# De dure AI eet zichzelf op: Agents’ Last Exam en de kans voor open modellen

11-06-2026


**Simpele kern:** de duurste AI-modellen worden steeds beter in benchmarks, maar falen nog op het werk dat bedrijven echt willen uitbesteden. Daardoor wordt de premium-tokenprijs geen statussymbool, maar zelfkannibalisme.

## Het nieuws: GPT-5.5 wint, maar niemand wint echt

VentureBeat 1 bericht dat OpenAI’s GPT-5.5, via de Codex-harness, bovenaan staat op de nieuwe **Agents’ Last Exam**-leaderboard met een pass rate van **24,0%** 1. Claude Fable 5, via Claude Code, komt op **22,0%** en eindigt derde, achter een tweede GPT-5.5-configuratie 13.

Dat klinkt als een nieuwsfeit over modelranglijsten. Maar het echte verhaal is groter: zelfs de beste setup haalt nog geen kwart van de echte professionele workflows volledig goed 3. Agents’ Last Exam, afgekort ALE, is juist gebouwd om te meten of AI-agents economisch waardevol, langdurig en controleerbaar werk kunnen uitvoeren 2.

Dit is geen quiz met meerkeuzevragen. ALE laat agents werken in echte softwareomgevingen: bestanden bewerken, GUI's bedienen, scripts draaien, domeinspecifieke tools gebruiken en uiteindelijk een controleerbaar resultaat opleveren 12. De benchmark kijkt daarbij naar meerdere lagen: redeneren, visie, orchestration, toolgebruik en runtime 1.

De uitslag is daarom hard: de modellen zijn beter geworden, maar de kloof tussen benchmarkprestatie en bruikbare arbeid blijft enorm.

## Waarom deze benchmark anders is

Eerdere AI-benchmarks hadden vaak twee zwakke plekken. Ten eerste waren ze te smal: losse codeeropdrachten, korte terminaltaken of statische vraag-antwoordsets 1. Ten tweede was de beoordeling vaak onbetrouwbaar, met automatische graders die goede oplossingen konden afkeuren of modellen die op een of andere manier langs de rand van de test konden schuiven 1.

ALE probeert dat dicht te timmeren. De benchmark gebruikt grotendeels deterministische checks in plaats van "LLM-as-a-judge" 1. Volgens VentureBeat wordt die methode nog maar voor **6,8%** van de workflows gebruikt 1. De paper benadrukt dat ALE is ontworpen als een instrument om de kloof tussen benchmarksuccessen en economisch relevante impact te dichten 2.

Dat maakt de uitkomst ongemakkelijk voor Big Tech. Dit is geen bench waar je met een mooie demo doorheen glijdt. Dit is een examen over werk dat iemand anders anders zelf zou moeten doen.

## De prijsclaim knapt op de werkelijkheid

Hier sluit dit nieuws direct aan bij de eerdere analyse op deze site: Big Tech geeft steeds duidelijker toe dat AI goedkoper moet worden 56. De nieuwe prijs van Claude Fable 5 en Claude Mythos 5 is **$10 per miljoen input tokens** en **$50 per miljoen output tokens** 7. Dat is niet alleen duur; het is strategisch riskant.

Want agents zijn tokenverbranders. Ze plannen, herhalen, controleren, roepen tools aan, lezen screenshots, schrijven bestanden, herstellen fouten en proberen opnieuw. Gizmodo signaleerde al dat bedrijven als Amazon en Uber intern remmen op tokengebruik, en citeerde een preprint-studie waarin agents tot **1.000 keer zoveel tokens** zouden gebruiken als andere AI-systemen 6.

ALE bevestigt de richting van dat probleem zonder dat je de exacte tokenclaim hoeft te geloven. De leaderboard toont bij de topconfiguratie Codex met GPT-5.5 een totale runtime van bijna **370 uur**, **1,6 miljard input tokens** en **7,2 miljoen output tokens** over de uitgevoerde runs 3. Claude Code met Claude Fable 5 komt uit op bijna **198 uur**, **886,6 miljoen input tokens** en **9,6 miljoen output tokens** 3.

Dat is de kern van het zelfkannibalisme: hoe ambitieuzer de agent, hoe langer hij werkt; hoe langer hij werkt, hoe meer tokens hij verbrandt; en hoe hoger de tokenprijs, hoe sneller de businesscase verdwijnt.

## De dure modellen eten hun eigen markt op

Premium-modellen kunnen zich verkopen als “de slimste”. Maar bedrijven kopen geen slimheid als abstracte eigenschap. Bedrijven kopen afgerond werk.

Als een agent op echte workflows maar rond de 20–24% perfect haalt, dan moet je als klant drie vragen stellen:

1. Hoeveel menselijke controle blijft er nodig?
2. Hoeveel tokenbudget verbrandt de agent voordat hij faalt?
3. Is de besparing nog echt groter dan de rekening?

Bij $10/$50 per miljoen tokens wordt dat lastig. Zeker wanneer output duurder is dan input. Een agent die veel denkt, veel terugleest, veel herprobeert en veel tooloutput produceert, kan de prijs van zijn eigen waarde opeten.

Dat is waarom “meer AI-gebruik” niet automatisch “meer productiviteit” betekent. Tokenmaxxing was even een meme, maar de stemming kantelt 6. De markt begint te merken dat onbegrensde tokenconsumptie geen strategie is, maar een lek.

## Waarom dit goed nieuws is voor open modellen

Open modellen hoeven niet per se de grootste te zijn om te winnen. Ze moeten wel bewijzen dat ze bruikbaar zijn. En precies daar ontstaat de opening.

### 1. De strijd verschuift van prestige naar werk

Als een benchmark echt meet of een agent werk afmaakt, wordt marketing minder belangrijk. Dan telt niet alleen “welk model is het grootste?”, maar “welk model haalt mijn workflow af voor een prijs die ik kan schalen?”

Dat is gunstig voor open-weight en open-source modellen, zolang ze hun prestaties net zo hard testen als de gesloten labs dat doen.

### 2. Goedkoop wordt een feature, geen troostprijs

De prijslijst in VentureBeat laat zien dat er modellen zijn die veel lager zitten dan de $10/$50-claim van Claude Fable 5 en Mythos 5 7. Voor veel workflows is

dat belangrijker dan het laatste procentje benchmarkglans.

Een model dat 90% van een specifieke taak haalt voor een fractie van de prijs kan zakelijk sterker zijn dan een duur model dat 10% beter redeneert maar de rekening onbetaalbaar maakt.

### 3. Edge en lokale modellen worden rationeler

De eerdere blog op deze site wees al op de pivot richting edge computing: kleinere modellen dichter bij het apparaat, minder afhankelijk van cloudtokens 5. Dat is geen romantisch open-source-verhaal. Het is economische zelfverdediging.

Niet elke taak heeft een frontiermodel nodig. Veel dagelijks werk heeft een slanker model nodig dat snel, voorspelbaar en betaalbaar genoeg is om continu te gebruiken.

### 4. Open evals worden belangrijker dan open branding

ALE laat zien dat de toekomst niet draait om vertrouwen op mooie screenshots van demo’s. De toekomst draait om reproduceerbare, verifieerbare workflows.

Open modellen kunnen daarop inspelen door eigen mini-ALE’s te bouwen: interne benchmarks gebaseerd op echte bedrijfsprocessen, met kosten per voltooide taak in plaats van alleen score per benchmark.

## De valkuil voor open modellen

Dit is geen vrije overwinning. “Goedkoop” is niet genoeg.

Open modellen moeten oppassen dat ze niet in de omgekeerde val trappen: roepen dat ze betaalbaar zijn, maar niet bewijzen dat ze werk afmaken. De

kans zit niet in goedkope tokens alleen. De kans zit in **goedkope tokens plus betrouwbare uitvoering**.

Daarom is de juiste strategie:

- meet kosten per voltooide taak;
- bouw eigen workflows, niet alleen benchmarkhype;
- gebruik kleine modellen waar kleine modellen genoeg zijn;
- houd een fallback voor moeilijke stappen;
- voorkom agent-loops zonder budgetplafond;
- vergelijk modellen op afgerond werk, niet op marketingclaims.

## Factcheck

| Claim | Status | Toelichting |
|---|---|---|
| GPT-5.5 via Codex staat bovenaan ALE met 24,0% pass rate | Bevestigd | De ALE-leaderboard toont Codex met GPT-5.5 op plaats 1 met 24,0% pass rate 3. |
| Claude Fable 5 via Claude Code haalt 22,0% en eindigt derde | Bevestigd | De leaderboard toont Claude Code met Claude Fable 5 op plaats 3 met 22,0% pass rate 3. |
| ALE test echte professionele workflows | Bevestigd | De paper beschrijft ALE als benchmark voor langdurige, economisch waardevolle taken met verifieerbare uitkomsten 2. |
| ALE omvat 55 subdomeinen en meer dan 1.000 taken | Bevestigd | De paper noemt 55 subfields, 13 industry clusters en 1K+ taken 2; de projectsite spreekt over 1.500+ verzamelde taken richting 5.000 4. |

| Claim | Status | Toelichting |
| --- | --- | --- |
| Claude Fable 5 en Mythos 5 kosten $10/$50 per miljoen tokens | Bevestigd | VentureBeat rapporteert deze prijzen en plaatst ze in een bredere API-prijstabel 7. |
| Agents kunnen extreem veel tokens gebruiken | Secundair bevestigd | Gizmodo citeert een april 2026-preprint over 1.000× tokengebruik; dit blijft een secundaire bron tot de primaire paper zelf is gecontroleerd 6. |

## Conclusie

GPT-5.5 wint Agents’ Last Exam, maar de echte winnaar is de prijsdruk.

De nieuwe benchmark maakt de lucht leeg. Niet “welk model klinkt het meest frontier?”, maar “welk model voltooit werk dat mensen anders zelf zouden doen?” is de vraag die overblijft.

En daar zit de kans voor open modellen. Niet omdat ze per definitie beter zijn. Wel omdat de markt begint te begrijpen dat duur niet automatisch waardevol is. De middenmoot verdwijnt. De demo’s worden duurder. De rekening wordt zichtbaarder.

De AI-industrie zit in zelfkannibalisme. Big Tech verkoopt premiumtokens voor werk dat zelfs de beste agents nog maar beperkt halen. Open modellen krijgen daardoor een opening: kleiner, scherper, beter meetbaar en vooral betaalbaar genoeg om echt te schalen.

De volgende vraag is niet: “Kunnen open modellen GPT-5.5 verslaan?”

De betere vraag is: **welk model haalt jouw werk af zonder je failliet te tokeniseren?**

## Bronnen


1 VentureBeat — Carl Franzen, “Surprise upset: GPT-5.5 beats Claude Fable 5 on brutal new Agents’ Last Exam benchmark” (10 juni 2026) — https://venturebeat.com/technology/surprise-upset-gpt-5-5-beats-claude-fable-5-on-brutal-new-agents-last-exam-benchmark — **Betrouwbaarheid: 6/10** — Technisch-nieuwsbron; secundair, maar met concrete leaderboard- en benchmarkdetails.

2 Sun et al. — “Agents’ Last Exam” (arXiv:2606.05405, 3 juni 2026) — https://arxiv.org/abs/2606.05405 — DOI: https://doi.org/10.48550/arXiv.2606.05405 — **Betrouwbaarheid: 8/10** — Primaire preprint van het ALE-team; niet peer-reviewed, maar essentieel als bron voor methodologie en benchmarkdoel.

3 Agents’ Last Exam Leaderboard — ALE-V1 leaderboard — https://agents-last-exam.org/leaderboard — **Betrouwbaarheid: 8/10** — Primaire leaderboard van het project; betrouwbaar voor gerapporteerde scores, maar onafhankelijke audit blijft wenselijk.

4 Agents’ Last Exam — Projectwebsite — https://agents-last-exam.org — **Betrouwbaarheid: 8/10** — Primaire projectpagina; deels promotioneel, maar nuttig voor scope, takenaantal en participatie.

5 Roelf Renkema — “Big Tech geeft het toe: AI moet goedkoper worden” — https://www.roelfrenkema.eu/index.php?s=big-tech-ai-goedkoper — **Betrouwbaarheid: 5/10** — Eerdere analyse op deze site; synthese van

secundaire bronnen, bruikbaar als context maar niet als primaire bron.

6 Gizmodo — Webb Wright, “Big Tech Is Quietly Admitting That If It Wants to Sell People on AI, It Better Be Cheap” (8 juni 2026) — https://gizmodo.com/big-tech-is-quietly-admitting-that-if-it-wants-to-sell-people-on-ai-it-better-be-cheap-2000768710 — **Betrouwbaarheid: 6/10** — Technisch-nieuwsbron; citeert Business Insider, OpenAI-evenement en een preprint over agent-tokengebruik.

7 VentureBeat — Carl Franzen, “Anthropic brings Mythos to the masses with Claude Fable 5, its most powerful generally available model ever” (9 juni 2026) — https://venturebeat.com/technology/anthropic-brings-mythos-to-the-masses-with-claude-fable-5-its-most-powerful-generally-available-model-ever — **Betrouwbaarheid: 6/10** — Secundaire tech-nieuwsbron met prijstabel en modelcontext.

---

✍ *Geschreven door Gemi & Roelf — AI-ondersteunde content creatie*