JH

AI-ARCHITECTUUR · WHITEPAPER

Welk AI-model kies je? Afweging voor 2026

Waarom een AI-workflow meer opschiet met een keten van kleine modellen dan met het nieuwste frontier-model

Jaap Hoeve

De Prompters B.V.

Inhoudsopgave

Elke keer als er een nieuw model uitkomt barst het weer opnieuw los op het internet: welk model moeten we nu hebben. Ik heb negen GenAI-assistenten naar productie gebracht en bij geen daarvan was het model de reden dat het wel of niet werkte. Het komt neer op:

- De harness of workflow bepaalt je resultaat, niet het model.
- Een workflow presteert beter met keten van kleine modellen achter elkaar.
- Open modelgewichten op Europese servers zijn branding geworden, geen techniek- of compliancevraag.

Het model van een half jaar geleden is goed genoeg

Met de modellen van een half jaar geleden bouw je vrijwel alles wat je vandaag wilt bouwen. Dan roep je een minder model meerdere keren aan met een taak per aanroep. Dit noemen we prompt chaining. Op deze manier lever je latency in en krijg je er nauwkeurigheid voor terug. Doordat elke aanroep een makkelijkere taak krijgt wordt de kwaliteit van output beter. Doordat je elke stap los hebt kun je ook beter optimaliseren. Je kan per stap een model kiezen dat het beste is voor die taak. Zo kun het model per taak optimaliseren en de prompt blijven verbeteren. Je ziet namelijk goed per aanroep wat er ingaat en wat eruit komt.

Begin bij de simpelste oplossing en voeg pas complexiteit toe als het simpelere tekortschiet. AI gaat niet altijd goed. Een keten van bijvoorbeeld twaalf stappen die elk 99 procent van de tijd goed gaan gaat als geheel 89 procent van de tijd goed.

Parallel waar het kan

Zodra je de taak hebt geknipt via prompt chaining zie je welke stappen onafhankelijk van elkaar zijn. Deze kun je dus parallel van elkaar laten verlopen en zo kun je de verhoogde latency verlagen. Bijvoorbeeld; vijf velden uit een polisblad halen is vijf keer hetzelfde werk op hetzelfde document. Serieel zou dit vijf keer drie seconden duren maar als je dit parallel doet duurt het maar drie seconden. Bij AI heb je twee verschillende vormen hiervan, dit zijn:

- Sectioning is het opdelen in onafhankelijke deeltaken die tegelijk draaien.
- Voting is dezelfde taak meerdere keren uitvoeren en de uitkomsten met elkaar vergelijken.

Voting gebruik je een fout duur is of grote gevolgen heeft. Laat bijvoorbeeld het verzekerd bedrag twee keer extraheren door twee verschillende modellen en accepteer het alleen als beide hetzelfde teruggeven. Als ze

afwijken worden ze doorgestuurd voor menselijk controle. Dat kost je één extra aanroep van een fractie van een cent en geeft je zekerheid terug.

Wat er het afgelopen jaar verbeterde

De coding agent van LangChain scoorde 52,8 procent op Terminal-Bench 2.0 en stond daarmee buiten de top 30. Na een reeks aanpassingen scoorde hij 66,5 procent en stond hij in de top 5. Zonder van model te wisselen. Ze hadden alleen het harness aangepast. Dat is wat er om de AI heen zit. Denk hierbij aan vaardigheden, connectoren en informatie injectie in de prompts. Bijvoorbeeld:

- Een verplichte controleronde voordat de agent mag zeggen dat hij klaar is.
- Informatie over de omgeving die vooraf werd meegegeven in plaats van dat de agent er zelf naar moest zoeken.
- Detectie van herhaling, zodat hij niet twintig keer hetzelfde bestand blijft bewerken.
- Een ongelijke verdeling van redeneerbudget met de meeste denkruimte bij het plannen en het controleren.

Door deze aanpassingen om de AI heen hebben ze grote winst geboekt. Maar dit gaat niet voor alle AI-modellen op. Deze verbeteringen helpen de middenmoot-modellen het meest en bij modellen die ouder zijn dan drie jaar maakt extra structuur het resultaat juist slechter.

De prompt

AI-modellen kunnen nu hele boekwerken aan de in de prompt. Groter dan de Lord of the rings Trilogie zelfs. Dat betekent niet dat je alles moet gebruiken. Uit diverse onderzoeken blijkt hoeveel informatie je er instopt, hoe slechter het resultaat is. Het adagium uit de dataverwerking houdt ook in de AI-wereld stand, "Garbage in, garbage out".

Hou de prompt dus taak gericht en de informatie die je upload relevant. Zo krijg je het beste antwoord en ook het meest stabiel qua output.

Wanneer wel een nieuw sterk model?

Een nieuw sterk model heeft zekers zijn use-cases. Als latency een probleem is of als de hulpvraag minder duidelijk is heeft een nieuwer sterk model zekers voordelen. Het geeft sneller antwoord en kan beter omgaan met

onduidelijke vragen. Het geeft sneller antwoord omdat er maar 1 aanroep naar het AI-model is in plaats van meerdere aanroepen. Dat maakt het veel sneller. In moderne AI-modellen zijn ook oplossingen gebouwd die de vraag herschrijven en redeneren over wat echt de klant vraag is. Hierdoor is de kwaliteit van antwoorden veel beter ook al is de vraagstelling slechter. In dit soort gevallen is de extra investering in een duurder model het ook echt waard.

Autonomous agents zijn AI's die zelf een plan bedenken om hun probleem op te lossen. Ze zoeken hierbij zelf informatie op en roepen aanvullende databronnen aan. Dit maakt modellen die minder krachtig zijn minder geschikt voor hen. Voor deze oplossingen wil je wel de nieuwe modellen gebruiken. In de praktijk zie ik weinig use-cases voor autonomous agents maar dit was wel het vermelden waard.

Het kostenplaatje

Een keten gebruikt meer tokens dan één grote prompt want per stap stuur je telkens opnieuw de gehele context mee. Het prijsverschil per token maakt dat ruimschoots goed.

	Input per miljoen tokens	Output per miljoen tokens
Frontier	$5,00 (GPT-5.5, Claude Opus 5)	$25 tot $30
Middenklasse	$2,00 (Claude Sonnet 5, Gemini 3 Pro)	$10 tot $12
Open gewichten	$0,10 tot $0,28 (DeepSeek V3.2, Mistral Small 3.2)	$0,30 tot $1,10
Klein en snel	$0,03 (Qwen3.7 Flash)	$0,13

Peildatum augustus 2026.

Een keten van vijf aanroepen op een open model kost minder dan één aanroep op een frontier-model. Met slim model management kan prompt chaining dus zelfs goedkoper worden. Je moet wel scherp zijn op welk model gebruikt en of deze taak voldoende kan uitvoeren.

Chinese modellen op Europese servers

De sterkste open source modellen komen op dit moment uit China. Open source modellen zijn modellen die je kunt downloaden en zelf kan draaien. Je hoeft dus niet afhankelijk te zijn van de Amerikaanse tech om goede modellen te kunnen gebruiken.

Model	Maker en herkomst	Licentie
DeepSeek V4	DeepSeek, China	MIT
Qwen 3.x	Alibaba, China	Apache 2.0
GLM-5.2	Z.ai, China	MIT
Kimi K3	Moonshot, China	Eigen licentie
Mistral Small en Large	Mistral, Frankrijk	Apache 2.0 en eigen licentie
Llama 4	Meta, Verenigde Staten	Community-licentie

Je hoeft ze niet per se zelf te hosten als je dat niet wilt. Het is namelijk nogal een technische aangelegenheid en dan moet je ook de security, monitoring enzovoorts regelen. Dat is een overhead waar weinig bedrijven op zitten te wachten. Deze aanbieders kunnen je er goed mee helpen:

- Scaleway host al zijn modellen in één datacenter in Parijs en analyseert, verzamelt, leest of hergebruikt de inhoud van je invoer, prompts en uitvoer niet. Deze zero-data retention is mijn mening een groot pluspunt.
- Nebius biedt een route via Finland waarbij je data in de EU blijft met een verwerkersovereenkomst en de toezegging dat er niet op je data wordt getraind. Dit is tegenwoordig een Nederlands bedrijf maar het is wel belangrijk om te weten dat dit vroeger onderdeel van Yandex was.

Open source modellen zijn veilig te gebruiken. Modellen zijn namelijk paden die tot een uitkomst lijden, de bestanden zijn niet uitvoerbaar. Wat betekent dat een Chinees model zijn data niet terug kan sturen naar China. Daarom is het veilig om deze modellen te gebruiken bij partijen waarvan je gecontroleerd hebt dat ze hun zaakjes goed voor elkaar hebben. Daarom gaf ik de bovenstaande bedrijven als tip. Die draaien al een tijdje mee en hebben hun zaakjes goed voor elkaar. In dit geval zou ik niet voor een compleet nieuwe partij gaan. Als je

toch liever zekerheid wilt hebben, kies Mistral. Wel Europees en degelijke modellen. Ze hebben niet de nieuwste modellen maar in de praktijk zie ik weinig taken wat zij niet zouden kunnen doen.