Cybernomics

RACI
Relative AI Cost Index

CYBERNOMICS PROPRIETARY BENCHMARK

# The Relative AI Cost Index 2026

A decision-maker report for understanding the real cost pressure behind AI agents.

Built from published provider pricing and Cybernomics benchmark calculations. Designed to help executives decide what to build, what to control, and what to measure before AI agent costs scale.

Pricing snapshot: May 29, 2026 | Cybernomics Research

EXECUTIVE ANSWER

# What leaders need to know

The RACI report turns AI agent cost into a simple management view. The goal is not to buy the cheapest model. The goal is to design agents that spend only where value requires it.

**100**

**Highest cost pressure in the model benchmark**
GPT-5.5 is the baseline highest mainstream model in this snapshot.

**2**

**Lowest score in the same benchmark**
Gemini 2.5 Flash-Lite is built for low-cost utility tasks.

**90%**

**Typical cached-input discount signal**
Cache hits can reduce repeated input-token cost by about 90%.

**50%**

**Batch discount for work that can wait**
Useful for reports, enrichment, QA, and cleanup.

**5**

**Control points every agent needs**
Model, context, tools, retries, and review.

### 1. The model is not the only cost.

Tool calls, search, file retrieval, containers, storage, latency, and human review can move the bill.

### 2. Long context is expensive.

Every extra page, transcript, tool result, and memory item can become input cost.

### 3. Agents can retry themselves into waste.

Without limits, agents search again, call tools again, and ask the model again.

### 4. Cheaper models work for many tasks.

Classification, extraction, routing, tagging, and formatting rarely need the strongest model.

### 5. The winning metric is cost per useful outcome.

Leaders should measure cost per resolved ticket, completed workflow, proposal, report, or decision.

Data pages use published provider pricing where available. RACI scores are Cybernomics normalized benchmark calculations.

CYBERNOMICS POINT OF VIEW

# AI cost control is becoming a board-level discipline

As businesses move from chatbots to agents, the financial question changes. The question is no longer what a model costs. The question is what a workflow costs when the model, tools, retries, context, and review all run together.

### The shift

Chatbots answer. Agents act. Once agents act across workflows, cost becomes operational, not experimental.

### The risk

Teams can make a good demo and still build a bad operating model if every request uses premium models, long context, live search, and open-ended retries.

### The opportunity

The best companies will build AI agents that know when to think, when to retrieve, when to use tools, when to stop, and when to escalate.

### The Cybernomics position

AI economics should be governed the same way leaders govern risk, workflow, and security: with standards, measurements, and accountability.

### Plain-English summary

Do not let every AI agent drive like a sports car. Most business work needs a reliable sedan. Save the race car for the few tasks that truly need it.

HOW THE RELATIVE AI COST INDEX WORKS

# A simple score built from real prices

RACI compares cost pressure across AI models, tools, and agent scenarios using a normalized score.

### Step 1 - collect prices

Use published provider prices for input tokens, output tokens, cached inputs, batch, search, file search, containers, and retrieval services.

### Step 2 - apply a workload

Apply the same workload to each item in a comparison. For model pricing, the standard workload is 10,000 agent steps.

### Step 3 - normalize

The highest-cost item gets 100. Everything else is expressed relative to that baseline.

### Core model benchmark formula

RACI Model Cost = 10,000 steps x (10,000 input tokens + 2,000 output tokens)
This creates an equal test across models. It does not claim every workflow uses exactly this pattern.

### Full cost view

AI agent cost pressure = model cost + context cost + tool cost + retry risk + governance burden.
The index helps leaders see where cost pressure is likely to appear before they scale.

DATA VALIDATION GUARDRAILS

# Real price inputs. Clear assumptions. Refreshable index.

The dollar inputs in this report come from published provider pricing. RACI scores are Cybernomics calculations that make those prices easier to compare.

### What is real

Published prices: input tokens, output tokens, cached inputs, batch discounts, web search, file search, container sessions, grounding, search credits, and reranking.

### What is modeled

Benchmark workloads, RACI scores, agent scenario cost examples, and maturity scoring.

### What is excluded

Enterprise discounts, reserved capacity, private contracts, region uplifts, taxes, and engineering labor unless noted.

### When to refresh

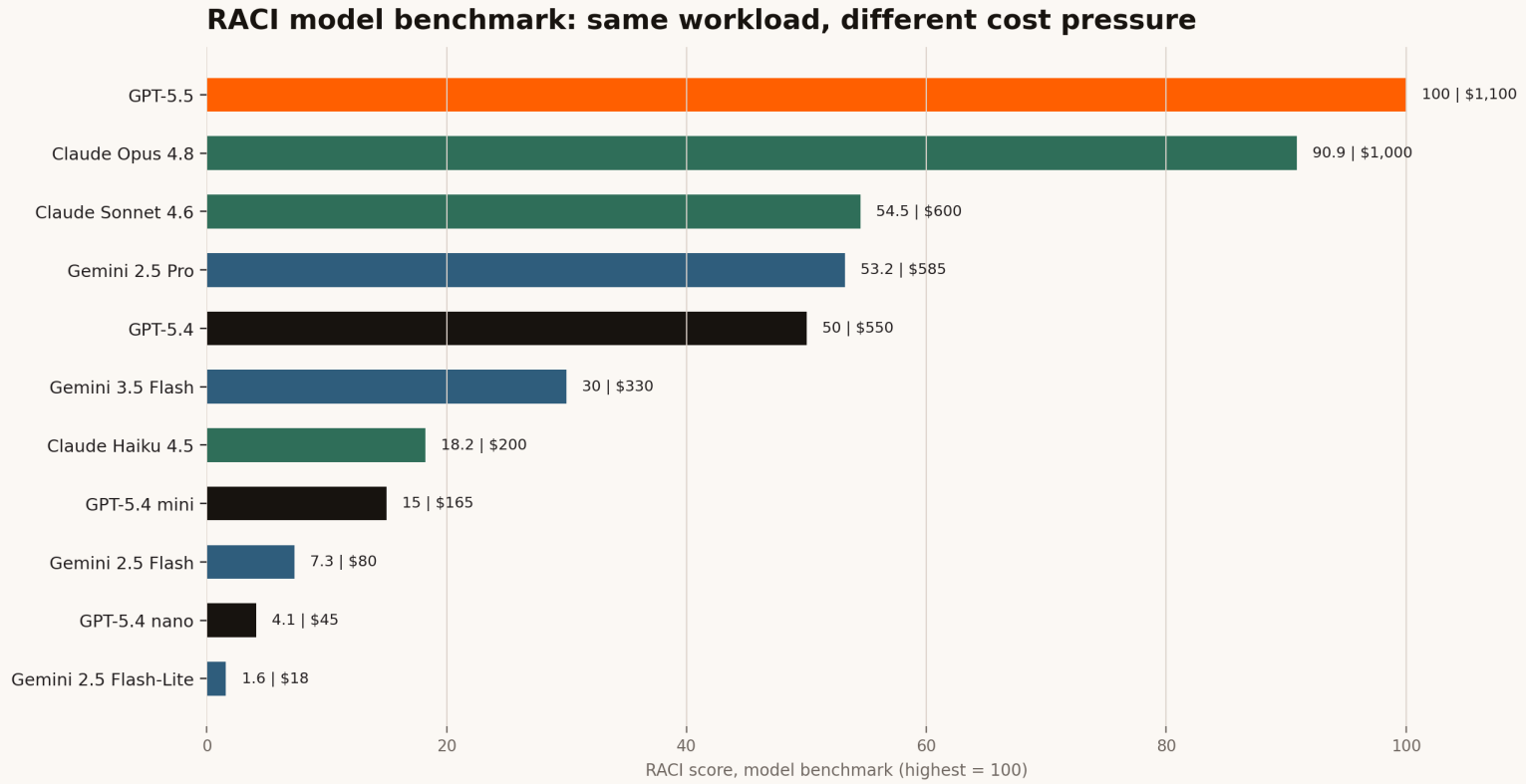
Before vendor negotiation, production rollout, board budget approval, or any major model/provider change.

| Provider | Price inputs used |
|---|---|
| OpenAI | GPT-5.5, GPT-5.4, GPT-5.4 mini/nano; cached input; Batch API; web search; file search; containers; realtime audio proxy. |
| Anthropic | Claude Opus 4.8, Sonnet 4.6, Haiku 4.5; cache write/read pricing. |
| Google | Gemini 3.5 Flash, Gemini 2.5 Pro, Gemini 2.5 Flash, Gemini 2.5 Flash-Lite; search grounding and context caching. |
| Tavily | Pay-as-you-go API credit pricing. |
| Pinecone | Rerank requests and sparse embedding token pricing. |

RACI MODEL BENCHMARK

# The same token workload can produce very different bills

This chart compares published model input/output prices using one standard agent workload.




### How to read it

100 is the highest cost pressure in this mainstream model set. Lower scores mean less cost pressure for the same workload.

### What it means

Do not use one premium model for every step. Use a routing policy that matches task difficulty.

### Leader question

Which agent steps truly require premium reasoning, and which can use efficient or utility models?

Model benchmark: 10,000 steps x 10k input tokens + 2k output tokens. Source inputs: OpenAI, Anthropic, and Google published pricing.

MODEL PRICE BANDS EXECUTIVES CAN REMEMBER

# A routing policy beats a single-model strategy

This page translates model pricing into a practical model-selection standard.

| Band | Use it for | Representative models | RACI range |
|---|---|---|---|
| Premium | Hard reasoning, executive output, complex coding, regulated decisions | GPT-5.5; Claude Opus 4.8 | 91-100 |
| High / mid | General agent work, analysis, customer-facing drafting | Claude Sonnet 4.6; Gemini 2.5 Pro; GPT-5.4 | 50-55 |
| Efficient | Summaries, Q&A, routing, support answers, repeatable workflows | Gemini 3.5 Flash; Claude Haiku 4.5; GPT-5.4 mini | 15-30 |
| Utility | Classification, extraction, tagging, formatting, background cleanup | Gemini 2.5 Flash; GPT-5.4 nano; Gemini Flash-Lite | 2-7 |

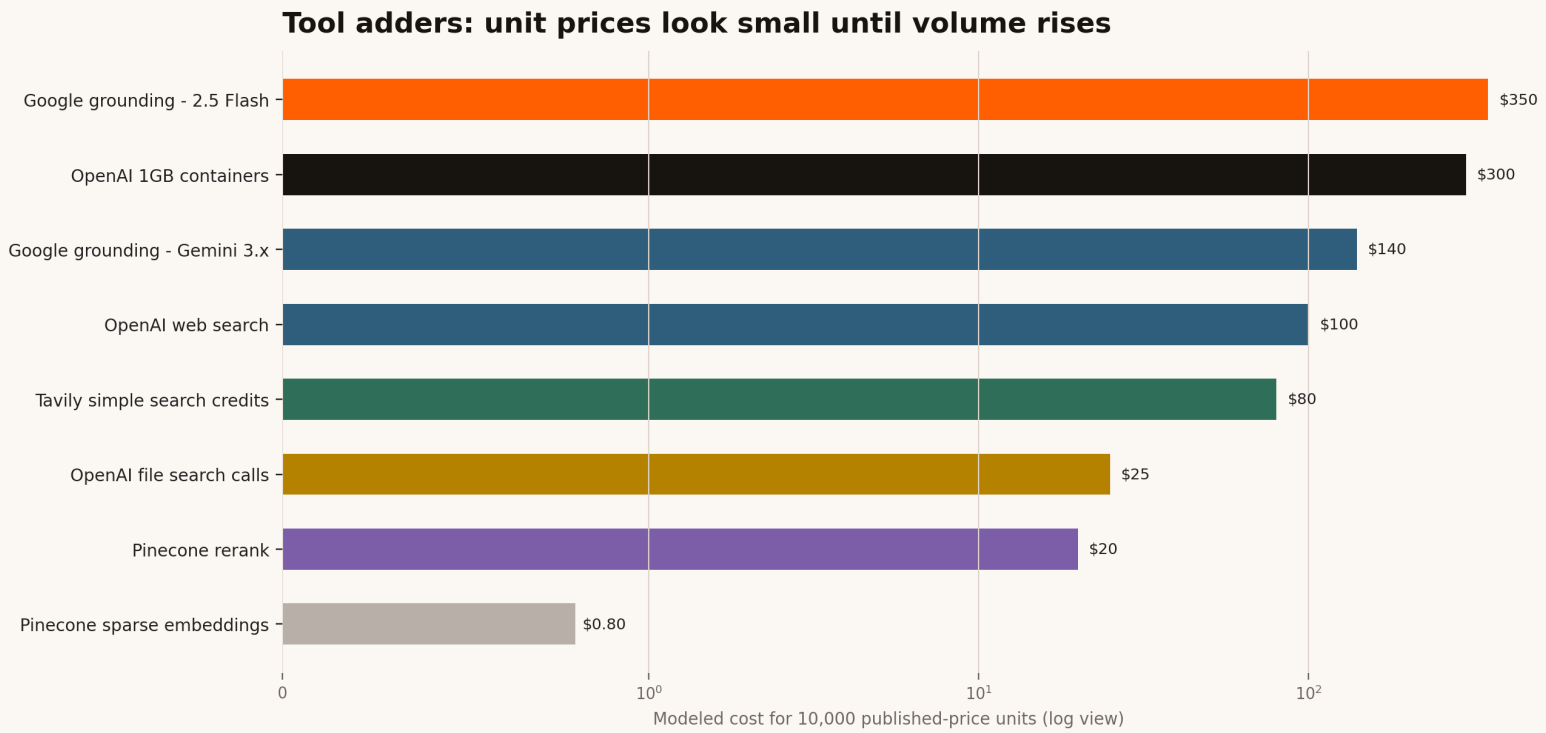
**Leadership rule**

The expensive model should be the escalation path, not the default path.

TOOL ADDERS CAN CHANGE THE ECONOMICS

# The model is only part of the bill

When agents search the web, call retrieval systems, open containers, or run repeated tool loops, the cost stack changes.




### Why this matters

A cheap model can still create an expensive workflow if it calls costly tools too often.

### Control rule

Set tool budgets, allowlists, retry caps, and clear stop conditions before launch.

### Watch closely

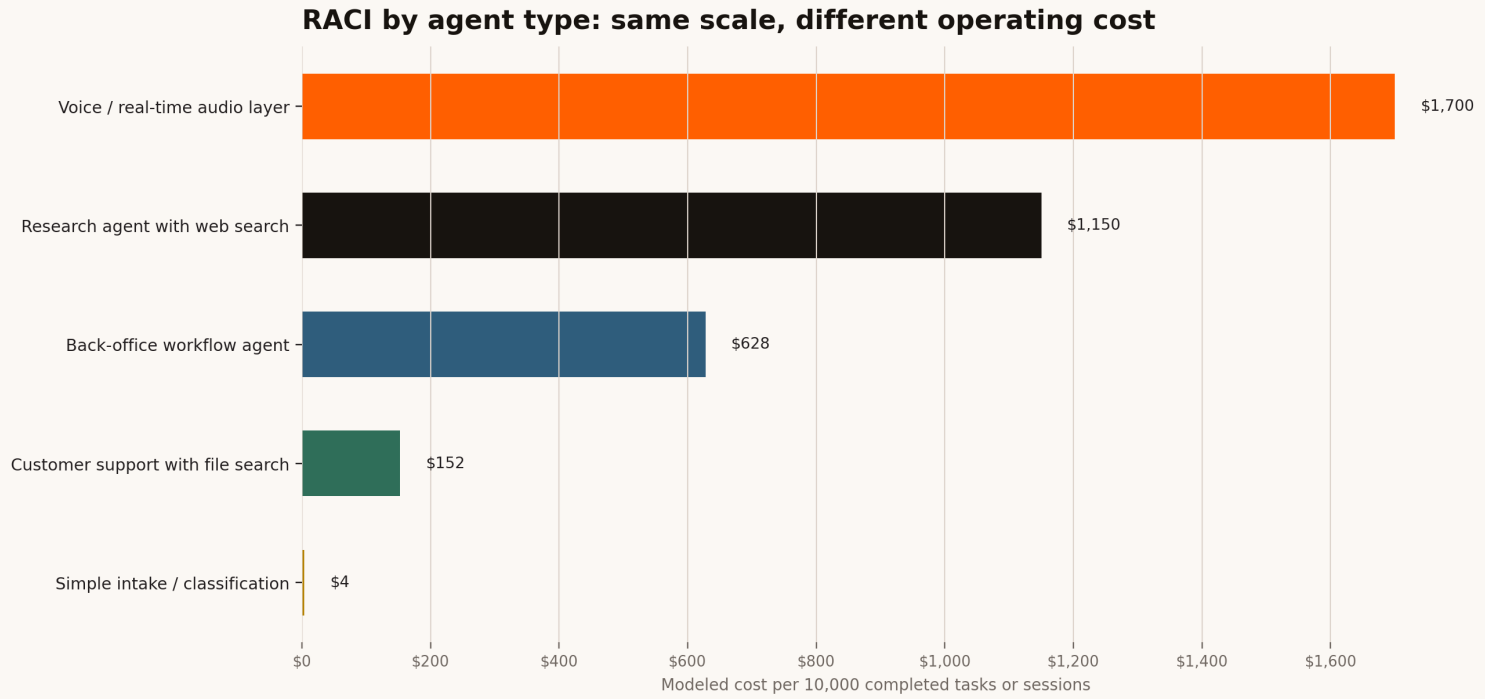
Web search, grounding, containers, and file search should be measured per completed outcome.

Tool units differ by provider and product. This page shows a 10,000-unit view using published unit prices. Search credit usage can vary by Tavily operation.

RACI BY AGENT TYPE

# Different agents carry different cost risks

Executives should compare agent economics by use case, not just by vendor.




### Voice / real-time audio layer

Modeled cost: $1,700
10k x 5-minute sessions | $0.034/min audio layer only

### Research agent with web search

Modeled cost: $1,150
20k input + 3k output | 2 web searches/task

### Back-office workflow agent

Modeled cost: $628
10k input + 2k output | 0.5 file search + 0.05 containers/task

### Customer support with file search

Modeled cost: $152
8k input + 1.5k output | 1 file search/case

### Decision use

Use this view to decide which agents need routing, caching, batching, and tool caps first.

Scenario costs are explicit Cybernomics benchmark assumptions built from published price inputs. Voice example uses OpenAI realtime translate price as an audio-layer proxy, not a full contact-center TCO.

FIVE REAL-WORLD AGENT SCENARIOS

# This is how leaders should think about cost risk

Every agent should be approved with a use case, a volume estimate, a model policy, and a tool policy.

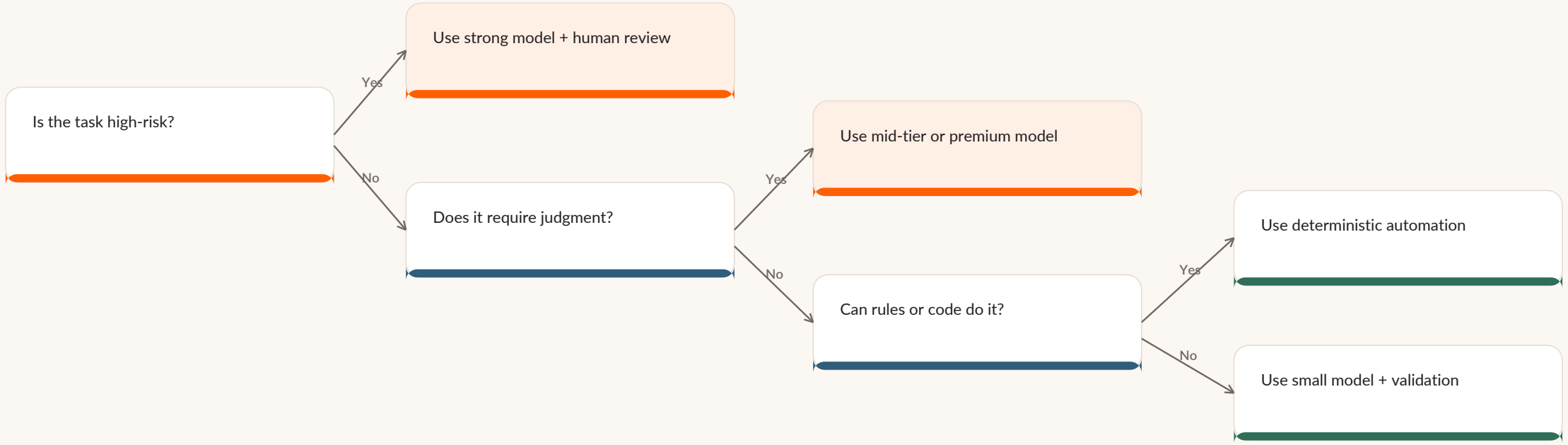
| Agent type | Cost risk | Why it gets expensive | Primary control |
|---|---|---|---|
| Simple intake / classification | Low | Short prompts, short outputs, few or no tools. | Utility model + deterministic rules |
| Customer support with file search | Medium | Many conversations and frequent retrieval calls. | RAG caps + cache + escalation rules |
| Back-office workflow agent | Medium / high | Tool calls, retries, approvals, and integration failures. | Tool allowlist + retry budget |
| Research agent | High | Web search, long context, repeated summarization, citations. | Search cap + source quality threshold |
| Voice / real-time audio agent | High | Real-time audio, latency needs, transcription, and follow-up actions. | Session budget + summary cache + escalation |

Scenario classification is a Cybernomics decision framework. Calibrate with real workflow traces before scaling.

SHOULD THIS AGENT USE A POWERFUL MODEL?

# A simple decision tree for executives

Use this before approving the model choice for a new AI agent.




**Simple rule**

Start cheap when the work is simple. Escalate only when the task is risky, ambiguous, customer-impacting, or strategically important.

AI COST CONTROL MATURITY MODEL

# Move from blind spend to cost advantage

This gives Cybernomics a simple way to assess where a company is today and what to fix next.

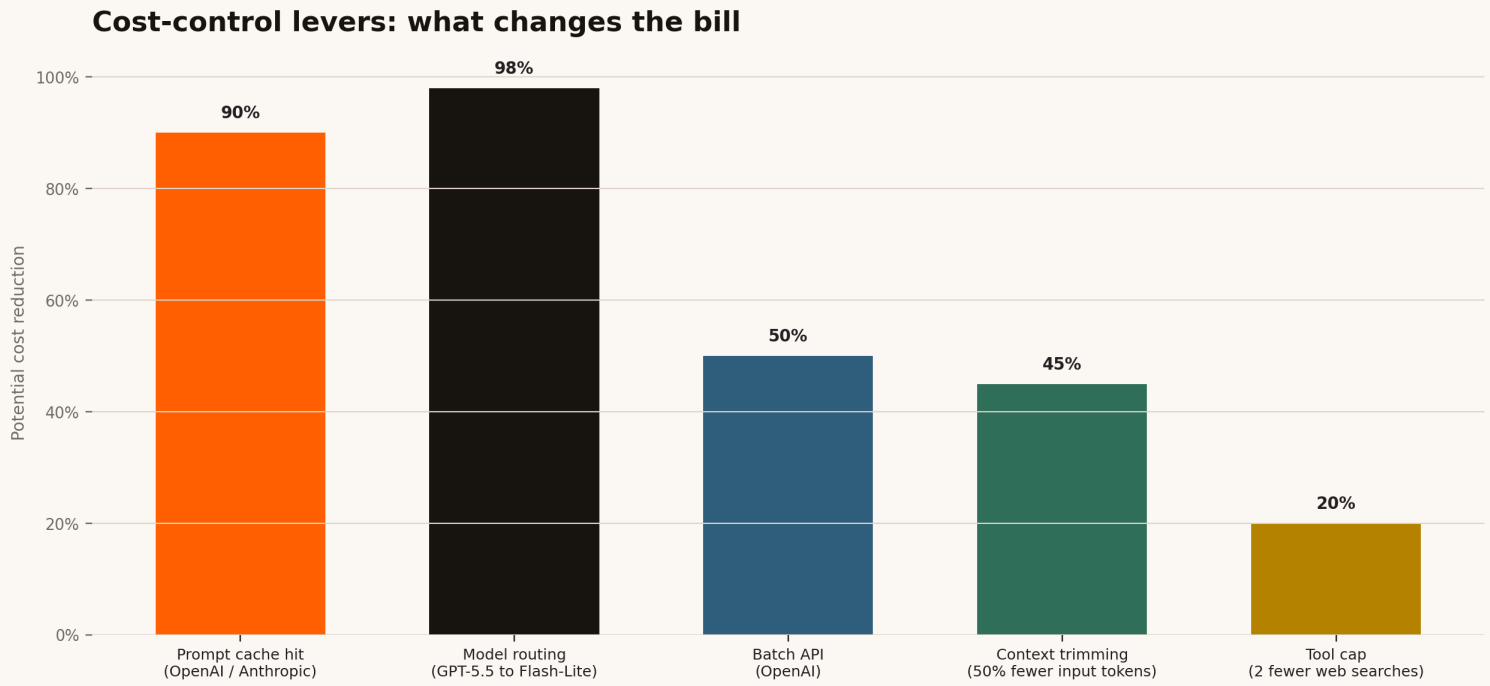
| Level | Name | What it means | Next move |
|---|---|---|---|
| 1 | Blind spend | The company sees the monthly AI bill, but not the cost per workflow. | Instrument usage by agent and workflow. |
| 2 | Basic tracking | Spend is tracked by provider or project. Unit economics are still unclear. | Add cost per outcome and owner accountability. |
| 3 | Workflow attribution | Cost is tied to teams, agents, customers, and workflows. | Add routing, caching, and budget limits. |
| 4 | Controlled scaling | Model routing, caching, batching, tool caps, and escalation rules exist. | Optimize cost per outcome. |
| 5 | AI cost advantage | The company uses cost discipline as a competitive advantage. | Refresh benchmarks quarterly and negotiate from data. |

Maturity model is a Cybernomics framework. Use it to prioritize governance and architecture work before scaling agents.

THE STRONGEST LEVERS ARE ARCHITECTURAL

# Cost control is designed into the workflow

Published discounts help, but the biggest leadership move is choosing the right operating pattern.




### Routing

Send easy work to cheaper models. Save premium models for hard work.

### Caching

Reuse stable instructions, policies, and context instead of repaying full input cost.

### Batching

Run reports, enrichment, QA, and cleanup in background mode where possible.

### Tool governance

Cap tool calls, retries, search, and container use before launch.

Savings chart combines published discounts with RACI benchmark calculations. Actual savings depend on workload and implementation.

90-DAY RACI IMPLEMENTATION ROADMAP

# Turn the report into an operating system

A leadership team can use this roadmap to move from analysis to governance.

1

### Days 1-15 - Instrument

Map agents, prompts, models, tool calls, token volume, and business outcomes.

2

### Days 16-30 - Baseline

Calculate cost per workflow and identify the top five cost-pressure areas.

3

### Days 31-60 - Control

Add model routing, prompt caching, context limits, tool caps, and budget alerts.

4

### Days 61-75 - Optimize

Shift non-urgent work to batch. Replace LLM steps with deterministic rules where possible.

5

### Days 76-90 - Govern

Create quarterly RACI reporting, owner accountability, and vendor decision rules.

**Executive output by day 90**

A RACI dashboard, cost-control policy, approved model-routing standard, quarterly refresh cadence, and a prioritized roadmap for reducing cost per business outcome.

EXECUTIVE RACI SCORECARD

# Six questions before approving an agent rollout

The answer should be clear before a workflow moves from pilot to production.

| Question | Green answer | Red flag |
|---|---|---|
| What is the cost per business outcome? | Known by workflow and reviewed monthly. | Only total vendor spend is known. |
| How are models selected? | Router uses task type, confidence, and risk. | One premium model does everything. |
| How is context controlled? | Context is filtered, capped, compressed, and cached. | Every transcript, file, and tool result is inserted. |
| How are tools controlled? | Allowlist, call caps, retry caps, and stop rules exist. | Agent can search, call, and retry freely. |
| What can run async? | Reports, enrichment, QA, cleanup, and indexing are batched. | Everything runs live in the user path. |
| Who owns unit economics? | Business owner and technical owner are named. | No owner for agent cost or quality. |

Cybernomics

THE CYBERNOMICS RACI ASSESSMENT

# Turn the benchmark into an executive decision system

The assessment helps companies find AI cost pressure before it becomes a budget problem.

### What we assess

AI model usage, agent workflows, prompt/context design, tool calls, retrieval architecture, governance controls, and cost-per-outcome metrics.

### What leaders receive

A RACI score by use case, cost pressure map, waste/risk findings, maturity score, vendor questions, and a 90-day control roadmap.

1. AI tool and model inventory
2. Agent workflow cost map
3. RACI score by use case
4. Waste and risk findings
5. 90-day cost-control roadmap
6. Executive summary for leadership

### Commercial use

Use this report as the opening asset. Use the RACI Assessment as the paid diagnostic. Use the maturity model as the roadmap for advisory work.

SOURCE APPENDIX AND ASSUMPTIONS

# Refresh these inputs quarterly

Pricing changes quickly. The RACI report should be refreshed before major budget or vendor decisions.

| Source | Validated facts used in this report |
|---|---|
| OpenAI API pricing | GPT-5.5: $5 input / $30 output per 1M tokens. GPT-5.4: $2.50 / $15. GPT-5.4 mini: $0.75 / $4.50. GPT-5.4 nano: $0.20 / $1.25. Web search: $10 / 1k calls. File search: $2.50 / 1k calls plus $0.10 / GB-day storage. Containers: from $0.03 per 20-minute 1GB session. Batch API: 50% discount on inputs and outputs. |
| Anthropic Claude pricing | Claude Opus 4.8: $5 input / $25 output. Sonnet 4.6: $3 / $15. Haiku 4.5: $1 / $5 per 1M tokens. Cache reads: 0.1x base input price. |
| Google Gemini pricing | Gemini 3.5 Flash: $1.50 input / $9 output. Gemini 2.5 Pro: $2.25 / $18 for prompts up to 200k tokens. Gemini 2.5 Flash: $0.30 / $2.50. Gemini 2.5 Flash-Lite: $0.10 / $0.40. Grounding prices use the published schedules. |
| Tavily API credits | Pay-as-you-go: $0.008 per credit. Monthly plans list lower per-credit ranges. |
| Pinecone pricing | Reranking: $2 / 1k requests. Sparse embedding: $0.08 / 1M tokens. |
| Cybernomics assumptions | Model benchmark = 10,000 agent steps x 10k input tokens + 2k output tokens. Scenario costs are explicit benchmark examples, not universal forecasts. |

Cybernomics

# The Relative AI Cost Index

A proprietary benchmark for AI agent economics, workflow design, and cost control.

Next step
Run the Cybernomics RACI Assessment before scaling AI agents across the business.

RACI

Cybernomics Research