

Codex-doorwerkproef: zes geslaagde rondes, herstel nog niet bewezen

In het kort

- Twee kleine, geïsoleerde testtaken. Beide met drie rondes.
- Baseline (bestaande werkwijze): 3 van de 3 rondes geslaagd.
- Variant met een expliciet actief doel (create_goal): ook 3 van de 3 geslaagd.
- Geen aangetoond verschil tussen de twee. Dus geen bewijs dat het actieve doel iets oplost.
- De oude fout (te vroeg stoppen) is in deze proef niet opnieuw opgetreden, dus ook niet gerepareerd bewezen.
- Geen instellingen, hooks, app's of agentconfig aangepast. Niets uitgerold.

Wat is getest?

Elke ronde had drie stappen: een invoerbestand maken en lezen, bedragen berekenen naar een tweede bestand, en onafhankelijk controleren met een derde bewijsbestand.

Tijdens het openstaande werk kwam steeds dezelfde tussentijdse vraag: "Wat is de stand nu?" Na die vraag ging het werk gewoon door, zonder extra aansporing.

CONDITIE	RONDE	ONDERBREKING NA	CORRECT TOTAAL	UITKOMST
Bestaande werkwijze	1	7,48 s	66,00	Geslaagd
Bestaande werkwijze	2	8,03 s	42,70	Geslaagd
Bestaande werkwijze	3	21,95 s	28,47	Geslaagd
Expliciet actief doel	1	15,56 s	66,00	Geslaagd
Expliciet actief doel	2	13,62 s	42,70	Geslaagd

CONDITIE	RONDE	ONDERBREKING NA	CORRECT TOTAAL	UITKOMST
Expliciet actief doel	3	13,40 s	28,47	Geslaagd

Alle achttien bestanden zijn onafhankelijk teruggelezen en gecontroleerd. Bedragen en totalen zijn apart nagerekend.

De minimale aanpassing

Alleen de tweede proef gebruikte per ronde de bestaande `create_goal`-functie. Die zette een actief doel aan; na controle bevestigde `update_goal` steeds dat het doel compleet was. Drie activeringen en drie voltooiingen zijn teruggevonden.

Let op: dit is alleen in de proef gebruikt. Er is niets algemeen ingesteld. De testdoelen zijn na afloop weer op voltooid gezet.

Wat verklaart het eerdere stoppen?

Eén eerder voorbeeld is opnieuw bekeken: een moment waarop het model zelf zei nog werk te moeten doen ("Ik moet zelf verder met de lege huiswerklíjst..."), en vlak daarna toch als klaar afsloot zonder dat er tussendoor iets is uitgevoerd.

Dat bevestigt dat het om een gewone, premature beurtafsluiting ging. Waaróm het model toen voor een eindantwoord koos, is met deze proef niet vastgesteld. Het verliezen van taakcontinuïteit tijdens een statusvraag is een aannemelijke verklaring, maar deze proef lokte die fout niet opnieuw uit. Het automatisch hervatten dat een actief doel mogelijk ondersteunt, is dus ook niet echt op de proef gesteld: de testagents gingen vanzelf door tot het einde.

Grenzen en aanbeveling

- Beide taken draaiden op hetzelfde model met dezelfde instelling (low).
- Het waren korte, nieuwe taken. Ronde 2 en 3 hergebruikten hun taakcontext.

- De onderbreking was aangekondigd en kwam binnen als delegatie-bericht, niet als een vraag die Jamal zelf rechtstreeks typte of insprak.
- De proef mist dus: lange sessies, hoge contextdruk, onduidelijke projectranden en overdrachten tussen bots. Precies de situaties waarin het probleem eerder wél optrad.

Advies: het actieve doel is een werkende kandidaat om verder te testen, geen bewezen oplossing. Niet fleet-breed uitrollen op basis van deze zes rondes.

De volgende zinnige proef:

1. Eén echte meerstapstaak met objectieve acceptatiecriteria.
2. Een gewone statusvraag via de gebruikersinterface (niet via delegatie).
3. Langere/drukkere context, om de oude fout eerst daadwerkelijk te reproduceren.
4. Pas dáárna kan een verbetering aangetoond worden. Extra instructietekst alleen is nu onvoldoende onderbouwd als oplossing.

Conclusie

De afgesproken beperkte proef is afgerond. Er is geen aangetoonde verbetering. Er draait geen vervolgtest automatisch.