

Din AI koster mest, når du prompter dårligt

Få styr på dine AI-tokens: samme svar, færre tokens, mindre strøm og vand.

GUIDE · AI-TOKENS · AUGUST 2026

Stefano Vincenti er AI-træner, konsulent og ekstern lektor ved ITU og DIS Copenhagen og co-founder af BotTellMe. Han hjælper nordiske ledelsesteams og fagfolk med at gå fra AI-nysgerrige til konkret, ansvarlig brug.

Følg på LinkedIn

Tilmeld nyhedsbrevet

linkedin.com/in/stefanovincenti · stefanovincenti.substack.com · aitrainer.dk

Jeg ramte Claude Code-loftet igen i sidste uge. Jeg vælger at se det som et tegn på produktivitet.

Men det fik mig til at kigge på, hvad jeg egentlig betaler for. Det meste af forbruget var spild. Jeg brugte AI skødesløst: lange tråde, løse prompts, hele filer proppet ind, hvor et enkelt afsnit havde rakt.

Da jeg skrev om det på LinkedIn, bad en masse af jer om tips. Så her er de. Teknikker der virker på tværs af de modeller, jeg selv bruger i hverdagen: Copilot, Claude, ChatGPT, Mistral, Gemini. Prisstrukturen ligner nok hinanden til, at de samme greb sparer penge alle steder.

PRINCIP

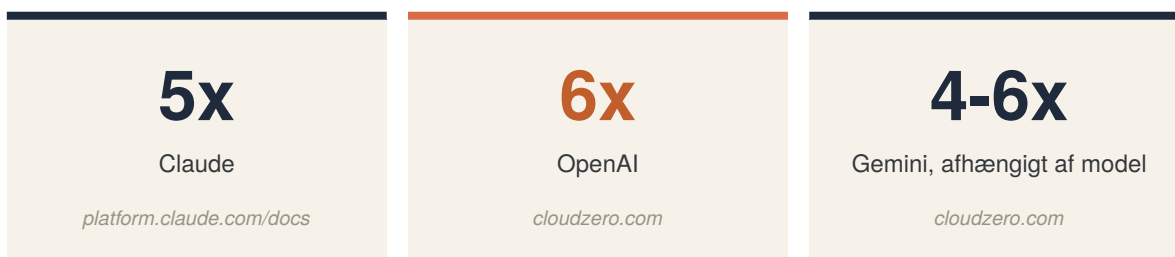
En token er et lille stykke af et ord. Du betaler per token, du sender ind, og per token, du får ud. Det lyder banalt. Det er lige præcis der, hele pointen gemmer sig.

■ 1. DET DU SENDER IND ER BILLIGT. DET DU FÅR RETUR ER DYRT

Hos stort set alle udbydere koster et svar mere per token, end det du spørger om.

Output koster et multiplum af input

Pris per output-token sammenlignet med input-token



En model der plaprer løs, koster et multiplum af en der svarer stramt

Bed om det korte svar, og sæt formen.

Model 1. Forholdet mellem output- og input-pris. Status pr. august 2026.

Så bed om det korte svar. Sæt formen. "Svar i tre punkter." "Maks 200 ord." "Giv mig kun ændringen, ikke hele filen igen." Og skru på den globale indstilling for din chatbot, hvor du kan: klik på din profil, gå under indstillinger og find stedet, hvor du vælger, at botten skal være kort og præcis.

Og pas på den skjulte post: reasoning-tokens, den "tænkning" modellen laver før svaret, afregnes som output. På en tung analyse kan tænkningen alene være 70 til 85 procent af output-regningen. Slå dyb tænkning fra, når opgaven ikke kræver den, og skru ikke for meget op på den reasoning-slider, flere af værktøjerne giver dig.

■ 2. EN LANG CHAT ER EN REGNING, DER VOKSER

Her er den misforståelse, der koster flest penge. Folk tror, modellen "husker" samtalen gratis. Det gør den ikke.

Hver gang du sender en besked, læser modellen hele samtalen igen for at kunne svare. Besked nummer 40 betaler også for besked 1 til 39. Derfor bliver en lang tråd dyrere per besked, jo længere den kører, selv om dine spørgsmål bliver ved med at være korte.

REGNEEKSEMPEL

Du har været i en chat en time. Konteksten fylder nu 50.000 tokens. Du stiller et spørgsmål på 15 ord. Modellen genlæser alle 50.000 tokens for at svare. På Claude Opus-listepris er inputtet til det ene spørgsmål omkring 0,25 dollar. Og du betaler det igen ved næste spørgsmål. Det korte spørgsmål var aldrig omkostningen. Bagagen var.

■ 3. DEN DYRESTE PROMPT ER DEN, DU SKAL SENDE TRE GANGE

Fordi hver tur frem og tilbage genafsender hele konteksten, er det i pingpongen, pengene brænder. Fem runder med løse spørgsmål koster mere end én skarp. Formulér opgaven og leverancen op front:

1. **Rolle.** Hvilken rolle AI'en skal spille.
2. **Minimal kontekst.** Kun de filer og vedhæftninger, opgaven kræver.
3. **Opgaven.** Forklar præcist, hvad opgaven er.
4. **Leverancen.** Hvilket format (fx Word), til hvem, med hvilke begrænsninger.

Det er kedeligt at skrive den prompt. Det er dyrt at lade være. Det er her, de fleste lækker mest. Og du behøver ikke et kursus i prompt engineering. Genlæs de fire punkter tre gange. Det er det, du skal vide.

■ 4. SKARP KONTEKST SLÅR STOR KONTEKST

Mere kontekst er ikke bedre kontekst. Giv modellen præcis det, den skal bruge: den relevante fil, det relevante afsnit, de relevante tal. Ikke hele mappen. Når folk spørger mig, hvorfor Copilot kun tager 20 filer, smiler jeg og forklarer, at de skal holde sig langt under de 20.

En stram, velvalgt kontekst giver bedre svar og koster mindre på samme tid. Måltrettet materiale slår et bulk-dump, også på kvaliteten af svaret. Vedhæft den ene side, ikke hele PDF'en.

■ 5. ENGELSK ER DET BILLIGSTE SPROG AT TALE MED DIN AI

De store modellers tokenizer er bygget på engelsk. Engelske ord og endelser er ofte én token. Dansk bliver hakket i flere. Det samme indhold på dansk fylder derfor flere tokens end på engelsk. Hvor meget

svinger fra model til model, men på nogle modeller kan det koste op mod 40 procent flere tokens. Æøå og lange sammensatte ord er de værste syndere.

Dansk er stadig fint til det meste. Kører du derimod en tung, gentagen opgave, hvor sproget ikke betyder noget for resultatet, for eksempel kode, dataudtræk eller struktureret analyse, så prøv at prompte på engelsk. Skriv "svar på dansk" til sidst, hvis leverancen skal være på dansk.

■ 6. PAK SAMMEN OG START FORFRA

Det her er den vane, der har rykket mest for mig. Når en chat bliver lang og træg, så stop op. Bed AI'en om at lave en handover-pakke, og åbn en frisk chat med den.

Lav en handover-pakke: et kort resumé af, hvor langt vi er, hvad der er besluttet, og hvad næste skridt er. Skriv det som en lille tekstfil, som en ny chat kan læse og fortsætte fra.

Du smider al den døde vægt og beholder kun det, der betyder noget. Billigere, hurtigere, og svarene bliver skarpere, fordi modellen ikke skal vade gennem 200 beskeder støj for at finde tråden.

■ 7. FEM GREB MERE, DER VIRKER ALLE STEDER

1. **Vælg den rigtige model til opgaven.** Simple opgaver på den lille, hurtige model, tung analyse på den store. Det alene kan skære 40 til 60 procent på en blandet arbejdsdag.
2. **Bed altid om det korte svar.** Output er den dyre ende. En længde- eller formatgrænse er den nemmeste besparelse, du har.
3. **Sluk for thinking, når du ikke skal bruge den.** Reasoning-tokens er output-tokens i forklædning.
4. **Start en ny chat til et nyt emne.** Slæb ikke gårsdagens samtale med ind i en ny opgave.
5. **Hold det stabile øverst.** De fleste udbydere cacher gentaget kontekst og giver op til 90 procent rabat på det. Læg det, der ikke ændrer sig, først i prompten.

■ 8. DER ER EN ANDEN REGNING END PENGENE

Tokens er regnekraft. Regnekraft er strøm og vand. Sam Altman har oplyst, at et gennemsnitligt ChatGPT-spørgsmål bruger omkring 0,34 watt-timer strøm og cirka en femtendedel af en teske vand. Det er OpenAI's egne tal fra juni 2025, ikke uafhængigt efterprøvet, så tag dem for hvad de er.

Ét spørgsmål er ingenting. Millioner af skødesløse spørgsmål om dagen er noget andet. Det gode er, at det er de samme vaner. Det, der sænker din regning, sænker også dit aftryk.

Skarp promptning er grøn promptning.

■ 9. TRE SPØRGSMÅL AT TAGE MED PÅ MANDAG

1. Hvor mange af jeres AI-samtaler er lange tråde, der burde have været pakket sammen for en time siden?
2. Hvem har lært jeres folk at bede om det korte svar?
3. Hvad ville I spare, i kroner og i strøm, hvis I gjorde det til en vane?

Svaret plejer at overraske. Det er den bedste anledning til at få skrevet et par enkle spilleregler for AI-brug, I nogensinde får.

■ KILDER

Priser og tal er verificeret mod disse kilder pr. august 2026.

- Claude-priser og prompt caching: platform.claude.com/docs, [about-claude/pricing](https://platform.claude.com/about-claude/pricing)
- OpenAI-priser (output 6x input): cloudzero.com/blog/openai-pricing
- Gemini-priser og thinking-tokens som output: cloudzero.com/blog/gemini-pricing
- Token-optimering (model-valg, output-grænser, caching, sessioner): tokenoptimize.dev
- Strøm- og vandtal per query (Sam Altman, juni 2025): datacenterdynamics.com
- Sprog og tokens: [arxiv.org 2305.15425](https://arxiv.org/abs/2305.15425), promptcost.org

© 2026 Stefano Vincenti · aitrainer.dk · Må deles med kildeangivelse.

Stefano Vincenti

AI-træner, konsulent og ekstern lektor · 20+ år i software og digital transformation · Co-founder af BotTellMe · Ekstern lektor ved ITU og DIS Copenhagen · Partner hos TryZone.

LinkedIn linkedin.com/in/stefanovincenti	Nyhedsbrev stefanovincenti.substack.com	Mere aitrainer.dk · tryzone.dk · bottellme.com
--	---	---

Dette er generel praktisk vejledning, ikke rådgivning du skal træffe beslutninger på alene. Alle priser er oplyst i USD ekskl. moms og kan være ændret; tjek altid leverandørernes egne prissider. Tallene for strøm og vand er OpenAI's egne, oplyst af Sam Altman i juni 2025 og ikke uafhængigt verificeret. Status pr. august 2026.

Stefano Vincenti · AI, Built Human