Reasoning Tokens and Thinking Budgets: Cost vs Accuracy

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · reasoning tokens · thinking budget · llm pricing · extended thinking · chain of thought · budget forcing · openai reasoning effort · gemini thinking budget · inference cost · test-time compute

Executive Summary

Every major large language model (LLM) provider examined in this report, OpenAI, Anthropic, Google, and DeepSeek, now bills internal "reasoning tokens" (also called "thinking tokens") at the same per-token rate as ordinary visible output, while giving callers a provider-specific control, OpenAI's `reasoning.effort`, Anthropic's `budget_tokens` (now legacy on its newest models, being replaced by an adaptive `effort` parameter), Google's `thinkingBudget` for Gemini 2.5, and Google's `thinkingLevel` for Gemini 3 models, to set how much a model reasons before answering (ai.google.dev) (learn.microsoft.com) (docs.anthropic.com) (ai.google.dev). As of September 2026, official pricing spans roughly $4.40 to $600 per million output tokens across OpenAI's o-series and GPT-5 models (developers.openai.com), $10 per million tokens for Claude Sonnet 5 (docs.anthropic.com), $10.00 to $15.00 per million tokens for Gemini 2.5 Pro depending on prompt length (ai.google.dev), and $1.98 to $3.96 per million tokens off-peak versus peak for DeepSeek's flagship reasoning model (api-docs.deepseek.com). Reasoning-token volume itself can range from "a few hundred to tens of thousands" of tokens per OpenAI's own documentation (developers.openai.com).

Published research reports that this spend does not translate linearly into accuracy. Forcing a fine-tuned 32-billion-parameter model to reason longer raised its accuracy on the 2024 American Invitational Mathematics Examination (AIME24) from 50% to 57% at a forced 7,320 reasoning tokens (arxiv.org), and OpenAI's own o1 launch data showed accuracy on the same exam climbing from 74% to 93% as test-time compute scaled from one sample to 1,000 re-ranked samples (openai.com). But multiple independent studies also document "overthinking": accuracy that plateaus within a few hundred tokens on easier tasks and then measurably declines past a crossover point, with one paper placing that crossover "at ~7K" tokens on average and another finding Claude Opus 4 accuracy fall from near 100% to 85 to 90% under artificially extended reasoning (arxiv.org) (arxiv.org).

Reasoning also carries a latency and reliability cost rarely captured on a pricing page. Independent benchmarking by Artificial Analysis measured a roughly 35-fold increase in time to first token (TTFT) when reasoning was enabled on an otherwise identical Gemini 2.5 Flash deployment, 14.94 seconds versus 0.43 seconds (artificialanalysis.ai) (artificialanalysis.ai), and a repeated-measures study found that identical prompts under identical settings can still produce different answers and different costs run to run, even at temperature zero (arxiv.org).

The evidence assembled in this report argues that any credible cost-versus-accuracy claim about a reasoning model must specify the model, task, effort setting, and number of repeated runs behind it, rather than a single quoted per-token price or a single benchmark run. IntuitionLabs, a life sciences and AI consultancy, applies a version of this discipline in its own enterprise AI adoption work, measuring "time recovered, quality, risk signals, reliability, and support burden before expanding the investment" in a given AI workflow rather than relying on a headline benchmark figure alone (intuitionlabs.ai).

Introduction and Background

A growing share of large language model deployments now run on "reasoning" models: systems that generate an internal chain of intermediate steps before producing a final answer, and that bill for those internal steps as if they were part of the visible response. OpenAI, Anthropic, Google, and DeepSeek each expose a control, called a reasoning effort level, an extended thinking budget, or a thinking budget, that lets a caller trade more of these internal "reasoning tokens" for potentially higher accuracy on hard problems (learn.microsoft.com) (docs.anthropic.com) (ai.google.dev). Gemini 2.5 uses `thinkingBudget`; Gemini 3 uses `thinkingLevel`. Reasoning-token volume can vary with task complexity and the selected reasoning effort.

This report is an analyst-oriented reference for engineering and finance teams who need to reason about reasoning: what a reasoning token actually is, how each major provider exposes and bills it as of September 2026, and what published research says about the relationship between token budget and answer accuracy. Epoch AI, an independent AI research organization, dates the start of this era to OpenAI's release of o1-mini and o1-preview "in September 2024" (epoch.ai), and finds that since then, model capability on its benchmark index has advanced at more than double the rate seen in the preceding non-reasoning era: "the Epoch Capabilities Index (ECI) frontier has advanced linearly by 14 points per year, compared with 6 points per year for non-reasoning models" (epoch.ai). The underlying technique, letting a model write out intermediate reasoning steps before an answer, traces to Wei et al.'s widely cited 2022 paper, which showed that generating a chain of thought, a series of intermediate reasoning steps, "significantly improves the ability of large language models to perform complex reasoning" (arxiv.org).

Because reasoning tokens are billed at the same rate as visible output tokens by every major provider examined here, and because their volume varies with task complexity, cost and accuracy cannot be assessed from a per-token price alone. The sections below define the mechanics provider by provider, publish the method behind the accuracy findings so a reader could attempt to reproduce them, and separate vendor claims from independently measured benchmarks throughout.

Key Changes in How Providers Expose and Bill Reasoning Tokens

The four providers examined in this report each use the same core billing mechanic: reasoning tokens are billed as output tokens. They diverge sharply on how much control a caller has, whether the reasoning text itself is visible, and how the billing control is named. The subsections below document each provider's mechanism as of September 2026, citing only the provider's own current documentation.

OpenAI: Reasoning Effort Levels and the Hidden Token Line

OpenAI's documentation states that "reasoning models introduce reasoning tokens in addition to input and output tokens" (developers.openai.com), which the model uses to work through the prompt, considering multiple possible approaches before generating a response. Critically, OpenAI's guide is explicit that while reasoning tokens are not visible via the API, they still occupy space in the model's context window and are billed as output tokens. Microsoft's Azure OpenAI Service, which hosts the same underlying models for enterprise customers, documents identical behavior: "reasoning tokens never appear in the message content" and "reasoning tokens are billed as output tokens" (learn.microsoft.com) (learn.microsoft.com).

Callers control reasoning volume with a `reasoning.effort` parameter, mirrored on Azure as `reasoning_effort` (learn.microsoft.com). OpenAI documents that "supported values are model-dependent and can include none, minimal, low, medium, high, xhigh, and max," with lower effort favoring speed and lower token usage and higher effort favoring response quality (developers.openai.com). The company also states its models "reason adaptively across reasoning efforts, using fewer tokens for simpler tasks and thinking harder for complex tasks" (developers.openai.com), producing volume that may span anywhere from a few hundred to tens of thousands of reasoning tokens, with the exact count reported per call under `output_tokens_details`.

For latency-sensitive deployments, OpenAI recommends generating a short preamble before continuing with deeper reasoning, to reduce perceived time to first visible token, and separately documents a higher-effort "pro" reasoning mode "for difficult tasks that need more model work and can tolerate higher latency and token usage" (developers.openai.com). As of September 2026, OpenAI's official pricing table lists standard per-million-token rates spanning $1.10 to $150 for input and $4.40 to $600 for output across its o-series and GPT-5 family, with output pricing applying identically to visible and reasoning tokens.

Anthropic: Extended Thinking Budgets and the Shift to Adaptive Effort

Anthropic calls its mechanism extended thinking: the model's reasoning "arrives in thinking content blocks ahead of the response" (docs.anthropic.com). A `budget_tokens` parameter sets a target for how many tokens Claude can use for its internal reasoning process, subject to a hard floor of "1,024 tokens" below which "the API rejects smaller values" (docs.anthropic.com). Anthropic documents that this figure is a target rather than a ceiling: "the budget is a

target rather than a strict cap. Actual token usage varies with the task, and Claude may stop reasoning well before the budget is exhausted" (platform.claude.com), a source of the run-to-run token variance discussed later in this report.

As with OpenAI, Anthropic bills the full internal reasoning, the tokens Claude spends reasoning are billed as output tokens, and this holds even when the caller only receives a condensed view, because "the thinking text you receive is a summary of Claude's full thinking process" (docs.anthropic.com) while "the billed output token count does not match the visible token count" (docs.anthropic.com). A related interleaved thinking mode "lets Claude think between tool calls, reasoning about each tool result before acting on it" (docs.anthropic.com), useful for multistep agentic workflows.

Anthropic frames larger budgets as a diminishing-returns tradeoff: "higher budgets enable more comprehensive reasoning, with diminishing returns that depend on the task" (platform.claude.com) (docs.anthropic.com), and separately warns that "extended thinking adds latency and should only be used when it will meaningfully improve answer quality" (docs.anthropic.com), including a hard operational caveat that "pushing the model to think beyond 32k tokens produces long-running requests that can hit system timeouts and open-connection limits" (platform.claude.com). As of the September 2026 documentation used for this report, `budget_tokens` is Anthropic's legacy manual control; the company's newest Claude models have moved to an adaptive `effort` parameter in place of a fixed token target, a shift worth noting for any reader building against the API today. Current published API pricing lists Claude Sonnet 5 at $2 per million input tokens and $10 per million output tokens, the latter inclusive of thinking tokens.

Google Gemini: Thinking Budgets and Dynamic Thinking

Google describes Gemini 2.5 as "a thinking model, designed to tackle increasingly complex problems" (blog.google), built around an internal "thinking process" that "significantly improves their reasoning and multi-step planning abilities" (ai.google.dev). For Gemini 2.5, the `thinkingBudget` parameter guides the model on the specific number of thinking tokens to use for reasoning. Gemini 3 models instead use the recommended `thinkingLevel` parameter. For Gemini 2.5 models that support it, setting `thinkingBudget` to 0 disables thinking, while setting it to -1 turns on dynamic thinking (ai.google.dev). For Gemini 2.5 Pro specifically, Google's own parameter table documents a dynamic-thinking range of 128 to 32,768 tokens and notes the model cannot fully disable thinking, a constraint independently confirmed on Google Cloud's Vertex AI platform: "thinking can't be turned off for Gemini 2.5 Pro" (cloud.google.com). Vertex AI's own documentation describes the parameter as one that "sets an upper limit on the number of tokens the model can use for its thought process" (cloud.google.com).

Billing follows the same pattern as OpenAI and Anthropic: "when thinking is turned on, response pricing is the sum of output tokens and thinking tokens," with usage reported via "the `thoughtsTokenCount` field" (ai.google.dev). Google is unusually explicit that this applies even when only a condensed thought summary is shown to the caller: "pricing is based on the full thought tokens the model needs to generate to create a summary, despite only the summary being output from the API" (ai.google.dev). Vertex AI's pricing page lists a single combined line item, "text output (response and reasoning)" (cloud.google.com), and the developer-facing Gemini API pricing page lists that combined output price, including thinking tokens, at $10.00 per million tokens for prompts up to 200,000 tokens and $15.00 above that threshold, against $1.25 to $2.50 for input.

Google's guidance ties budget size to task difficulty, recommending "a high thinking budget" for "complex math problems or coding tasks" (ai.google.dev), while a low thinking level exists specifically because it "minimizes latency and cost" (ai.google.dev), a control Google frames explicitly around developer "latency constraints" (ai.google.dev), the same constraint-driven framing OpenAI and Anthropic use elsewhere in this report. Google's own Gemini 2.5 launch announcement separately claims the model "leads in math and science benchmarks like GPQA and AIME 2025" without added test-time techniques such as majority voting (blog.google), a vendor claim not independently verified in this report.

DeepSeek: Reasoning Content, Default-On Thinking, and Budget Forcing

DeepSeek's API returns internal reasoning in a dedicated `reasoning_content` field, documented as "the reasoning contents of the assistant message, before the final answer" (api-docs.deepseek.com), separately from the visible `content` field, with token accounting reported through a `reasoning_tokens` field defined as "tokens generated by the model for reasoning" (api-docs.deepseek.com). Unlike OpenAI, Anthropic, and Google, DeepSeek documents an on-by-default thinking mode rather than an opt-in effort scale: "thinking mode is enabled by default, with the default effort being high". Billing again follows the reasoning-as-output convention, since DeepSeek states "we will bill based on the total number of input and output tokens" (api-docs.deepseek.com), with current official pricing of $0.66 per million output tokens off-peak and $1.32 peak on one tier, and $1.98 off-peak and $3.96 peak on another, inside a documented context length of 1 million tokens and a maximum output cap of 384,000 tokens that bounds total reasoning plus visible generation (api-docs.deepseek.com).

DeepSeek's R1 model, and the broader open-reasoning research community around it, is also the primary setting for "budget forcing," a decoding-time technique described in the preprint literature as "forcing a maximum and/or minimum number of thinking tokens" (arxiv.org) by either truncating the model's reasoning with an end-of-thinking marker or lengthening it by repeatedly appending a continuation cue to the model's generation to force additional reasoning. Applied to a 32-billion-parameter fine-tuned model, s1-32B, researchers reported this technique "exceeds o1-preview on competition math questions by up to 27%" (arxiv.org), and reported a scaling relationship in which forcing more reasoning tokens raised AIME24 accuracy from 50% to 57%, a specific, load-bearing data point covered in more depth in the Data Analysis section below.

Table 1 below summarizes each provider's control parameter, documented range, billing treatment, and current pricing as of September 2026.

Provider	Control Parameter	Documented Range / Behavior	Reasoning Billed As	Reasoning Text Visible to Caller	Current Output Price (per 1M tokens, includes reasoning)
OpenAI	`reasoning.effort` (`reasoning_effort` on Azure)	none, minimal, low, medium, high, xhigh, max; a few hundred to tens of thousands of tokens generated	Output tokens	Hidden by default; count reported via `output_tokens_details`	$4.40 (o4-mini) to $600 (o1-pro)

Provider	Control Parameter	Documented Range / Behavior	Reasoning Billed As	Reasoning Text Visible to Caller	Current Output Price (per 1M tokens, includes reasoning)
Anthropic	`budget_tokens` (legacy); adaptive `effort` on newest models	Minimum 1,024 tokens; target, not a hard cap; no fixed maximum documented	Output tokens	Summarized by default; full token count still billed	$10 / MTok (Claude Sonnet 5)
Google Gemini	Gemini 3: `thinkingLevel`; Gemini 2.5: `thinkingBudget`	Gemini 3 uses level-based controls; Gemini 2.5 supports 128 to 32,768 for 2.5 Pro, which cannot disable thinking	Output tokens, combined line item	Optional thought summary only; full thought tokens still billed	$10.00 to $15.00 / MTok (Gemini 2.5 Pro, tiered by prompt length)
DeepSeek	Thinking mode on/off; default effort "high"	On by default; bounded by a 384K max-output cap	Output tokens, single combined bill	Visible by default via dedicated `reasoning_content` field	$1.98 to $3.96 / MTok, off-peak versus peak

Two patterns stand out. DeepSeek is the only provider among the four that surfaces the full reasoning text by default rather than hiding or summarizing it, a difference that matters for teams that need to audit a model's reasoning path rather than trust a black-box answer. And documented output pricing spans roughly two orders of magnitude, from DeepSeek's sub-$4 per million tokens to OpenAI's $600 per million tokens for its highest-effort o1-pro tier, before accounting for the token-volume differences discussed next.

Measuring the Cost-Accuracy Tradeoff: Method and Findings

A defensible measurement of the cost-accuracy tradeoff requires holding three things fixed and varying one: the same model, the same fixed set of task inputs, and the same scoring rubric, while varying only the reasoning-token budget or effort level, then repeating each condition multiple times because reasoning models are not fully deterministic even at a temperature of zero. The research reviewed for this report follows one of two designs. The first is budget forcing, where researchers intervene at decode time to cap or extend the number of tokens a model spends reasoning on identical prompts, then measure the resulting accuracy on a fixed benchmark, isolating token volume as the only variable. A separate vendor test-time-compute design changes the number of sampled traces and the selection method. At o1's launch, OpenAI compared a single sample, 64-sample consensus voting, and re-ranking of 1,000 samples on the same 15 AIME24 problems; this is not an isolated single-trace reasoning-token-budget sweep (openai.com).

Two unreviewed preprints, each limited to its reported benchmark scope, complicate that smooth-improvement narrative. A repeated-measures benchmark, ReasonBENCH, found that "the same model, strategy, and task can produce meaningfully different answers and costs across repeated executions, even under greedy decoding", and that the best-performing reasoning strategy in its test set "wins only 77% of head-to-head runs against its nearest

competitor" (arxiv.org), meaning a single benchmark run, of the kind many procurement comparisons rely on, can misrank two systems. A separate reliability study of coding tasks found that "run-level pass rate overstates retry-free coverage by up to 17.8 percentage points" (arxiv.org), a caution against reading a single reported accuracy figure as the rate a production deployment will actually achieve without retries.

Several unreviewed preprints report non-monotonic accuracy under their tested conditions, a phenomenon they term overthinking or inverse scaling in test-time compute. One paper places a crossover point at roughly 7,000 tokens on average, past which extended thinking becomes harmful, with easier problems crossing into net-harmful territory at far lower token counts than hard problems, while noting the underlying compute cost itself scales roughly linearly with token count. A second team found that forcing Claude Opus 4 to reason longer than necessary on a misleading-math overthinking task caused accuracy to fall from nearly 100% to around 85 to 90%, and found a substantially larger effect on DeepSeek R1, whose accuracy dropped "from 70% to 30%" once five distractor sentences were added to an otherwise simple counting task (arxiv.org). These findings are tabulated together in the Data Analysis and Evidence section below, alongside the specific benchmarks, models, and token counts each study reports, so that a reader could attempt to reproduce them.

Implementation Considerations and Process Changes

Teams adopting reasoning models for production workloads face a set of practical controls, not a single dial. Because every major provider bills reasoning tokens at the output rate while several hide the reasoning text itself, cost observability requires reading the correct usage field rather than the visible response length: OpenAI reports the count under `output_tokens_details`, Google under `thoughtsTokenCount`, and DeepSeek under a dedicated `reasoning_tokens` field alongside the full `reasoning_content` text. Teams that only log visible output length will systematically undercount spend, since billed and visible token counts can diverge.

Effort-level selection should be treated as a per-task-class configuration rather than a global default. OpenAI's own guidance to reserve high-effort or pro modes for tasks that can tolerate higher latency and token usage, and Google's parallel framing that low thinking levels exist to minimize latency and cost, both imply that a single effort setting applied uniformly across a heterogeneous workload will overspend on easy requests and potentially underperform on hard ones. Anthropic's explicit operational warning that budgets above 32,000 tokens risk system timeouts and open-connection limits is a concrete engineering constraint: streaming, retry, and timeout logic must be sized to the maximum configured thinking budget, not the median one.

Because reasoning-token volume, and resulting accuracy, vary run to run on identical inputs, a production evaluation pipeline should score multiple repeated runs per test case rather than a single pass, mirroring the coding-reliability study's distinction between single-run and retry-free pass rates. For regulated or audit-sensitive workflows, the kind IntuitionLabs advises life sciences organizations on, this argues for tracking quality, risk, and reliability signals as ongoing adoption metrics alongside raw token spend, rather than substituting a one-time benchmark score for continuous measurement, consistent with the consultancy's stated approach to "measure adoption, time recovered, quality, and support before scaling" a given AI workflow (intuitionlabs.ai).

Data Analysis and Evidence

Table 1 above established how each provider prices reasoning; the more consequential question for a technical buyer is what that spend buys in accuracy. *Table 2* below summarizes the reported token-budget-versus-accuracy findings referenced in the Measuring section above, each drawn from a named benchmark, model, and study so that a reader could attempt replication.

Study	Model(s)	Token Budget Manipulation	Reported Accuracy Result
s1: budget forcing	s1-32B (fine-tuned Qwen2.5-32B)	Budget forcing	The authors report that budget forcing scaled AIME24 performance from 50% to 57% and that s1-32B exceeded o1-preview on competition math by up to 27% (arxiv.org)
Phi-4-reasoning capacity study	Reasoning and quantized variants across basic-math tasks	Different reasoning budgets	The authors report that reasoning models generated about 18 times more tokens, sometimes with lower accuracy, and that constraining tokens could reduce accuracy by up to about 36% (arxiv.org)
Mirage of test-time scaling	Multiple models and benchmarks	Extended thinking compared with parallel thinking	The authors report initial performance improvements from additional thinking followed by a decline due to overthinking (arxiv.org)
Token budget saturation	DeepSeek-R1-Distill-Qwen-7B	Termination within a token budget versus budget exhaustion	The authors report a 62.0% overall convergence rate; converged generations reached 90.3% accuracy on AIME 1983–2024, versus 6.6% for non-converged generations (arxiv.org)
Inverse scaling in test-time compute	Large reasoning models across four task categories	Extended reasoning length	The authors report evaluation tasks in which extending reasoning length deteriorated performance, yielding an inverse relationship between test-time compute and accuracy (arxiv.org)
OpenAI o1 launch test-time-compute benchmark	OpenAI o1 vs GPT-4o	Number of complete samples and selection method: single sample, 64-sample consensus, or 1,000-sample reranking	AIME24 accuracy: GPT-4o 12%, o1 single-sample 74%, 64-sample consensus 83%, 1,000-sample re-ranked 93%; not a single-trace token-budget curve (openai.com)

Read across five cited studies and one OpenAI vendor benchmark, two conclusions recur despite different models, benchmarks, and manipulation methods. First, on problems below a model's natural difficulty threshold, additional reasoning tokens produce fast, then flat, then negative returns, a pattern the token budget saturation and mirage of test-time scaling studies both quantify with accuracy peaks well under 2,000 tokens (arxiv.org) (arxiv.org). Second, on problems above that threshold, deliberately forcing more reasoning tokens can still buy real accuracy in the s1 result, but only up to a point; OpenAI's o1 result instead illustrates test-time sampling and reranking, not a direct single-trace token-budget curve. No study reviewed here reports a universal optimal token budget; each finds a task-dependent one, which is a large part of why vendor documentation from OpenAI, Anthropic, and Google converges on adaptive, per-task effort controls rather than a single fixed setting.

Latency compounds the tradeoff. *Table 3* below reports independently measured time-to-first-token figures from Artificial Analysis, a benchmarking firm that publishes its measurement methodology, defining TTFT for a reasoning model as measured "through the first reasoning token" (artificialanalysis.ai), comparing reasoning and non-reasoning variants of the same base models.

Model	Mode	Measured Time to First Token (TTFT)
Gemini 2.5 Flash	Non-reasoning	0.43 seconds
Gemini 2.5 Flash	Reasoning enabled	14.94 seconds
OpenAI o3	Reasoning (default)	6.34 seconds
Gemini 2.5 Flash-Lite	Non-reasoning	0.29 seconds, the lowest TTFT Artificial Analysis measured across its full model set

Enabling reasoning on an otherwise identical Gemini 2.5 Flash deployment raised measured TTFT by roughly 35 times in Artificial Analysis's benchmark, 14.94 seconds versus 0.43 seconds. OpenAI's o3 measured 6.34 seconds to first token on the same methodology (artificialanalysis.ai), and the lowest-latency model Artificial Analysis measured across its entire catalog was a non-reasoning model, at 0.29 seconds (artificialanalysis.ai). This is a latency cost that does not appear on a per-token pricing page and that vendor documentation describes only qualitatively; Anthropic advises callers to "expect longer response times when thinking is active, because generating thinking blocks adds processing time" (docs.anthropic.com), and OpenAI's preamble recommendation exists specifically to offset it. The underlying technique is decades younger than the cost it now carries: Wei et al.'s original 2022 paper demonstrated that intermediate reasoning steps improve accuracy using a fraction of the compute a modern reasoning model consumes on a comparable problem, meaning the core idea has not changed even as its price has grown by orders of magnitude.

Implications and Future Directions

The clearest implication for buyers is that a per-token price alone does not describe the cost of a reasoning-model workload; effective cost is price multiplied by a token volume that varies with task complexity and can vary run to run even holding the parameter fixed. Procurement and finance processes built around a single quoted per-million-token rate should instead model a distribution of likely reasoning-token counts per task class, using each provider's documented range, a few hundred to tens of thousands for OpenAI, 128 to 32,768 for Gemini 2.5 Pro, a 1,024-token floor for Claude, as bounds.

The overthinking findings summarized above point toward a second-order engineering shift already visible in vendor roadmaps: away from a single, manually set token budget and toward adaptive, per-request effort selection, the direction OpenAI's adaptive reasoning behavior, Anthropic's newer effort parameter, and Google's Gemini 3 `thinkingLevel` control and Gemini 2.5 dynamic `thinkingBudget = -1` each represent (ai.google.dev). Independent research on parallel test-time strategies, such as running several shorter reasoning traces and combining them by majority vote instead of extending one long trace, reports accuracy gains of "up to 20% higher accuracy compared to extended thinking" at a comparable token budget (arxiv.org), suggesting that how a fixed reasoning-token budget is spent, not only how large it is, will remain an active area of both vendor and independent research through the remainder of 2026.

For enterprise buyers in regulated industries, where a wrong or inconsistent answer carries compliance exposure beyond its raw error rate, the run-to-run variance documented in independent reliability studies argues for evaluation processes that score repeated executions rather than a single demonstration run. IntuitionLabs, a life sciences and AI consultancy, applies a comparable discipline in its own AI Acceleration Program for life sciences organizations, which is built around implementing "governed information, specialist implementation, role-based adoption, and measured results" for one workflow at a time rather than committing budget on the strength of a single benchmark (intuitionlabs.ai). As reasoning-capable models become the default rather than the exception across major providers, the operative cost question shifts from what a token costs to how many tokens a given task will actually consume, how reliably, and at what latency, a question current per-token pricing pages do not answer on their own.

Frequently Asked Questions (FAQs)

What are reasoning tokens? Reasoning tokens are the internal, model-generated tokens a reasoning model produces while working through a problem before it writes its final answer; OpenAI defines them as tokens the model uses to "think," in addition to standard input and output tokens. Every major provider examined in this report, OpenAI, Anthropic, Google, and DeepSeek, bills these tokens at the same rate as ordinary output tokens.

What is a thinking budget or reasoning effort setting? It is a caller-set parameter, named `reasoning.effort` at OpenAI, `budget_tokens` (legacy) or an adaptive `effort` value at Anthropic, `thinkingBudget` for Gemini 2.5, and `thinkingLevel` for Gemini 3, that controls reasoning behavior before answering (ai.google.dev). DeepSeek instead exposes a simpler on-by-default thinking mode with a documented effort level.

How many reasoning tokens does a model actually use? OpenAI documents a range of a few hundred to tens of thousands of reasoning tokens depending on task difficulty. Independent studies report specific averages on fixed benchmarks; for example, roughly 379 tokens for a non-reasoning model versus 6,066 for its reasoning-tuned counterpart on the same task set, a 16-fold difference (arxiv.org).

What is budget forcing? Budget forcing is a research technique, not a production feature offered by any major vendor, that manipulates a model's reasoning length at decode time by truncating or extending its internal chain of thought. It has been used to demonstrate that AIME24 accuracy for one fine-tuned model rises from 50% to 57% as forced reasoning length increases to 7,320 tokens.

Do more reasoning tokens always improve accuracy? No. Multiple independent studies find accuracy plateaus, and in some cases declines, past a task-dependent token threshold, a pattern researchers call overthinking or inverse scaling in test-time compute. The Data Analysis and Evidence section above tabulates six such findings across different models and benchmarks.

How should reasoning-token cost be measured for a real workload? By holding the model, task set, and scoring rubric fixed, varying only the effort or budget setting, and running each condition multiple times rather than once, because both token counts and final answers can vary run to run even under identical settings.

Conclusion

Reasoning tokens have become the least transparent line item in LLM pricing precisely because every major provider treats them identically at the billing layer, as output tokens, while diverging in how much of that spend a caller can see or control. OpenAI, Anthropic, Google, and DeepSeek each expose a distinct parameter for the same underlying tradeoff, and each documents, in its own words, that larger budgets raise cost and latency with accuracy gains that are real but bounded and task-dependent.

The reviewed studies and vendor results provide measurements within their reported evaluation scopes: reasoning-token volume can vary 16-fold on an identical task depending on whether a reasoning mode is engaged at all, accuracy can rise sharply and then fall past a crossover point that arrives sooner for easy problems than hard ones, and the same model, prompt, and settings can still produce a different token count and a different answer on repeated runs. None of this makes reasoning models a poor choice; OpenAI's own launch data for o1 reports an accuracy gain from added test-time compute on hard math problems. It does mean that a defensible cost or accuracy claim about a reasoning model needs to specify the model, the task, the effort setting, and the number of repeated runs behind it, exactly the discipline this report's method section lays out.

For organizations evaluating whether and how to deploy reasoning models in cost-sensitive or compliance-sensitive workflows, the practical takeaway is to budget for the reasoning-token distribution a workload will actually produce, not the headline per-token price, and to measure accuracy and reliability across repeated runs before committing a workflow to a fixed effort setting.
