workflow into a practical evaluation and production plan.

aiapplied.ca/contact