

GUIA DE MODELOS E CUSTOS

GitHub Copilot vs. APIs oficiais • Revisão manual de código

Versão 2.2 | Lista de modelos consultada em 26 de outubro de 2026

[🔗 Acesse o guia online](#)

Objetivo desta revisão

Esta versão inclui os 28 modelos listados na documentação oficial de modelos com suporte no GitHub Copilot, identifica o fornecedor e se o modelo é proprietário, compara preços de referência por milhão de tokens no Copilot e na API direta quando há uma tarifa pública verificável e corrige os exemplos matemáticos de custo por token.

Leitura rápida

API direta	Você paga ao fornecedor por tokens processados segundo a tarifa do endpoint, com possíveis diferenças de cache, contexto longo, processamento prioritário e região.
GitHub Copilot	O consumo de modelos é convertido em GitHub AI Credits. 1 crédito equivale a US\$ 0,01; a assinatura pode incluir uma franquia de créditos. A tabela representa tarifas de referência para uso adicional, não necessariamente uma cobrança extra em toda interação.
Open source	Nenhum dos 28 modelos desta lista está identificado aqui como modelo de pesos abertos. Todos são tratados como modelos proprietários/serviços hospedados. A licença exata deve ser verificada na fonte do fornecedor.
Preços indisponíveis	Quando a API pública não mostra uma tarifa para o nome exato do modelo, o documento marca N/D. Não foi usada uma estimativa inventada para preencher a lacuna.

Como ler a tabela: todos os preços são USD por 1 milhão de tokens. “Entrada” é o texto/contexto enviado ao modelo; “entrada em cache” é conteúdo reutilizado quando elegível; “gravação em cache” é o custo de criar o cache em modelos que cobram por isso; “saída” são os tokens gerados. Os custos de cache não devem ser somados à entrada normal para os mesmos tokens: são categorias alternativas conforme o tipo de token faturado.

1. Todos os modelos suportados e seus custos

A coluna API direta mostra entrada/saída padrão do fornecedor para o modelo correspondente quando foi possível verificar uma tarifa pública. A tabela Copilot é a fonte autoritativa para valores usados no Copilot. Tarifas especiais, cache e contexto longo podem mudar o custo final.

Modelo	Empresa tipo	API direta Entrada / saída	Copilot Entrada	Entrada em cache	Gravação em cache	Copilot Saída	Contexto, categoria e observações
GPT-5.3-Codex	OpenAI Proprietário	\$1.75 / \$14.00	\$1.75	\$0.175	—	\$14.00	Categoria Copilot: Powerful. Sem tier longo listado na tabela de preços.
GPT-5.4 nano	OpenAI Proprietário	\$0.20 / \$1.25	\$0.20	\$0.02	—	\$1.25	Categoria: Lightweight.
GPT-5.6 Luna	OpenAI Proprietário	\$0.20 / \$1.20	\$0.20	\$0.02	\$0.25	\$1.20	Padrão ≤200K: mostrado. Longo >200K: entrada \$0.40, cache \$0.04, gravação \$0.50, saída \$1.80.
GPT-5.6 Sol	OpenAI Proprietário	\$4.00 / \$20.00	\$4.00	\$0.40	\$5.00	\$20.00	Padrão ≤272K. Longo >272K: entrada \$8, cache \$0.80, gravação \$10, saída \$30. API: tarifa padrão consultada; preço promocional pode variar.
GPT-5.6 Terra	OpenAI Proprietário	\$2.00 / \$12.00	\$2.00	\$0.20	\$2.50	\$12.00	Padrão ≤272K. Longo >272K: entrada \$4, cache \$0.40, gravação \$5, saída \$18.
GPT-6 Astra	OpenAI Proprietário	\$10.00 / \$50.00	\$10.00	\$1.00	\$12.50	\$50.00	Padrão ≤272K. Longo >272K: entrada \$20, cache \$2, gravação \$25, saída \$75.
GPT-6 Luna	OpenAI Proprietário	\$0.10 / \$0.50	\$0.10	\$0.01	\$0.125	\$0.50	Padrão ≤272K. Longo >272K: entrada \$0.20, cache \$0.02, gravação \$0.25, saída \$0.75.
GPT-6 Sol	OpenAI Proprietário	\$2.00 / \$10.00	\$2.00	\$0.20	\$2.50	\$10.00	Padrão ≤272K. Longo >272K: entrada \$4, cache \$0.40, gravação \$5, saída \$15.
GPT-6.1 Sol	OpenAI Proprietário	\$2.00 / \$10.00	\$2.00	\$0.10	\$2.50	\$10.00	Padrão ≤272K. Longo >272K: entrada \$4, cache \$0.20, gravação \$5, saída \$15.
GPT-6.1 Terra	OpenAI Proprietário	\$1.00 / \$5.00	\$1.00	\$0.05	\$1.25	\$5.00	Padrão ≤272K. Longo >272K: entrada \$2, cache \$0.10, gravação \$2.50, saída \$7.50.
Claude Haiku 4.5	Anthropic Proprietário	\$1.00 / \$5.00	\$1.00	\$0.10	\$1.25	\$5.00	Categoria Copilot: Versatile.
Claude Haiku 5.5	Anthropic Proprietário	\$0.10 / \$0.50 (≤100K)	\$0.10	\$0.01	\$0.125	\$0.50	Padrão ≤100K. Longo >100K: entrada \$0.50, cache \$0.05, gravação \$0.625, saída \$2.50. API direta também tem tarifa longa.
Claude Sonnet 4.6	Anthropic Proprietário	\$3.00 / \$15.00	\$3.00	\$0.30	\$3.75	\$15.00	Categoria Copilot: Versatile.
Claude Opus 4.8	Anthropic Proprietário	\$5.00 / \$25.00	\$5.00	\$0.50	\$6.25	\$25.00	Categoria Copilot: Powerful.
Claude Opus 4.8 (fast mode)	Anthropic Proprietário	N/D*; base \$5 / \$25	\$10.00	\$1.00	\$12.50	\$50.00	Preview no Copilot. *Tarifa direta da variante fast não confirmada com SKU equivalente.
Claude Opus 5	Anthropic Proprietário	\$5.00 / \$25.00	\$5.00	\$0.50	\$6.25	\$25.00	Categoria Copilot: Powerful.
Claude Opus 5.5	Anthropic Proprietário	\$4.00 / \$20.00	\$4.00	\$0.20	\$5.00	\$20.00	Categoria Copilot: Powerful. Cache de leitura: \$0.20/M.

Modelo	Empresa tipo	API direta Entrada / saída	Copilot Entrada	Entrada em cache	Gravação em cache	Copilot Saída	Contexto, categoria e observações
Claude Sonnet 5	Anthropic Proprietário	\$2.00 / \$10.00	\$2.00	\$0.20	\$2.50	\$10.00	Categoria Copilot: Versatile.
Claude Sonnet 5.5	Anthropic Proprietário	\$2.00 / \$10.00	\$2.00	\$0.10	\$2.50	\$10.00	Categoria Copilot: Versatile. Preço de cache de leitura foi reduzido para \$0.10/M.
Claude Fable 5	Anthropic Proprietário	N/D*	\$10.00	\$1.00	\$12.50	\$50.00	*Não localizei tarifa pública de API para o nome exato; preço Copilot oficial informado.
Claude Fable 5.1	Anthropic Proprietário	N/D*	\$10.00	\$0.25	\$12.50	\$50.00	*Não localizei tarifa pública de API para o nome exato; preço Copilot oficial informado.
Gemini 3.8 Flash	Google Proprietário	\$0.75 / \$3.75**	\$0.75	\$0.075	—	\$3.75	**API: preço promocional até 31/12/2026; passa a \$1.50/\$7.50 em 01/01/2027.
Gemini 4 Argon	Google Proprietário	\$2.00 / \$10.00**	\$2.00	\$0.10	—	\$10.00	**API: preço promocional até 30/11/2026; passa a \$4.00/\$20.00 em 01/12/2026.
MAI-Code-1.1-Flash	Microsoft Proprietário	N/D*	\$0.20	\$0.02	—	\$1.20	*Modelo hospedado no Azure; catálogo Foundry oferece cotação/seleção de preço por implantação.
Kimi K3	Moonshot AI Proprietário	\$3.00 / \$15.00	\$3.00	\$0.30	—	\$15.00	Categoria Copilot: Powerful. API direta informa \$0.30/M para entrada em cache.
Grok 4.5	xAI Proprietário	\$2.00 / \$6.00	\$2.00	\$0.50	—	\$6.00	Padrão ≤200K. Copilot longo >200K: entrada \$4, cache \$1, saída \$12. API direta tem entrada em cache de \$0.30/M.
Grok 4.6	xAI Proprietário	\$2.00 / \$6.00	\$2.00	\$0.50	—	\$6.00	Padrão ≤200K. Longo >200K: entrada \$4, cache \$1, saída \$12.
Grok 4.7	xAI Proprietário	\$2.00 / \$6.00	\$2.00	\$0.50	—	\$6.00	Padrão ≤200K. Longo >200K: entrada \$4, cache \$1, saída \$12.

Legenda: “—” significa que a tabela oficial do Copilot não informa uma tarifa separada de gravação em cache para aquele modelo. “N/D” indica que não foi possível confirmar uma tarifa pública para a API direta com o nome exato; não significa que nunca possa haver acesso por outro provedor ou contrato.

2. Como entender o custo por token

O preço por milhão é o formato usado pelas tabelas oficiais. Para obter o custo unitário por token, divida o valor por 1.000.000. O exemplo abaixo considera entrada não armazenada em cache, saída normal, contexto padrão e nenhuma chamada de ferramenta adicional.

Modelo (tarifa Copilot padrão)	Entrada por 1M	Saída por 1M	Entrada por token	Saída por token	100 mil entrada + 10 mil saída
GPT-6 Luna	\$0.10	\$0.50	\$0.00000010	\$0.00000050	\$0.015 = \$0.010 + \$0.005
Claude Haiku 5.5 (prompt ≤100K)	\$0.10	\$0.50	\$0.00000010	\$0.00000050	\$0.015 = \$0.010 + \$0.005
Claude Sonnet 5.5	\$2.00	\$10.00	\$0.00000200	\$0.00001000	\$0.30 = \$0.200 + \$0.100
GPT-6 Sol	\$2.00	\$10.00	\$0.00000200	\$0.00001000	\$0.30 = \$0.200 + \$0.100
GPT-6 Astra	\$10.00	\$50.00	\$0.00001000	\$0.00005000	\$1.50 = \$1.000 + \$0.500

Fórmula correta

Custo = (tokens de entrada ÷ 1.000.000 × tarifa de entrada) + (tokens de saída ÷ 1.000.000 × tarifa de saída).

Exemplo — GPT-6 Luna

Entrada: 100.000 ÷ 1.000.000 × US\$ 0,10 = US\$ 0,010. Saída: 10.000 ÷ 1.000.000 × US\$ 0,50 = US\$ 0,005. Total: US\$ 0,015. O total é 1,5 centavo de dólar, não US\$ 0,15 nem US\$ 1,50.

Atenção ao contexto longo

O preço pode aumentar quando o prompt ultrapassa o limite específico do modelo. Exemplo: Claude Haiku 5.5 muda de US\$ 0,10/US\$ 0,50 para US\$ 0,50/US\$ 2,50 por milhão de tokens acima de 100K tokens de entrada. Para GPT-6 Luna, a tabela Copilot dobra as tarifas de entrada/saída no tier longo acima de 272K tokens. Use a linha do tier correto para estimar.

Entrada em cache

Quando a plataforma informa tokens faturados como entrada em cache, aplique a tarifa de cache àquela parcela. Não calcule o mesmo token simultaneamente pela tarifa normal e pela tarifa cache. Modelos com gravação de cache podem ainda cobrar um valor separado ao criar o cache.

3. API direta vs. GitHub Copilot

Contratação direta por API	Uso no GitHub Copilot
O fornecedor cobra pelo consumo de tokens, usando as regras do endpoint. Pode haver descontos ou sobretaxas para cache, processamento em lote, contexto longo, prioridade e região.	O GitHub converte o uso em AI Credits. A tarifa por milhão de tokens é expressa em créditos equivalentes a dólares: 1 AI Credit = US\$ 0,01. Os planos incluem franquias de créditos conforme as condições do plano.
As tarifas na tabela são do serviço de API correspondente quando publicadas. Um nome de modelo disponível no Copilot pode não estar disponível com o mesmo SKU em uma API pública direta.	Quando os créditos incluídos cobrem o uso, o custo marginal pago pode ser zero até o limite da franquia. Ao exceder a franquia, aplicam-se as regras de cobrança adicional vigentes.

Nota para revisão de código: a função “Copilot code review” pode selecionar automaticamente o modelo, e a cobrança também pode incluir minutos de GitHub Actions. Portanto, esta comparação de tokens é uma estimativa para entender as tarifas de modelos, não uma previsão garantida do custo de uma revisão automatizada.

4. Recomendações para code review

1. Triagem econômica	Use um modelo leve para enumerar riscos, funções suspeitas, casos extremos e testes ausentes. Limite o contexto aos arquivos relevantes. Modelos sugeridos: GPT-6 Luna ou Claude Haiku 5.5, opções de menor custo na tabela.
2. Investigação complexa	Use um modelo versátil ou potente para revisar fluxos entre arquivos, concorrência, autenticação, estado compartilhado e regressões. Forneça o diff e o comportamento esperado. Modelos sugeridos: Claude Opus 5.5 ou GPT-6 Astra para análise profunda; GPT-6 Sol como alternativa de menor custo.
3. Verificação	Solicite referências a arquivo/linha, hipótese reproduzível e teste que falharia antes da correção. Valide as recomendações executando testes e conferindo o código. Modelos sugeridos: Claude Sonnet 5.5 ou GPT-6 Sol para uma segunda análise equilibrada; execute os testes independentemente do modelo.
4. Controle de contexto e custo	Evite enviar o repositório inteiro se apenas alguns arquivos forem relevantes. Ao mudar de modelo, não presuma que todo cache sempre será invalidado: isso depende do produto e da implementação. Modelos sugeridos: GPT-6 Luna para contextos menores e GPT-6.1 Terra quando uma análise realmente exigir mais contexto; confira os tiers e custos de cache na tabela.

5. Fontes oficiais e metodologia

Lista de modelos e tarifas consultadas em 26/10/2026. Os fornecedores podem alterar nomes, disponibilidade, limites de contexto, descontos e preços. A coluna de API direta reflete tarifas públicas que puderam ser verificadas para o nome exato ou correspondente do modelo; “N/D” foi usado quando isso não foi confirmado.

GitHub Docs — Modelos e preços para GitHub Copilot

<https://docs.github.com/pt/copilot/reference/copilot-billing/models-and-pricing>

Fonte das tarifas por 1 milhão de tokens no Copilot, incluindo entrada, cache, gravação de cache, saída e tiers de contexto.

GitHub Docs — Modelos de IA com suporte no Copilot

<https://docs.github.com/en/copilot/reference/ai-models/supported-models>

Fonte da lista de modelos com suporte e seus provedores. A disponibilidade pode depender do plano, cliente e políticas da organização.

OpenAI API — Pricing

<https://developers.openai.com/api/docs/pricing>

Fonte de referência para os modelos GPT e suas tarifas de API.

Anthropic — Claude Haiku 5.5

<https://www.anthropic.com/claude-haiku-5-5>

Fonte para tarifas de Haiku 5.5 e comparação de tarifas atuais de Haiku/Sonnet.

Anthropic — Claude Sonnet 5.5

<https://www.anthropic.com/claude-sonnet-5-5>

Fonte para tarifas de Sonnet 5.5 e preços de cache.

Anthropic — Claude Opus 5.5

<https://www.anthropic.com/claude-opus-5-5>

Fonte para tarifas de Opus 5.5 e comparação com Opus 5.

Google — Gemini API Pricing

<https://ai.google.dev/gemini-api/docs/pricing>

Fonte para preço da API Gemini, incluindo promoções e datas de mudança de tarifa.

Microsoft Foundry — MAI-Code-1.1-Flash

<https://ai.azure.com/catalog/models/mai-code-1.1-flash>

Catálogo do modelo; o preço de implantação deve ser confirmado no portal Azure.

xAI — Pricing

<https://docs.x.ai/developers/pricing>

Fonte de preços de Grok 4.5, 4.6 e 4.7, incluindo entrada em cache e contexto longo.

Moonshot AI — anúncio do Kimi K3

<https://forum.moonshot.ai/t/kimi-k3-is-here-our-most-capable-model/480>

Fonte da tarifa de API publicada para Kimi K3.

Limitações importantes

- Preços em USD por 1 milhão de tokens; custos locais dependem de câmbio, impostos, cache efetivamente utilizado, contexto, volume e modalidade de API.
- “Proprietário” identifica o modelo como serviço/modelo fechado na comparação; não é uma avaliação jurídica completa de termos e licenças.
- O total estimado por interação não inclui ferramentas, pesquisa web, chamadas de função, minutos de GitHub Actions, execução agentic ou outros custos fora dos tokens listados.
- Para orçamento real, registre consumo por modelo e calcule o custo de várias revisões típicas.

Conclusão: escolha o modelo pelo tipo de revisão e não só pelo preço por token. Uma análise curta com modelo potente pode sair mais barata que uma sequência longa de tentativas com um modelo inadequado; compare o custo por tarefa com evidências e resultados reais.