GRATIS KOSTENWIJZER · RUDYJELLESMA.NL

Ai-kostenwijzer voor MKB

Wat kost Ai nou echt? Prijsverschillen, een beslisboom voor modelkeuze, en een eerlijke rekensom voor je eerste toepassing.

Rudy Jellesma

rudyjellesma.nl

Werkdocument · 2026

Waarom dit kader?

De meestgestelde vraag die ik krijg over Ai is niet “wat kan het?” maar “wat kost het?”. En dat is een terechte vraag, want het antwoord verrast bijna iedereen: tussen de duurste en de goedkoopste manier om exact dezelfde Ai-taak uit te voeren zit een factor van meer dan tweehonderd. Wie dat niet weet, betaalt vrolijk het volle pond voor werk dat een fractie mag kosten.

Ik draai zelf tientallen Ai-toepassingen dag en nacht, naast al het andere werk in mijn bedrijven. Dat is alleen betaalbaar omdat ik elke taak naar het goedkoopste model stuur dat hem nog goed genoeg doet. Deze wijzer laat zien hoe die rekensom werkt, zodat je hem voor je eigen bedrijf kunt maken.

De belangrijkste regel staat meteen bovenaan:

> **Stuur niet elke Ai-vraag naar het duurste model. Routinewerk mag goedkoop; alleen echt oordeel is het dure model waard.**

De prijsverschillen: waar de factor 200 vandaan komt

Ai-modellen rekenen af per “token”, een stukje tekst van ongeveer een halve tot een hele woord. Je betaalt voor wat je erin stopt (de vraag plus context) en voor wat eruit komt (het antwoord). De prijs per miljoen tokens verschilt enorm per model.

Neem een alledaagse taak: een document van zo’n 10.000 tokens laten samenvatten. Diezelfde taak, drie prijskaartjes:

Aanpak	Prijs per miljoen (in / uit)	Kosten van deze taak	Verhouding
Duur topmodel	grofweg €14 in / €69 uit	ongeveer €0,17	1x (basislijn)
Goedkoop cloudmodel	€0,07 in / €0,28 uit	ongeveer €0,0008	222x goedkoper
Ander goedkoop cloudmodel	€0,13 in / €0,26 uit	vergelijkbaar laag	153x goedkoper
Lokaal model op eigen hardware	nul	nul	gratis

De aanbieders rekenen deze API-tarieven in dollars af; de eurobedragen hierboven zijn omgerekend tegen de koers van het moment en schommelen dus mee. Eén samenvatting van ongeveer zeventien eurocent klinkt onschuldig. Maar bij tientallen calls per dag loopt het verschil op tot honderden euro’s per maand, voor precies dezelfde uitvoer. Dat is de kern: niet de Ai is duur, maar de verkeerde Ai voor de verkeerde taak.

Het gratis lokale model heeft een addertje. Op een gewone kantoor-pc zonder zware videokaart haalt zo’n model maar drie tot acht tokens per seconde. Prima voor werk dat op de achtergrond mag draaien (nachtelijke batches, rapportages), maar te traag voor iets waar iemand live op zit te wachten. Gratis betekent hier: langzaam maar kosteloos.

De beslisboom: welk model bij welke taak?

De sleutel is dat de meeste Ai-taken helemaal geen topmodel nodig hebben. Grofweg valt werk in twee soorten.

Routinewerk (goedkoop of gratis): samenvatten, classificeren en labelen, logbestanden doorlezen, standaardteksten opstellen, gegevens uit documenten halen. Hier is het goedkope model bijna niet te onderscheiden van het dure.

Oordeelswerk (het dure model waard): meerdere risico's tegen elkaar afwegen en op volgorde zetten, complexe analyse, strategische keuzes, lastige uitzonderingen. Hier is de kwaliteit van het topmodel het geld waard.

Type taak	Voorbeelden	Aanbevolen model
Routine, veel volume	Mail labelen, facturen uitlezen, teksten samenvatten, logs scannen	Goedkoop cloudmodel of gratis lokaal
Nuance, middelmatig	Klantvraag begrijpen, concept opstellen met toon	Middenmodel
Oordeel, weinig volume	Risico's prioriteren, architectuur, uitzonderingen	Duur topmodel

De praktische regel: begin elke taak bij het goedkoopste model en schuif alleen omhoog als de kwaliteit aantoonbaar tekortschiet. Niet andersom. De valkuil is dat het dure model vaak de standaard is, en dan wint gemak het van de portemonnee.

Meet het, verzin het niet: bouw een mini-benchmark

"Aantoonbaar tekortschiet" is het sleutelwoord. Kies niet op gevoel welk model goed genoeg is, maar meet het. Dat kan verrassend simpel.

Ik heb voor mijn eigen werklast een mini-benchmark gebouwd: achttien echte taken uit de praktijk, elk met een vastgesteld goed antwoord en een beoordelingsregel. Vervolgens laat je elk kandidaat-model alle taken doen en laat je een model als scheidsrechter de antwoorden scoren tegen het goede antwoord. Zo krijg je een eerlijk cijfer per model, plus de kosten en de snelheid ernaast.

De uitkomst was leerzaam. Het goedkoopste model scoorde 87 procent, het duurste 91 procent. Op de gewone taken (triage, logs doorlezen, standaardclassificatie) zaten ze praktisch gelijk, tussen de 89 en 100 procent. Het grote verschil zat op één categorie: het rangschikken van risico's op urgentie. Daar zakte het goedkope model naar 44 procent, terwijl het dure model overeind bleef.

De conclusie schrijft zichzelf: stuur de gewone taken naar het goedkope model, en houd het dure model voor het oordeelswerk. Die gelaagde aanpak sneed fors in mijn maandelijkse Ai-rekening (in de orde van vele honderden euro's per maand) zonder dat de kwaliteit op die taken merkbaar daalde.

Een eerlijke kanttekening: achttien taken is een kleine steekproef, en een Ai die andere Ai beoordeelt kan zichzelf licht bevoordelen. Lees kleine verschillen dus niet als hard bewijs. Maar de grote lijn (routinewerk is inwisselbaar, oordeelswerk niet) komt er glashelder uit.

Koop nog geen dure hardware: de ROI-som van lokale Ai

Zodra ondernemers horen dat een lokaal model gratis draait, komt de vraag: moet ik dan een dure Ai-computer kopen? Bijna altijd is het antwoord nee, en de rekensom laat zien waarom.

Serieuze Ai-hardware kost al gauw tussen de 2.500 en 6.000 euro. Zet daar tegenover wat je aan een cloudmodel kwijt bent voor het Ai-werk rond één klant of project: in mijn ervaring zo'n 2 tot 5 euro per jaar. Op dat volume verdient de hardware zich pas over vijf tot veertig jaar terug. Dan koop je geen besparing, je koopt een hobby.

Eigen hardware wordt pas verdedigbaar als je hem als platform inzet: meerdere zware werklasten tegelijk, dag in dag uit, waarbij de cloudrekening écht oploopt. Tot die tijd is de cloud goedkoper, sneller en zorgelozer. De regel: reken de terugverdientijd uit vóór je koopt, niet erna.

De kosten-envelop: een hard budget per klantinteractie

De laatste bouwsteen maakt Ai-kosten voorspelbaar in plaats van eng. Stel voor elke terugkerende Ai-taak een maximum vast: een kosten-envelop.

Voor een Ai-gesprek met een klant (bijvoorbeeld een intake of een adviesvraag) hanteer ik een hard budget van maximaal 50 eurocent per sessie. Dat haal je door het gesprek in fasen te knippen: de simpele stappen door een goedkoop model, de stappen die nuance vragen door een middenmodel, en door slim om te gaan met herhaalde context zodat je niet steeds opnieuw betaalt voor hetzelfde. Zo blijft een complete klantinteractie ruim onder een euro.

De onderliggende regel is breder dan dit ene voorbeeld: een Ai-functie is pas verantwoord als je vooraf weet wat één transactie maximaal mag kosten. Zonder envelop weet je pas achteraf wat je uitgeeft, en dan is het te laat.

In het kort: de kostenwijzer op één pagina

1. **Dezelfde taak, tot 200x prijsverschil.** Het model kiest de rekening, niet de taak.
2. **Splits werk in routine en oordeel.** Routine goedkoop of gratis; oordeel het dure model waard.
3. **Begin goedkoop, schuif alleen omhoog als het meetbaar moet.** Niet andersom.
4. **Meet welk model goed genoeg is** met een mini-benchmark van echte taken, in plaats van te gokken.
5. **Gratis lokaal is langzaam, niet slecht.** Prima voor achtergrondwerk, niet voor live.
6. **Koop pas dure hardware als hij een platform draagt.** Anders is de terugverdientijd jaren.

7. **Zet een hard budget per klantinteractie** (bij mij maximaal 50 cent) en ken je kosten per transactie vóór je live gaat.

Tot slot

Ai hoeft geen open eind te zijn op je begroting. Wie begrijpt dat routinewerk een fractie kost van oordeelswerk, en dat vooraf meet in plaats van achteraf schrikt, houdt de kosten volledig in eigen hand. De techniek is niet duur; onnadenkend gebruik is duur.

Meer praktijkverhalen over Ai in het MKB, inclusief wat een toepassing echt kostte en opbracht, lees je op **rudyjellesma.nl**. Wil je hulp bij de rekensom voor jouw eerste Ai-toepassing? Ook daarvoor kun je daar terecht.

© Rudy Jellesma, rudyjellesma.nl. Dit document mag je vrij delen binnen je eigen organisatie.

Over de auteur

Rudy Jellesma is ondernemer en CTO. Hij bouwt en beheert AI-systemen die dag en nacht meewerken in zijn eigen bedrijven, van monitoring en codecontrole tot boekhouding en mail-afhandeling. Op rudyjellesma.nl deelt hij wat daarbij werkt en wat misgaat, steeds vertaald naar de praktijk van het MKB.

Verder lezen

Praktijkverhalen, tips en stappenplannen over AI voor het MKB vind je op `rudyjellesma.nl`.

Andere gratis downloads:

- Het Ai-stoplicht
- Ai-startklaar-checklist voor MKB
- BTW-checklist voor Ai-abonnementen
- Ai-boekhouden met akkoord-gate
- Beveilig je Ai-agent in 7 stappen
- Start je bedrijfs-kennisbank voor Ai in 1 dag

© 2026 Rudy Jellesma, rudyjellesma.nl. Dit document mag je vrij delen binnen je eigen organisatie.