ChatGPT

Prijsstellingsstrategieën van AI-bedrijven in 2023–2026 vergeleken met Kotler

Executive summary

De meeste grote AI-aanbieders zijn in 2023–2026 geconvergeerd op **usage-based pricing** (meestal per **input- en outputtokens**) voor API-aanbod, aangevuld met **abonnementen** (B2C/B2B) en **commitment-/capaciteitsmodellen** voor voorspelbaarheid en korting (bijv. gereserveerde throughput). [1] Dit sluit sterk aan bij de economische logica van cloud: in plaats van een eenmalige licentie is “pay-as-you-go” de standaard, met een complexe set opties en meters. [2]

Vergeleken met **klassieke prijsstrategieën van Philip Kotler** (o.a. skimming/penetration, product-mix-pricing zoals bundling en captive/two-part pricing, plus prijsaanpassingen zoals kortingen en segmented pricing) zijn er duidelijke overeenkomsten: AI-bedrijven doen intensief aan **productlijn-prijsstelling** (model-tiers), **bundling** (abonnementen met meerdere features) en **(twee-delige) tariefstructuren** (vaste fee + variabel gebruik). [3]

Tegelijkertijd ontstaan hybride en “cloud-native” varianten die Kotlers indeling slechts gedeeltelijk dekt: **token-metering** als micro-unit, **context-window-drempels** (prijs verandert bij >200k tokens), **caching-prijzen** (goedkopere “cached input” of “cache read”), en **service tiers** (Flex/Batch/Reserve) die vraag en latency sturen—een combinatie van prijsstrategie en capaciteitsmanagement. [4]

Afbakening en methode

Dit rapport vergelijkt actuele prijspraktijken (primair 2023–2026) van vijf aanbieders: OpenAI [5], Google [6] (via Vertex AI/Google Cloud), Anthropic [7], Microsoft [8] (Azure OpenAI/Azure AI) en Cohere [9]. Prijsdata zijn waar mogelijk gebaseerd op **officiële pricing- of documentatiepagina’s**; waar prijzen contractueel/regio-afhankelijk zijn, wordt dat expliciet benoemd. [10]

Belangrijke beperkingen: - **Valuta en regio**: prijzen kunnen lokaal worden weergegeven (bijv. CHF/EUR) of afhangen van deployment type/regionale data-zones. [11]
- **List price vs. enterprise**: enterprise-/volumeafspraken zijn vaak “contact sales”, waardoor kortingen niet publiek vergelijkbaar zijn. [12]
- **Unit-definities**: “token”, “character”, “query”, “training hour” en “cached token” zijn niet volledig uitwisselbaar; dat is juist een kernpunt van de strategie. [13]

Huidige prijsstellingspraktijken per aanbieder

Onderstaande mermaid-flow vat de dominante prijslogica samen: abonnementen en trials bestaan, maar productie-gebruik wordt vrijwel altijd gemeten en gefactureerd op verbruik (tokens/requests/compute), met differentiatie via tiers, caching en commitments.

```
flowchart LR
A[Toegang] -->|Free / Trial| B[Beperkingen: quota/rate limits]
```

```
A -->|Abonnement| C[Vaste fee per gebruiker/maand]
A -->|API pay-as-you-go| D[Metering]
C --> D
D --> E{Metriek}
E -->|Tokens in/out| F[LLM inference]
E -->|API calls / queries| G[Search/rerank/agents]
E -->|Compute-uren| H[Fine-tuning / provisioned capacity]
F --> I[Prijslevers: model-tier, contextdrempels, caching]
G --> I
H --> I
I --> J[Kortingen: batch/flex/commitments + enterprise deals]
J --> K[Factuur/chargeback]
```

[14]

OpenAI (API + fine-tuning + ChatGPT abonnementen)
Voor API's is het model typisch: prijs per **1M tokens**, apart voor **input**, **cached input** en **output**, met meerdere modellen en tiers. Bijvoorbeeld: *gpt-4o* $2.50/1M input, $1.25/1M cached input, $10/1M output; *o3* $2.00/1M input, $0.50/1M cached input, $8.00/1M output. [15] Dezelfde pagina toont ook meerdere **service tiers** (o.a. Flex) met andere tarieven, wat prijs inzet als keuzevariabele voor latency/throughput. [16]
Fine-tuning introduceert extra meters: **training per 1M tokens** of zelfs **training per uur** (bij reinforcement fine-tuning). Voorbeelden: GPT-4.1 fine-tuning training $25/1M tokens; o4-mini reinforcement fine-tuning training $100 per training hour. [17]
Aan de abonnementskant is er een expliciete freemium- en tier-ladder: ChatGPT Go ($8/maand, US-prijs), Plus ($20/maand) en Pro ($200/maand). [18] Voor zakelijke teams is "Business" per gebruiker per maand (weergave is gelokaliseerd; in de bron wordt CHF 29/user/month bij jaarfacturatie getoond) en "Enterprise" is contact-sales. [19]

Google (Vertex AI / Google Cloud Generative AI)
Vertex AI hanteert pay-as-you-go met **prijs per 1M tokens**, maar met een opvallende variatie: prijs verschilt bij **contextwindow-drempels** (≤200k tokens versus >200k tokens) en er zijn expliciete tarieven voor **cached input**. Bijvoorbeeld voor Gemini 3.0 Pro (tekst): input $1.25/1M tokens (≤200k) en $2.50/1M tokens (>200k); output $5.00/1M tokens (≤200k) en $10.00/1M tokens (>200k); cached input is lager geprijsd.
Daarnaast bestaat "Model Optimizer" met een per-response prijs en drempel: er wordt pas gefactureerd na 200 responses; errors worden niet doorbelast.
Qua gratis instap gebruikt Google Cloud klassieke cloud-mechanismen: **$300 trial credit** voor nieuwe klanten (met verdere free-tier opties). [20]

Anthropic (Claude API + Claude abonnementen)
Anthropic's API-prijsbasis is eveneens tokens in/out per 1M, met duidelijke "value ladder" tussen modellen: Claude Haiku 3.5 start bij $0.80/1M input tokens en $4/1M output tokens; Claude Sonnet 4 bij $3/1M input en $15/1M output; Claude Opus 4.5 bij $15/1M input en $75/1M output. [21] Daarbovenop documenteert Anthropic expliciet aanvullende meters zoals **prompt caching** en gerelateerde prijsregels (cache-economics).
Aan de B2C-kant is Claude Pro een klassiek abonnement: $20/maand (US) of $17/maand bij jaarabonnement ($200 upfront); prijzen zijn regio-afhankelijk. [22]

Microsoft Azure AI / Azure OpenAI Service
Azure OpenAI combineert **pay-as-you-go tokens** met capaciteits-/commitment-modellen: Standard (on-demand) rekent input/output tokens; Provisioned Throughput Units (PTUs) reserveren capaciteit met maand-/jaarreserveringen om kosten te verlagen; Batch API biedt 50% korting op Global Standard pricing (met langere doorlooptijd). [23] Azure maakt prijs bovendien expliciet stuurbaar via **deployment types** (Global / Data Zone EU-US / Regional). [23]
Concreet voorbeeld uit de Azure pricing-informatie (let op: weergave kan verschillen per regio/selectie): GPT-4o-Transcribe noemt tekst-input $2.50/1M tokens en tekst-output $10/1M tokens, naast audio-meters. [24] Voor instap gebruikt Azure ook credits ("Try Azure for free" met $200 credit voor 30 dagen). [25]

Cohere (API + trial keys + enterprise private deployment)
Cohere maakt expliciet onderscheid tussen **trial usage** en **production usage**: trial API keys zijn gratis maar beperkt; production keys zijn betaald en minder beperkt. [26] Voor generatieve modellen is pricing token-gebaseerd; Command R+ (08-2024) vermeldt $2.50/1M input tokens en $10/1M output tokens. [27] Cohere documenteert bovendien dat "billed units" (billed input/output tokens) kunnen afwijken van ruwe token-tellingen door model-/systeemgedrag, wat pricing en cost-forecasting minder triviaal maakt dan "tokens × tarief". [28]
Voor enterprise-behoeften (model customization, private deployment) is pricing "custom" en afhankelijk van eisen—klassiek 'contact sales'. [29]
Ter illustratie van alternatieve meters buiten tokens: via Amazon Web Services [30] Bedrock wordt Cohere Rerank 3.5 bijvoorbeeld per **1.000 queries** geprijsd (met constraints per query/document chunk). [31]

Vergelijkende tabel en kernverschillen

Aanbieder	Primaire prijsmodellen	Hoofdmeter(s)	Voorbeeld list price (tekst-LLM)	Gratis instap	Volume/ enterprise-mechanism
OpenAI [5]	API pay-as-you-go + fine-tuning tarieven + abonnementen (B2C/B2B) [32]	1M tokens (input/ cached/output), training tokens, training hours [33]	gpt-4o: $2.50 in / $10 out per 1M tokens (cached input goedkoper) [15]	ChatGPT Free + betaalde tiers (Go/ Plus/Pro) [34]	Enterprise "contact sales"; business per se [19]
Google [6]	Cloud pay-as-you-go (Vertex AI) + trial credits [35]	1M tokens (in/out), cached tokens; soms per response (optimizer)	Gemini 3.0 Pro (≤200k): $1.25 in / $5 out per 1M tokens; >200k duurder	$300 trial credit (90 dagen) + free tier (select producten) [20]	Enterprise cloud-contracten; prijs vaak integraal onderde van cloud spend management [36]
Anthropic [7]	API pay-as-you-go + abonnementen (Claude Pro e.a.) [37]	1M input/output tokens; aanvullende meters (o.a. caching)	Sonnet 4: $3 in / $15 out per 1M tokens; Opus 4.5 hoger [38]	Claude Free; Pro $20/mo of $17/mo bij jaarplan [39]	Team/enterprise contractueel (niet altijd publiek) [40]

Aanbieder	Primaire prijsmodellen	Hoofdmeter(s)	Voorbeeld list price (tekst-LLM)	Gratis instap	Volume/ enterprise-mechanism
Microsoft [8]	Azure OpenAI: pay-as-you-go + PTU-reserveringen + batch-korting [23]	Tokens (input/ output), throughput-capaciteit (PTU), batch jobs [23]	GPT-4o-Transcribe voorbeeld: $2.50 tekst-input / $10 tekst-output per 1M tokens [24]	Azure free trial ($200 credit/30 dagen) [25]	PTU maand/ jaar-reserveringen; region/data zone keuz [23]
Cohere [9]	API pay-as-you-go + trial keys (gratis/ beperkt) + enterprise private deployment [42]	1M tokens (in/out); billed units kunnen afwijken [43]	Command R+ (08-2024): $2.50 in / $10 out per 1M tokens [27]	Trial/ evaluation keys gratis maar beperkt [26]	Custom pricing voor customization/private deployment [29]

Kernobservatie: hoewel "$ per 1M tokens" een ogenschijnlijk uniforme meeteenheid is, blijken de **échte strategische verschillen** te zitten in (i) welke tokens worden geteld (raw vs billed, cached vs non-cached), (ii) hoe prijs verandert met contextlengte en latency-eisen, en (iii) hoe sterk abonnementen/credits/capaciteitsreserveringen de marginale prijs dempen of juist escaleren. [45]

Kotler-mapping en waar het schuurt

Kotlers klassieke prijsstrategieën als referentiekader

Kotler's (en Kotler-gerelateerde Pearson-literatuur) onderscheidt onder meer: - **New-product pricing**: *market-skimming* vs *market-penetration*. [46]
- **Product-mix pricing**: product line, optional product, **captive product** en **product bundle**; bij diensten vaak **two-part pricing** (vaste fee + variabele usage rate). [47]
- **Price-adjustment strategies**: *discount and allowance*, **segmented pricing** (prijsdiscriminatie) en **psychological pricing**. [46]

Deze categorieën zijn historisch product-/marktgericht. Cloud- en SaaS-literatuur laat zien dat digitale diensten in de praktijk vaak juist worden geprijsd als **variabele kostenmodellen** met veel opties (pay-as-you-go, reserveringen, etc.). [48]

Expliciete overeenkomsten: "AI pricing" als Kotler-productmix + cloud-metering

Product line pricing (Kotler) ↔ model-tiering / capability tiering
Bij alle aanbieders bestaan meerdere modellen (mini/flash/haiku vs pro/opus/flagship) met sterke prijsstappen. Dit is textbook *product line pricing* (verschillende prijsstappen binnen één productlijn), maar hier gemeten in tokens in/out. [49]

Product bundle pricing (Kotler) ↔ abonnementen die meerdere features bundelen
Chat/assistants worden gebundeld met extra tools (analysis, voice, agents, admin controls) tegen een per-user prijs, waardoor de marginale tokenprijs impliciet wordt. Zowel OpenAI Business/Enterprise als Claude Pro zijn hier voorbeelden. [50]

Captive product pricing / two-part pricing (Kotler) ↔ vaste fee + variabele meters / credits
Het klassieke "razor-and-blades" idee (laag geprijsd basisproduct, hoger geprijsde verbruikscomponent) verschijnt in AI als: (a) abonnement per seat plus (b) extra credits/usage voor zwaardere modellen of enterprise-features; bij diensten is dit precies Kotlers "two-part pricing". [51]

Segmented pricing (Kotler) ↔ price discrimination via tiers, regio's, latency en commitments
Prijsdifferentiatie gebeurt niet alleen via "klantsegmenten" maar via: deployment type (regional vs global/data zone), reserved throughput (PTU), batch-kortingen en flex-tiers. Dit is functioneel prijsdiscriminatie op basis van service level en contractduur. [52]

Skimming vs penetration (Kotler) ↔ premium frontier models + goedkope instapmodellen/free tiers
De ladder "Haiku/Flash/mini" versus "Opus/Pro/Thinking" en daarnaast trials/free tiers, past op een skimming/penetration-hybride: hoge marges op topsegmenten, brede adoptie via goedkope of gratis instap. [53]

Waar Kotler minder volstaat: nieuwe of hybride modellen

De onderstaande mermaid-mapping toont waar klassieke labels te grof worden: veel moderne AI-pricing combineert meerdere Kotler-strategieën én voegt "cloud-mechanismen" toe die tegelijk prijsstrategie en capaciteitssturing zijn.

```
flowchart TB
K1[Kotler: product line pricing] --> A1[Model-tiers: haiku/mini vs opus/pro]
K2[Kotler: bundle pricing] --> A2[Abonnementen: tools + modellen +
governance]
K3[Kotler: captive/two-part pricing] --> A3[Seat fee + variabele tokens/
credits]
K4[Kotler: segmented pricing] --> A4[Tiers: batch/flex/priority + regio/data
zone + commitments]
K5[Kotler: skimming/penetration] --> A5[Premium frontier + free/trial instap]
A6["Nieuwe variant: token- & context-metering"] --> A4
A7["Nieuwe variant: caching-tarieven"] --> A3
A8["Nieuwe variant: optimizer (charged after N)"] --> A4
```

[54]

Wat Kotler conceptueel minder scherp maakt in deze markt: - **De meeteenheid is niet "product" maar "compute-proxy"**: tokens/cached tokens/context buckets zijn operationele proxies voor inference-kosten en capacity. Cloud-economics benadrukt dat cloud-diensten pay-as-you-go en multi-option zijn; AI-pricing erft dat patroon. [55]
- **Prijs als vraag- en latency-sturing**: Flex/Batch/Reserved/Provisioned tiers zijn geen simpele korting, maar sturen doorlooptijd en resource-allocatie (een mengvorm van pricing en operations). [56]
- **"Freemium" is vaak credit-gedreven i.p.v. permanent gratis**: cloud-trials ($300) en trial keys zijn economisch anders dan een blijvende freemium-laag; ze zijn acquisitie- en leerinstrumenten met harde grenzen. [57]
- **Marginale prijs is context-afhankelijk**: bij Vertex AI verandert de prijs bij >200k tokens; bij meerdere aanbieders bestaan cache-tarieven die de marginale prijs bewust verlagen voor hergebruik—een expliciete prikkel tot architectuurkeuzes (prompt reuse, caching). [58]

Casestudies

Casus OpenAI: “service tiers” als dynamische prijsdiscriminatie op latency
OpenAI’s pricing-pagina laat zien dat dezelfde (of vergelijkbare) modellen in meerdere tiers kunnen bestaan (bijv. “Flex” met lagere tarieven). Dit lijkt op Kotlers *segmented pricing*, maar het segment is hier: **service level** (wachttijd/priority) in plaats van demografie. [59] In dezelfde prijstabellen is “cached input” substantieel goedkoper, waardoor productarchitectuur (cache-geschikte prompts) een economisch ontwerpdoel wordt—iets wat Kotler wel in algemene zin als “price adjustment” kan vangen, maar niet op micro-meter niveau. [60]

Casus Google Vertex AI: context-drempels en caching als price-for-constraints
Vertex AI publiceert expliciet verschillende tarieven voor ≤200k versus >200k tokens, plus aparte cached-input prijzen. Dit is geen klassieke “bundel” of “discount” maar een prijsmechanisme dat de echte kostendrijver (extreme contextlengte) internaliseert. In Kotler-termen is dit te plaatsen onder product line/price adjustment, maar het is in feite een **non-lineaire tariefstructuur** (marginale prijs stijgt na een drempel) die typisch is voor cloud-pricing. [61]

Casus Cohere: trial-vs-production keys en billed_units
Cohere maakt pricing-governance expliciet: trial API key usage is gratis maar beperkt, production keys zijn betaald; daarnaast kan billing plaatsvinden op “billed input/output tokens” die afwijken van ruwe tokens doordat Cohere soms tokens “onder de motorkap” toevoegt of anders telt. [26] Dit is praktisch een moderne variant van freemium + two-part pricing: gratis evaluatie, daarna verbruiksfacturatie; maar met een additionele complexiteit (billing model) die forecasting en chargeback beïnvloedt. [62]

Casus Azure OpenAI: PTU-reserveringen en batch-korting als cloud-commitment strategie
Azure OpenAI positioneert drie paden: Standard pay-as-you-go tokens, Provisioned Throughput Units (PTUs) met maand/jaar-reserveringen om spend te verlagen, en Batch API met 50% korting (met langere doorlooptijd). [23] Dit is klassiek cloud-mechanisme (commitment/slotting), en past in Kotler deels onder *discount/allowance* en *segmented pricing*, maar het is tegelijk een operationeel contractmechanisme (capaciteit reserveren). [63]

Conclusies: hybride modellen, implicaties en “Kotler-plus”

De actuele prijsstrategieën van AI-bedrijven zijn het best te begrijpen als een **hybride van Kotler’s product-mix-denken en cloud/SaaS-metering**. Kotler blijft nuttig voor het “waarom” (segmenten, bundels, productlijnen), maar schiet tekort in het “hoe” van moderne digitale metering: tokens, caching, context-drempels en latency-tiers zijn fijnmaziger dan klassieke retail-/FMCG-analogieën. [64]

Een bruikbaar “Kotler-plus” raamwerk voor 2023–2026 is daarom: - **Product line pricing → model-/capability-versioning** (meerdere modellen met oplopende waarde en oplopende $/token). [65]
- **Two-part / multi-part tariffs → seat fee + inbegrepen quota + overage** (abonnementen en enterprise credits) én parallel pay-as-you-go API. [51]
- **Segmented pricing → service-level, regio en commitment-gebaseerde discriminatie** (Flex/Batch/Reserved/PTU, data zones). [66]
- **Freemium → credit/trial-gedreven acquisition** (trial keys, $300 cloud credits) in plaats van “gratis voor altijd”, omdat inferencekosten structureel variabel zijn. [67]

Praktische implicatie voor afnemers (FinOps/inkoop/architectuur): de grootste kostenverschillen komen zelden uit “welke aanbieder is het goedkoopst per token” alleen, maar uit **architectuur-compatibiliteit**

met prijslevers (caching, batch, context-discipline), plus contractkeuzes (commitments/PTU) en governance (chargeback op billed units). [68]

---

[1] [9] [10] [15] [16] [32] [33] [45] [49] [56] [58] [59] [60] https://developers.openai.com/api/docs/pricing/
https://developers.openai.com/api/docs/pricing/

[2] [14] [48] [55] [61] “The Economics of the Cloud”
https://www.tse-fr.eu/sites/default/files/TSE/documents/doc/wp/2024/wp_tse_1520.pdf?utm_source=chatgpt.com

[3] [46] [47] [54] [63] [64] https://www.pearson.com/content/dam/one-dot-com/one-dot-com/hei-samples/marketing/9781488626203-toc.pdf
https://www.pearson.com/content/dam/one-dot-com/one-dot-com/hei-samples/marketing/9781488626203-toc.pdf

[4] [13] https://cloud.google.com/vertex-ai/generative-ai/pricing
https://cloud.google.com/vertex-ai/generative-ai/pricing

[5] [27] https://docs.cohere.com/docs/command-r-plus
https://docs.cohere.com/docs/command-r-plus

[6] [37] [38] [41] https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro
https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro

[7] [8] [23] [24] [25] [30] [52] [66] [68] https://azure.microsoft.com/en-us/pricing/details/azure-openai/
https://azure.microsoft.com/en-us/pricing/details/azure-openai/

[11] [12] [19] [50] https://openai.com/business/chatgpt-pricing/
https://openai.com/business/chatgpt-pricing/

[17] https://openai.com/api/pricing/
https://openai.com/api/pricing/

[18] https://openai.com/index/introducing-chatgpt-go/
https://openai.com/index/introducing-chatgpt-go/

[20] [35] [57] [67] Free Trial and Free Tier Services and Products
https://cloud.google.com/free?utm_source=chatgpt.com

[21] https://developers.openai.com/api/docs/models/gpt-4.1
https://developers.openai.com/api/docs/models/gpt-4.1

[22] [39] Plans & Pricing | Claude by Anthropic
https://claude.com/pricing?utm_source=chatgpt.com

[26] [28] [42] [43] [44] [62] https://docs.cohere.com/docs/how-does-cohere-pricing-work
https://docs.cohere.com/docs/how-does-cohere-pricing-work

[29] https://cohere.com/pricing
https://cohere.com/pricing

[31] https://aws.amazon.com/bedrock/pricing/
https://aws.amazon.com/bedrock/pricing/

[34] https://chatgpt.com/pricing/
https://chatgpt.com/pricing/

[36] https://azure.microsoft.com/en-us/pricing
https://azure.microsoft.com/en-us/pricing

40 Anthropic intensifies AI competition with $200 plan for Claude models
https://www.reuters.com/technology/artificial-intelligence/anthropic-intensifies-ai-competition-with-200-plan-claude-models-2025-04-09/?utm_source=chatgpt.com

51 https://nscpolteksby.ac.id/ebook/files/Ebook/Business%20Administration/Principles%20of%20Marketing%2014th%20Edition%20-%20Kotler%20and%20Amstrong%20%282012%29/Chapter%2011%20-%20Pricing%20Strategies.pdf
https://nscpolteksby.ac.id/ebook/files/Ebook/Business%20Administration/Principles%20of%20Marketing%2014th%20Edition%20-%20Kotler%20and%20Amstrong%20%282012%29/Chapter%2011%20-%20Pricing%20Strategies.pdf

53 65 https://www.anthropic.com/claude/haiku
https://www.anthropic.com/claude/haiku