Vertere AI

COMPARATIVO DE MODELOS DE IA
NOVEMBRO DE 2026

INTELIGÊNCIA · CUSTO · VELOCIDADE · CONFIABILIDADE

Qual modelo de IA escolher: o que os números significam

19 modelos de 10 empresas comparados em inteligência, velocidade, custo e confiabilidade, em linguagem simples.

0,57 *ou* 26,54

Dólares por tarefa: o modelo mais barato e o mais caro entre os 19.
Artificial Analysis v5.0.1.

Baseado em: Artificial Analysis v5.0.1 e páginas oficiais de preço, 19/11/2026
Análise: Vertere AI · novembro de 2026

01 Em uma página

Não existe um modelo melhor em tudo. O mais inteligente, Claude Fable 5.2 (64,6 pontos), custa US$ 26,54 por tarefa, 6,6 vezes o GPT-6.2 Astra (62,1 pontos, US$ 4,00). Entre os modelos de nível alto, o melhor custo-benefício é o GPT-6.2 Sol: 57,3 pontos por US$ 1,01 por tarefa.

19

modelos de 10 empresas, medidos pela mesma régua.

3 pts

é a distância do 1º ao 3º em inteligência (Fable 5.2, Astra e Neon).

89%

da inteligência do Fable no GPT-6.2 Sol, por 3,8% do custo por tarefa.

47x

de diferença entre o custo por tarefa mais alto e o mais baixo.

1. **Mais inteligente: Claude Fable 5.2.** 64,6 pontos, mas US$ 26,54 por tarefa, 6,6 vezes o Astra.

2. **Melhor custo-benefício: GPT-6.2 Sol.** 57,3 pontos por US$ 1,01 por tarefa, e conclui a tarefa mais rápido (8,9 min).

3. **Orçamento mínimo:** Qwen3.8-Flash-Next‡, DeepSeek V4.1 Flash† e GLM-5.3-Flash custam de US$ 0,57 a US$ 0,60 por tarefa, mas têm índice de 28 a 33 e alucinação de 38% a 64%.

4. **Mais confiável em fatos: a família Gemini 4.** Alucinação de 4% a 14%, contra 59% do Fable e do Astra.

5. **Cuidado com o preço por token.** Ele não diz quanto você vai pagar: o custo depende de quantos tokens o modelo gasta pensando (seção 07).

02 Por que existe e como ler

Escolher um modelo de IA é uma decisão de custo e de qualidade ao mesmo tempo, e as opções mudam quase toda semana: 9 dos 19 modelos foram lançados nos 30 dias anteriores à consulta. Preço por token e nota de benchmark, isolados, enganam: o modelo mais barato por token pode ser o mais caro para concluir uma tarefa. Aqui estão nota, velocidade, custo e confiabilidade, com o significado de cada número.

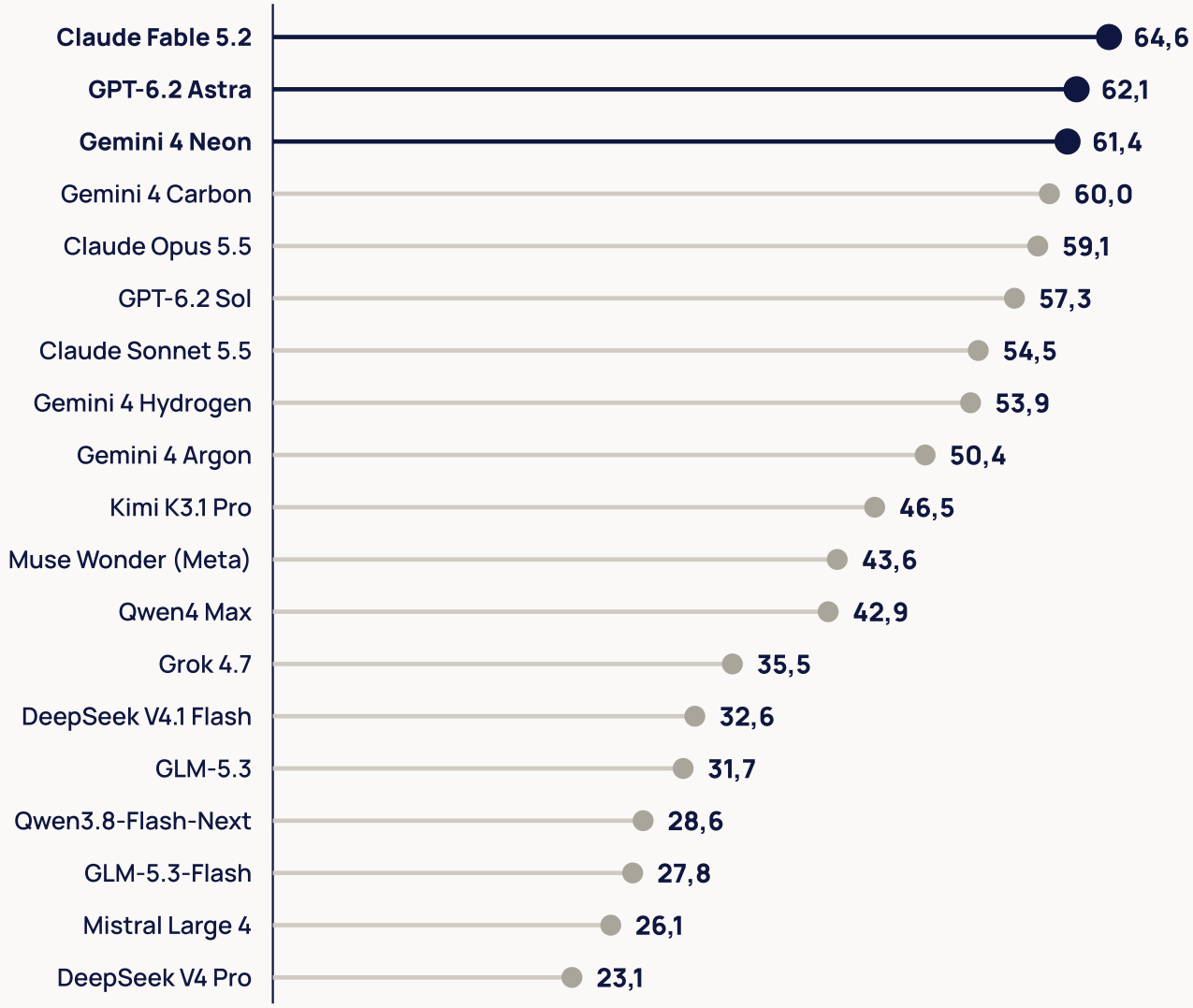
A Artificial Analysis (AA) mede modelos de IA com as mesmas provas, de forma independente. Todos os números são de 19/11/2026, no maior nível de esforço de raciocínio que a AA lista para cada modelo: o teto de qualidade, de custo e de espera. Preços de API vêm das páginas oficiais dos fabricantes, no mesmo dia. Interpretações da Vertere AI aparecem como **leitura**. Nomes de modelos e de testes ficam no original.

INDICADOR	O QUE MEDE, EM LINGUAGEM SIMPLES	POR QUE IMPORTA PARA A EMPRESA
Índice de Inteligência	Nota de 0 a 100 que resume 9 provas da AA: trabalho de escritório com agentes, automação, código, conhecimento, raciocínio e ciência.	Resume quanto o modelo resolve sozinho. Nota maior tende a significar menos retrabalho humano.
Terminal-Bench 4.0 (%)	Tarefas reais de programação e administração de sistemas feitas num terminal, com o modelo agindo como agente.	Indica se o modelo executa trabalho técnico de ponta a ponta, não só escreve trechos de código.
FrontierCode-AA 2.0 (%)	Engenharia de software em repositórios de código aberto, com problemas reais e difíceis.	Mede a capacidade de mexer em sistemas grandes. É a prova mais exigente, por isso as notas são menores.
AutomationBench-AA (%)	Fluxos de trabalho em sistemas de empresa (SaaS), com pontuação parcial por etapa cumprida.	Parece-se com automatizar CRM, planilhas, tickets e e-mail, o uso mais comum de agentes.
Velocidade de escrita (tokens por segundo)	Quantos tokens o modelo escreve por segundo depois de começar. Mil tokens equivalem a cerca de 750 palavras em inglês.	Quanto maior, mais rápido o texto aparece. Pesa em chat com cliente e textos longos.
Espera até a resposta (TTFT, s)	Tempo entre pedir e ver o início da resposta final. Nos modelos que pensam no esforço máximo, inclui o raciocínio.	É a espera percebida. De 200 a 340 s indica que o modelo pensa por minutos antes de responder.
Custo por tarefa (US$)	O que custa rodar uma tarefa do índice: tokens de entrada, saída e raciocínio, com cache, vezes os tokens realmente usados.	É o número mais honesto para orçamento, porque junta preço e consumo.
Preço por 1 milhão de tokens (entrada / saída)	Tabela de preço do fabricante: o que custa cada milhão de tokens enviados (entrada) e escritos pelo modelo (saída).	Serve para comparar fornecedores e tarefas simples, mas não mostra quantos tokens o modelo vai gastar.
Taxa de alucinação (%)	Em perguntas de conhecimento, a frequência com que o modelo erra em vez de admitir que não sabe. Menor é melhor.	Mede o risco de inventar com segurança. Crítico em jurídico, financeiro e atendimento.

03 Quem lidera e por quê

A inteligência geral tem um líder claro e um segundo grupo quase empatado. Em código e agentes, não há campeão único.

Claude Fable 5.2 lidera, com um segundo grupo logo atrás



Fonte: Artificial Analysis v5.0.1, 19/11/2026; 19 modelos, no maior esforço de raciocínio listado pela AA.

Inteligência geral: Claude Fable 5.2 (64,6). Lidera também o FrontierCode (48,8%) e tem a maior taxa de acerto em conhecimento (79%). Leitura: vence nas tarefas mais difíceis e longas. O custo é pensar muito: 273.776 tokens de saída por tarefa, contra 41.029 do Astra.

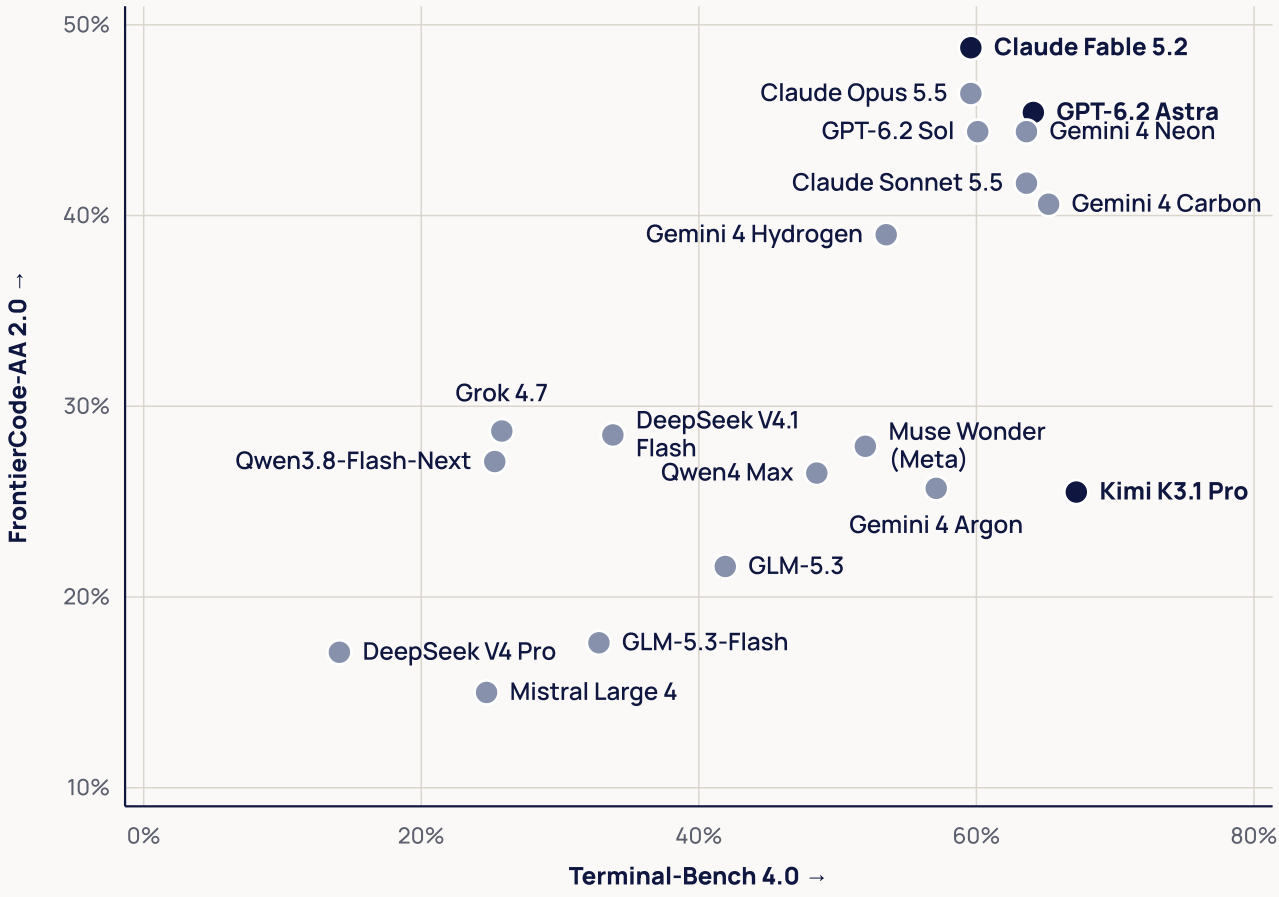
Segundo grupo, quase empatado: GPT-6.2 Astra, Gemini 4 Neon, Gemini 4 Carbon e Claude Opus 5.5 ficam de 2,5 a 5,5 pontos do líder. Para a maioria das empresas, esse tamanho de diferença não decide sozinho.

Modelos abertos e de outros países: pesos abertos são os que a empresa pode rodar na própria infraestrutura: Kimi K3.1 Pro, Qwen3.8-Flash-Next, DeepSeek V4.1 Flash, DeepSeek V4 Pro, Mistral Large 4, GLM-5.3 e GLM-5.3-Flash. O melhor deles é o Kimi K3.1 Pro, com 46,5 pontos, 72% do Fable. Qwen4 Max (42,9) e Muse Wonder (Meta) (43,6) vêm logo atrás, mas não são de pesos abertos segundo a AA. DeepSeek, GLM e Mistral ficam entre 23 e 33 pontos.

Código e agentes: o melhor depende do tipo de tarefa

Kimi K3.1 Pro lidera o Terminal-Bench (67,2%), à frente de Gemini 4 Carbon (65,2%) e GPT-6.2 Astra (64,1%). No FrontierCode, a ordem é Fable (48,8%), Opus (46,4%) e Astra (45,4%). O Kimi vai bem em tarefas de terminal e mal no FrontierCode (25,5%): bom em código depende do tipo de tarefa. Em automação de sistemas, Gemini 4 Neon lidera o AutomationBench (79,3%), com Carbon, Argon e Astra entre 77% e 78%.

Quem vai bem em um teste de código nem sempre vai bem no outro

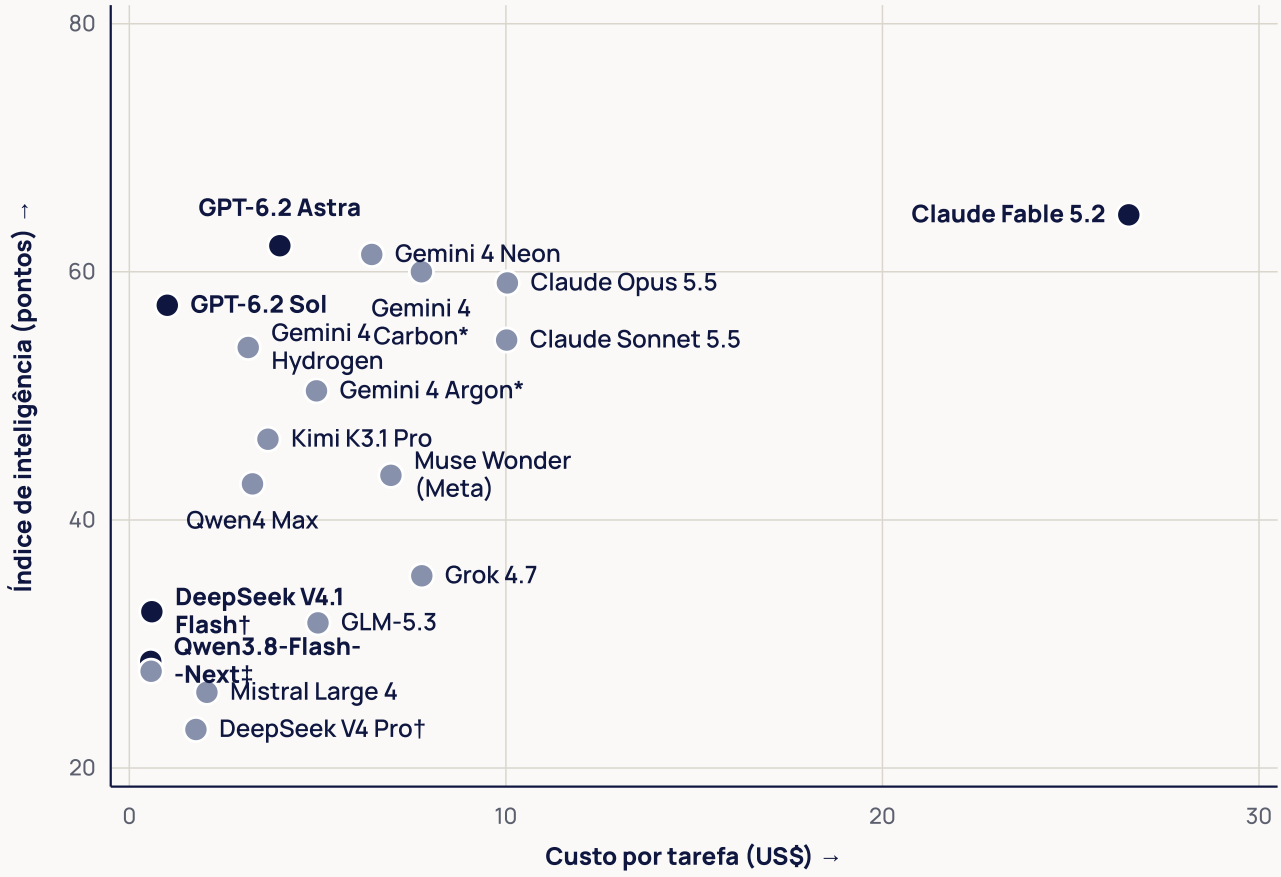


Fonte: Artificial Analysis v5.0.1, 19/11/2026; 19 modelos, no maior esforço de raciocínio listado pela AA. Terminal-Bench 4.0 e FrontierCode-AA 2.0 substituem os antigos Coding Index e Agentic Index, que a AA não publica mais.

04 Inteligência por dólar

Só 5 dos 19 modelos não perdem para nenhum outro em inteligência e em custo por tarefa ao mesmo tempo: Qwen3.8-Flash-Next‡, DeepSeek V4.1 Flash†, GPT-6.2 Sol, GPT-6.2 Astra e Claude Fable 5.2. Os outros 14 perdem para algum deles nos dois critérios.

A fronteira eficiente: custo por tarefa contra inteligência



Fonte: Artificial Analysis v5.0.1, 19/11/2026; 19 modelos, no maior esforço de raciocínio listado pela AA. Em azul-marinho, os modelos da fronteira. O ponto sem nome, quase sobre o Qwen3.8-Flash-Next, é o GLM-5.3-Flash (US$ 0,58; índice 27,8). * Gemini 4 Carbon e Argon: preço promocional do Google até 30/11/2026; depois, US$ 4 / US$ 20. † DeepSeek: preço de horário de pico. ‡ Qwen3.8-Flash-Next: correspondência com a página oficial não confirmada. Notas completas na seção 10.

Quanto custa subir de nível: do Sol ao Astra, são US$ 2,99 a mais por tarefa para ganhar 4,8 pontos. Do Astra ao Fable, US$ 22,54 a mais para ganhar 2,5 pontos.

Como ler: três dos quatro primeiros (DeepSeek V4.1 Flash, Qwen3.8-Flash-Next e GLM-5.3-Flash) têm índice abaixo de 33: a razão alta vem do custo baixo, não da inteligência. Entre os de índice acima de 50, o GPT-6.2 Sol lidera (56,7), seguido de Gemini 4 Hydrogen (17,1), GPT-6.2 Astra (15,5) e Gemini 4 Argon (10,1).

Ranking de pontos de índice por dólar de custo por tarefa

#	Modelo	Pontos por US$	Índice	Custo/tarefa
1	**GPT-6.2 Sol**	**56,7**	57,3	US$ 1,01
2	DeepSeek V4.1 Flash†	**54,3**	32,6	US$ 0,60
3	Qwen3.8-Flash-Next‡	**50,2**	28,6	US$ 0,57
4	GLM-5.3-Flash	**47,9**	27,8	US$ 0,58
5	Gemini 4 Hydrogen	**17,1**	53,9	US$ 3,16
6	GPT-6.2 Astra	**15,5**	62,1	US$ 4,00
7	Qwen4 Max	**13,1**	42,9	US$ 3,27
8	DeepSeek V4 Pro†	**13,1**	23,1	US$ 1,77
9	Mistral Large 4	**12,7**	26,1	US$ 2,06
10	Kimi K3.1 Pro	**12,6**	46,5	US$ 3,68
11	Gemini 4 Argon*	**10,1**	50,4	US$ 4,97
12	Gemini 4 Neon	**9,5**	61,4	US$ 6,44
13	Gemini 4 Carbon*	**7,7**	60,0	US$ 7,76
14	GLM-5.3	**6,3**	31,7	US$ 5,01
15	Muse Wonder (Meta)	**6,3**	43,6	US$ 6,95
16	Claude Opus 5.5	**5,9**	59,1	US$ 10,04
17	Claude Sonnet 5.5	**5,4**	54,5	US$ 10,02
18	Grok 4.7	**4,6**	35,5	US$ 7,77
19	Claude Fable 5.2	**2,4**	64,6	US$ 26,54

Fonte: Artificial Analysis v5.0.1, 19/11/2026; 19 modelos, no maior esforço de raciocínio listado pela AA. Conta da Vertere AI (leitura): índice dividido pelo custo por tarefa em US$; quanto maior, mais inteligência por dólar. * Gemini 4 Carbon e Argon: preço promocional do Google até 30/11/2026; depois, US$ 4 / US$ 20. † DeepSeek: preço de horário de pico. ‡ Qwen3.8-Flash-Next: correspondência com a página oficial não confirmada. Notas completas na seção 10.

05 Melhor modelo por categoria e por uso

CATEGORIA	MELHOR	VALOR	OBSERVAÇÃO
Inteligência geral	Claude Fable 5.2	64,6	2,5 pontos à frente de GPT-6.2 Astra
Código e agentes (média de Terminal-Bench e FrontierCode)	GPT-6.2 Astra	54,8%	Fable 5.2 fica em 54,2%. Terminal-Bench puro: Kimi K3.1 Pro (67,2%). FrontierCode puro: Fable 5.2 (48,8%)
Automação de sistemas (AutomationBench)	Gemini 4 Neon	79,3%	Carbon, Argon e Astra a menos de 2,1 pontos
Velocidade de escrita	Gemini 4 Hydrogen	170 tok/s	DeepSeek V4.1 Flash quase igual (165)
Menor tempo por tarefa	GPT-6.2 Sol	8,9 min	Claude Fable 5.2 é o mais demorado (41,6 min)
Menor custo por tarefa	Qwen3.8-Flash-Next‡	US$ 0,57	Empate técnico com GLM-5.3-Flash (US$ 0,58) e DeepSeek V4.1 Flash (US$ 0,60). Com índice acima de 50: GPT-6.2 Sol
Melhor custo-benefício	GPT-6.2 Sol	56,7 pontos por US$	1º entre os de índice acima de 50 (seção 04)
Menor alucinação	Gemini 4 Argon	4%	Hydrogen 8%, Carbon 11%, Neon 14%

Qual modelo para qual uso

Leitura da Vertere AI a partir dos números. Teste com os seus dados antes de decidir.

CASO DE USO	ESCOLHA	POR QUE, COM OS NÚMEROS
Triagem e resposta de e-mails e WhatsApp de suporte, volume alto, perguntas repetidas	GPT-6.2 Sol ou Gemini 4 Hydrogen	GPT-6.2 Sol: US$ 1,01 por tarefa, índice 57,3. Gemini 4 Hydrogen: o mais rápido (169,5 tok/s) e só 8% de alucinação. Evite os modelos baratos de índice baixo se um erro de fato custar caro (alucinação de 38% a 64%).
Classificação de leads e extração de dados de notas e contratos em grande volume	Qwen3.8-Flash-Next, DeepSeek V4.1 Flash e GLM-5.3-Flash	US$ 0,57 a US$ 0,60 por tarefa. Índice 28 a 33: serve para tarefas bem definidas, com revisão por regra ou por amostragem.
Agente que corrige bugs e mexe em sistemas internos	GPT-6.2 Astra (equilíbrio) ou Claude Fable 5.2 (máxima qualidade)	GPT-6.2 Astra: Terminal-Bench 64,1%, FrontierCode 45,4%, US$ 4,00 por tarefa. Claude Fable 5.2: FrontierCode 48,8%, mas US$ 26,54 por tarefa.
Automação de CRM, planilhas, tickets e fluxos entre sistemas	Gemini 4 Neon ou GPT-6.2 Astra	AutomationBench 79,3% e 77,2%. O Neon custa US$ 6,44 por tarefa e tem 14% de alucinação.

CASO DE USO	ESCOLHA	POR QUE, COM OS NÚMEROS
Revisão de contratos, análise financeira ou jurídica de documentos longos	Claude Fable 5.2 ou Gemini 4 Carbon* / Neon, sempre com revisão humana	Fable tem o maior acerto (79%), mas chuta mais (59%); os Gemini erram menos ao chutar (11% a 14%) e o contexto é de 1 milhão de tokens. Nenhuma prova do índice é específica de direito brasileiro.
Assistente interno de consulta rápida, com perguntas e respostas ao vivo	Gemini 4 Hydrogen ou Muse Wonder (Meta)	Respostas começam cedo (espera de 45 s e 20 s no esforço máximo; tende a ser menor com esforço menor) e escrevem rápido (169,5 e 117 tok/s). Evite os Claude e o GPT-6.2 Astra em chat ao vivo no esforço máximo (espera de 207 a 339 s).
Dados sensíveis que não podem sair da empresa	Kimi K3.1 Pro (ou DeepSeek, GLM ou Mistral Large 4) em infraestrutura própria	São de pesos abertos. Kimi tem o maior índice entre eles (46,5); os outros ficam entre 23 e 33.
Orçamento fixo e pouco volume, mas tarefas difíceis	Claude Fable 5.2 só nos casos críticos, GPT-6.2 Astra ou GPT-6.2 Sol no resto	1.000 tarefas como as do índice custariam cerca de US$ 26,5 mil no Fable, US$ 4 mil no Astra e US$ 1 mil no Sol (ilustração, não é previsão).

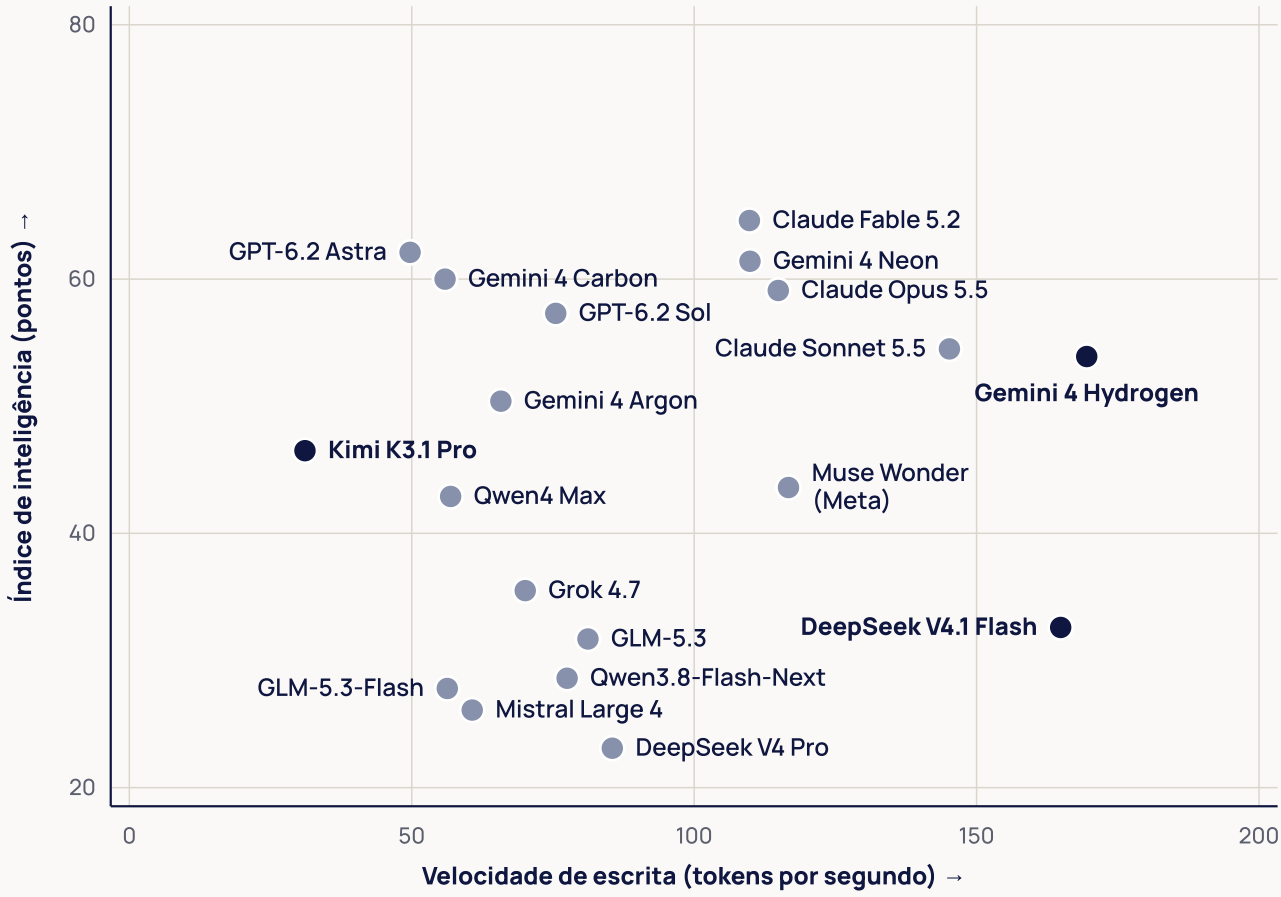
O que você ganha e a que renuncia

SE VOCÊ ESCOLHE	GANHA	ABRE MÃO DE
Claude Fable 5.2	Maior inteligência, melhor código difícil (FrontierCode), maior acerto em conhecimento (79%)	Custo (US$ 26,54 por tarefa), espera longa (TTFT de 222 s; 41,6 min por tarefa), tendência a chutar (59% de alucinação)
GPT-6.2 Astra	Quase a inteligência do líder por 15% do custo (US$ 4,00); boa em Terminal-Bench e AutomationBench	Escreve devagar (50 tok/s) e demora a começar (TTFT 207 s)
GPT-6.2 Sol	Melhor custo-benefício, mais rápido por tarefa, preço baixo (US$ 2 / 10)	4,8 pontos a menos que o Astra; alucinação de 55%
Gemini 4 Neon e Carbon	Perto do topo, menor alucinação entre os grandes (14% e 11%), AutomationBench alto	O Carbon é lento (56 tok/s) e seu preço dobra em 01/12/2026; o Neon custa US$ 6,44 por tarefa
Gemini 4 Hydrogen	Mais rápido de todos (169,5 tok/s), alucinação de 8%, índice 53,9	Custa US$ 3,16 por tarefa, 3 vezes o do Sol
Kimi K3.1 Pro	Melhor Terminal-Bench, pesos abertos, preço baixo por token (US$ 2 / 10)	Mais lento (31 tok/s) e fraco no FrontierCode (25,5%)
Modelos baratos de DeepSeek, Qwen e GLM (Flash)	Custo por tarefa de US$ 0,57 a US$ 0,60	Índice de 28 a 33 e alucinação de 38% a 64%
Grok 4.7	Contexto de 500 mil tokens e preço baixo por token (US$ 2 / 6)	Índice de 35,5: 9 modelos são ao mesmo tempo mais inteligentes e mais baratos por tarefa

06 Velocidade e espera

O que o usuário sente é o tempo até concluir a tarefa (tempo total, incluindo o raciocínio), não só a velocidade de escrita. Gemini 4 Hydrogen (169,5 tok/s) e DeepSeek V4.1 Flash (165 tok/s) escrevem mais rápido; o mais lento é o Kimi K3.1 Pro (31 tok/s). Mas o menor tempo por tarefa é do GPT-6.2 Sol, 8,9 min, porque usa poucos tokens, apesar de escrever a 75,5 tok/s.

Velocidade de escrita não é velocidade de entrega



Fonte: Artificial Analysis v5.0.1, 19/11/2026; 19 modelos, no maior esforço de raciocínio listado pela AA.

Espera até a resposta: os Claude e o GPT-6.2 Astra têm TTFT de 207 a 339 s no esforço máximo, ou seja, pensam por minutos antes de responder, o que não serve para chat ao vivo. Os de resposta mais cedo são Muse Wonder (Meta) (20 s), DeepSeek V4.1 Flash (44 s) e Gemini 4 Hydrogen (45 s). Com esforço menor, espera, custo e nota caem juntos.

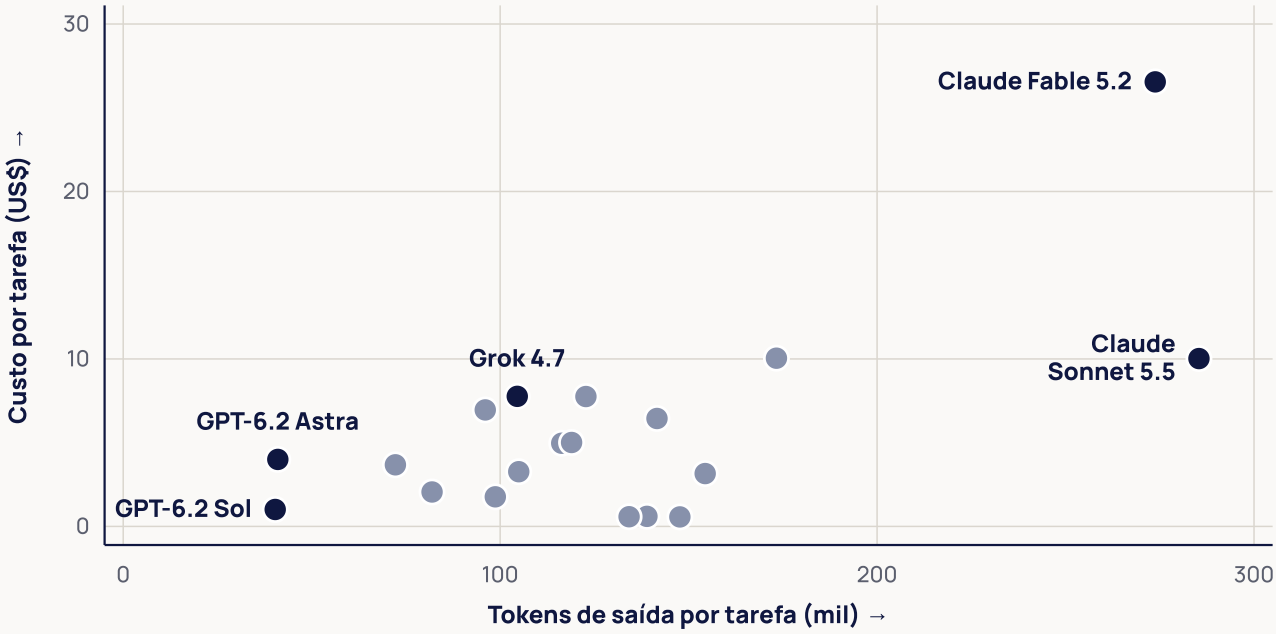
07 Preço por token e custo real por tarefa

O preço por 1 milhão de tokens é o preço do litro; o custo por tarefa é quanto você gasta no trajeto. Dois modelos com o mesmo preço por token podem ter custos muito diferentes, porque um pensa e escreve muito mais que o outro. A coluna de tokens de saída mostra quantos tokens o modelo escreve, incluindo o raciocínio, para resolver uma tarefa do índice.

MODELO	ENTRADA / SAÍDA POR 1 MI	PREÇO MISTO 3:1 (AA)	TOKENS DE SAÍDA POR TAREFA	CUSTO POR TAREFA
GPT-6.2 Sol	US$ 2 / 10	US$ 4	40.360	US$ 1,01
Claude Sonnet 5.5	US$ 2 / 10	US$ 4	285.387	US$ 10,02
GPT-6.2 Astra	US$ 8 / 40	US$ 16	41.029	US$ 4,00
Claude Fable 5.2	US$ 8 / 40	US$ 16	273.776	US$ 26,54
Claude Opus 5.5	US$ 4 / 20	US$ 8	173.313	US$ 10,04
Gemini 4 Hydrogen	US$ 1,50 / 7,50	US$ 3	154.406	US$ 3,16
Grok 4.7	US$ 2 / 6	US$ 3	104.531	US$ 7,77

Fonte: Artificial Analysis v5.0.1, 19/11/2026; preços oficiais dos fabricantes, 19/11/2026.

Quanto mais tokens, mais caro, e o preço por token multiplica



Fonte: Artificial Analysis v5.0.1, 19/11/2026; 19 modelos, no maior esforço de raciocínio listado pela AA. Só os modelos citados no texto têm nome.

- **Sol × Sonnet:** os dois custam US$ 2 / 10 por milhão. O Sonnet gasta 285.387 tokens de saída por tarefa e o Sol, 40.360 (7,1 vezes menos). Resultado: o Sonnet custa US$ 10,02 por tarefa e o Sol, US$ 1,01, quase 10 vezes mais, com inteligência menor.
- **Astra × Fable:** mesmo preço (US$ 8 / 40). O Fable gasta 273.776 tokens por tarefa contra 41.029 (6,7 vezes mais) e custa US$ 26,54 contra US$ 4,00.
- **Grok 4.7:** tem um dos menores preços por token entre os modelos de ponta (US$ 2 / 6), mas gasta tokens demais para a sua inteligência e custa US$ 7,77 por tarefa, mais que o Astra, que cobra de 4 a 6,7 vezes mais por token.
- **DeepSeek:** o preço oficial muda com o horário (fora do pico é metade). A tabela usa o preço de pico.
- **Cache e lote:** os fabricantes descontam entrada repetida (cache) e processamento em lote (50% de desconto na OpenAI, Anthropic e Google; no Kimi o lote custa 60% do preço padrão). A AA já considera cache no custo por tarefa, mas o lote não entra.

08 O que muda em 01/12/2026: Gemini 4 Carbon e Argon

O Google cobra US$ 2 / US$ 10 por milhão de tokens no Carbon e no Argon só até 30/11/2026. A partir de 01/12/2026, o preço de entrada e de saída dobra, para US$ 4 / US$ 20. A AA já lista o Carbon a esse valor futuro; o Argon ela lista ainda ao preço promocional. Por isso o custo por tarefa dos dois no ranking não está na mesma base.

Custo estimado por tarefa antes e depois de 01/12/2026



Fonte: Estimativa Vertere AI sobre Artificial Analysis v5.0.1, 19/11/2026; 19 modelos, no maior esforço de raciocínio listado pela AA. Pressupostos: custo proporcional ao preço (mistura 3:1 de entrada e saída), mesmo consumo de tokens e nenhum desconto de cache ou lote.

- **Gemini 4 Carbon:** a AA mede US$ 7,76 por tarefa, já a US$ 4 / US$ 20. Pago ao preço promocional até 30/11, o custo seria de cerca de US$ 3,88, ou 15,5 pontos por dólar. Depois de 01/12/2026, volta a 7,7.
- **Gemini 4 Argon:** a AA mede US$ 4,97 por tarefa, ao preço promocional. A partir de 01/12/2026, a estimativa é de cerca de US$ 9,94 por tarefa, e a razão cai de 10,1 para 5,1 pontos por dólar.
- **Efeito na fronteira (seção 04):** ao preço promocional, Gemini 4 Carbon entraria na fronteira; a partir de 01/12/2026, deixa de estar.

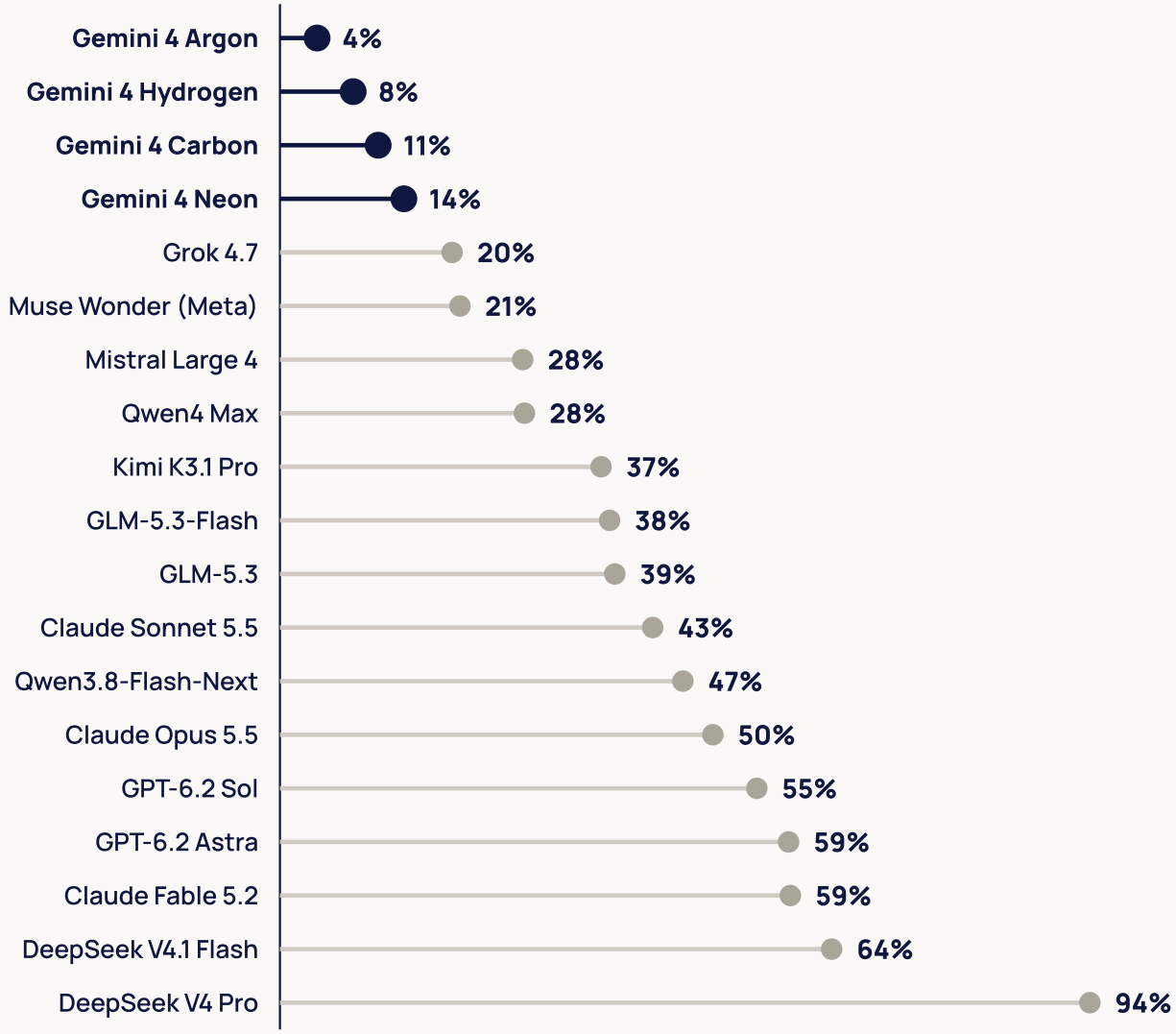
COMO A ESTIMATIVA FOI FEITA

Entrada e saída mudam na mesma proporção, então o custo por tarefa muda na proporção do preço misto 3:1 usado pela AA. Pressupostos: mesmo consumo de tokens por tarefa e nenhum desconto de cache ou lote. É estimativa da Vertere AI, não dado da AA.

09 Confiabilidade e contexto

Acertar mais não significa errar menos. Fable e Astra têm as maiores taxas de acerto em conhecimento (79% e 69%), mas chutam errado em 59% das vezes quando não sabem. Os Gemini 4 acertam menos (48% a 60%) e inventam muito menos (4% a 14%). Onde o erro custa caro, peça fontes ao modelo e preveja checagem.

A família Gemini 4 inventa menos; os modelos baratos e o Fable inventam mais



Fonte: Artificial Analysis v5.0.1, 19/11/2026; 19 modelos, no maior esforço de raciocínio listado pela AA. Quanto menor, melhor. Mede conhecimento factual em prova, não o risco em todas as tarefas.

Janela de contexto (texto máximo considerado de uma vez): 1 milhão de tokens na maioria, cerca de 1.500 páginas A4; Grok 4.7 tem 500 mil, Qwen4 Max 984 mil e Qwen3.8-Flash-Next 256 mil.

10 Recomendação, limites e fontes

RECOMENDAÇÃO

Comece pelo uso, não pelo ranking. Para a maioria das tarefas de nível alto, comece por GPT-6.2 Sol ou GPT-6.2 Astra. Reserve o Claude Fable 5.2 para casos críticos, em que a diferença de qualidade paga o custo. Para volume alto de tarefas simples e bem definidas, os modelos baratos servem, com revisão por regra ou amostragem. Onde o erro de fato custa caro, considere a família Gemini 4, sempre com revisão humana.

Antes de contratar, teste dois ou três finalistas com tarefas reais da empresa e meça o custo por tarefa, não só o preço por token. Reavalie a escolha quando houver novo modelo ou mudança de preço, como a do Google em 01/12/2026.

LIMITES

- **Benchmark não é o seu trabalho.** As 9 provas são genéricas; nenhuma é específica de português do Brasil, de leis brasileiras ou do seu sistema.
- **Esforço máximo.** A AA lista a versão de maior esforço: é o teto de qualidade, de custo e de espera. Em esforço menor, custo e tempo caem, e a nota também (o Astra vai de 62 no Max para 57 no Medium e 46 no Low, segundo a AA).
- **Custo por tarefa é o do índice.** Sua tarefa pode ter entradas maiores, menos raciocínio ou mais cache; a ordem entre modelos tende a se manter, o valor absoluto muda.
- **Velocidade e espera variam** com fornecedor, horário e carga; são medições da AA em 19/11/2026.
- **Diferenças pequenas não decidem.** De 1 a 2 pontos de índice, ou custos de US$ 0,57, 0,58 e 0,60, são empates práticos (leitura).
- **Preço muda.** O Google tem preço promocional até 30/11/2026, o DeepSeek varia por horário, e a AA pode listar preço diferente do oficial (caso do Carbon).
- **Alucinação** mede conhecimento factual em perguntas de prova, não o risco em todas as tarefas.
- **O índice v5.0.1 mudou.** Não há mais Coding Index nem Agentic Index separados; usamos os testes de código e agentes que compõem o índice. Não compare com versões antigas.
- **Cobertura.** Os Llama 4 não têm Índice de Inteligência na AA, por isso a Meta aparece com o Muse Wonder. O Muse Wonder Full (índice 50,1 na AA) não tem preço de API nem dados de velocidade e custo, e ficou fora.

Lacunas e divergências nos dados

- Preços de OpenAI, Anthropic e xAI vêm da verificação da manhã de 19/11/2026, e os valores que a AA lista batem com eles.
- Gemini 4 Carbon: a AA lista US$ 4 / US$ 20; a página oficial mostra US$ 2 / US$ 10 até 30/11. Usamos o oficial.
- Qwen3.8-Flash-Next: a correspondência com "qwen3.8-flash" da página oficial não está confirmada (mesmo preço, US$ 0,15 / US$ 0,47). A página da AA mostra contexto de 256 mil; a

da Alibaba cita faixa de 0 a 1 milhão para o "qwen3.8-flash".

- Qwen4 Max: usamos a região Internacional (Singapura), US$ 1,50 / US$ 4,50, faixa única até 1 milhão. A região China tem US$ 1,413 / US$ 4,24.
- Mistral Large 4: a página geral cita "Mistral Large" a US$ 0,5 / US$ 1,5 só como exemplo; o valor na página do modelo é US$ 0,70 / US$ 2,10, igual ao da AA.
- Z.ai: o contexto de 1 milhão vem da AA; a página de preços não o informa. Nenhuma página oficial consultada informou faixa extra de preço por contexto longo.

Fontes

Desempenho: Artificial Analysis, 19/11/2026: artificialanalysis.ai/leaderboards/models e artificialanalysis.ai/models/<modelo>; definições de alucinação e Omniscience em artificialanalysis.ai/evaluations/omniscience. **Preços oficiais** (horário de Fortaleza):

EMPRESA	PÁGINA OFICIAL DE PREÇO	ACESSO EM 19/11/2026
OpenAI	platform.openai.com/docs/pricing	19/11 (manhã)
Anthropic	docs.claude.com/en/docs/about-claude/pricing	19/11 (manhã)
SpaceX / xAI	docs.x.ai/docs/models	19/11 (manhã)
Google	ai.google.dev/gemini-api/docs/pricing	19/11 16:34
Moonshot (Kimi)	platform.kimi.ai/docs/pricing/chat	19/11 16:41
Alibaba (Qwen)	www.alibabacloud.com/help/en/model-studio/model-pricing	19/11 16:41
DeepSeek	api-docs.deepseek.com/quick_start/pricing	19/11 16:41
Mistral	docs.mistral.ai/models/mistral-large-4	19/11 16:41
Z.ai (GLM)	docs.z.ai/guides/overview/pricing	19/11 16:41
Meta	dev.meta.ai/docs/pricing-rate-limits	19/11 16:42

* Gemini 4 Carbon e Argon: preço promocional até 30/11/2026; a partir de 01/12/2026 será US$ 4 / US$ 20 (página oficial do Google). A AA já lista o Carbon a US$ 4 / US$ 20. † DeepSeek: preço de horário de pico (cache miss); fora do pico é metade. ‡ Qwen3.8-Flash-Next: a página da Alibaba lista "qwen3.8-flash" a US$ 0,15 / US$ 0,47; a correspondência com o nome da AA não está confirmada.

A Dados completos dos 19 modelos

MODELO	EMPRESA	ÍNDICE	TERM. BENCH	FRONTIER CODE	AUTOM. BENCH	TOK/S	TTFT (S)	TEMPO (MIN)	CUSTO / TAREFA US$	ENTR. / SAÍDA US$	CONTEXTO	ALUCIN.	PESOS ABERTOS
Claude Fable 5.2	Anthropic	64,6	59,6%	48,8%	76,4%	110	222	41,6	26,54	8 / 40	1 milhão	59%	não
GPT-6.2 Astra	OpenAI	62,1	64,1%	45,4%	77,2%	50	207	13,8	4,00	8 / 40	1 milhão	59%	não
Gemini 4 Neon	Google	61,4	63,6%	44,4%	79,3%	110	56	21,5	6,44	3 / 15	1 milhão	14%	não
Gemini 4 Carbon*	Google	60,0	65,2%	40,6%	77,8%	56	73	36,6	7,76	2 / 10	1 milhão	11%	não
Claude Opus 5.5	Anthropic	59,1	59,6%	46,4%	69,5%	115	339	25,1	10,04	4 / 20	1 milhão	50%	não
GPT-6.2 Sol	OpenAI	57,3	60,1%	44,4%	71,3%	76	79	8,9	1,01	2 / 10	1 milhão	55%	não
Claude Sonnet 5.5	Anthropic	54,5	63,6%	41,7%	71,8%	145	262	32,7	10,02	2 / 10	1 milhão	43%	não
Gemini 4 Hydrogen	Google	53,9	53,5%	39,0%	76,1%	170	45	15,2	3,16	1,50 / 7,50	1 milhão	8%	não
Gemini 4 Argon*	Google	50,4	57,1%	25,7%	77,5%	66	48	29,5	4,97	2 / 10	1 milhão	4%	não
Kimi K3.1 Pro	Moonshot (Kimi)	46,5	67,2%	25,5%	73,3%	31	157	38,8	3,68	2 / 10	1 milhão	37%	sim
Muse Wonder (Meta)	Meta	43,6	52,0%	27,9%	68,1%	117	20	13,7	6,95	4 / 20	1 milhão	21%	não
Qwen4 Max	Alibaba (Qwen)	42,9	48,5%	26,5%	72,6%	57	146	30,8	3,27	1,50 / 4,50	984 mil	28%	não
Grok 4.7	SpaceX / xAI	35,5	25,8%	28,7%	65,6%	70	52	24,9	7,77	2 / 6	500 mil	20%	não
DeepSeek V4.1 Flash†	DeepSeek	32,6	33,8%	28,5%	66,6%	165	44	14,0	0,60	0,30 / 1,20	1 milhão	64%	sim
GLM-5.3	Z.ai (GLM)	31,7	41,9%	21,6%	62,2%	81	72	24,4	5,01	1,40 / 4,40	1 milhão	39%	sim
Qwen3.8-Flash-Next‡	Alibaba (Qwen)	28,6	25,3%	27,1%	55,9%	78	80	31,8	0,57	0,15 / 0,47	256 mil	47%	sim
GLM-5.3-Flash	Z.ai (GLM)	27,8	32,8%	17,6%	60,4%	56	90	39,7	0,58	0,15 / 0,50	1 milhão	38%	sim
Mistral Large 4	Mistral	26,1	24,7%	15,0%	57,8%	61	80	22,5	2,06	0,70 / 2,10	1 milhão	28%	sim
DeepSeek V4 Pro†	DeepSeek	23,1	14,1%	17,1%	56,7%	86	59	19,3	1,77	1,32 / 3,96	1 milhão	94%	sim

Fonte: Artificial Analysis v5.0.1, 19/11/2026; 19 modelos, no maior esforço de raciocínio listado pela AA. Preços oficiais em US$ por 1 milhão de tokens (entrada / saída), 19/11/2026. * Gemini 4 Carbon e Argon: preço promocional do Google até 30/11/2026; depois, US$ 4 / US$ 20. † DeepSeek: preço de horário de pico. ‡ Qwen3.8-Flash-Next: correspondência com a página oficial não confirmada. Notas completas na seção 10.

Vertere AI · Comparativo de modelos de IA · 19/11/2026.

PRÓXIMO PASSO

Qual modelo faz sentido para a sua operação?

Podemos analisar com você as tarefas da sua empresa, comparar os modelos que fazem sentido e desenhar o caminho de adoção, com custo e qualidade claros.

www.vertereai.com.br

Acesse vertereai.com.br e agende um bate-papo com um dos nossos especialistas.

Vertere AI

Vertere AI, parte do ecossistema do Grupo Gomes de Matos.