

Welk agent-harnas werkt best met welk model?

Onderzoek voor Jamal naar wat andere mensen concludeerden na het testen van meerdere modellen. Focus: Hermes (jouw harnas), Codex, en de bredere vergelijking. Alleen bronnen die echt gelezen zijn.

De kern (als je maar 6 regels leest)

- **Hermes is een echt, publiek harnas van Nous Research.** Bewust model-onafhankelijk (300+ modellen). Er is geen publieke "beste model voor Hermes"-test. Jij kiest zelf per taak.
 - **Let op de naam:** "Nous Hermes" = taalmodellen. "Hermes Agent" = het harnas. Zelfde bedrijf, twee producten. Jouw Hermes = het harnas.
 - **Codex CLI is gebouwd op OpenAI (GPT-codex)**, maar kan ook andere/lokale modellen draaien via een vlag en config.
 - **Het harnas telt vaak zwaarder dan het model.** Zelfde model door 8 harnassen gaf 68% tot 88% succes: 20 punten verschil, puur door het harnas.
 - **"Harnas hoort bij één model" is half waar.** Claude Code is geoptimaliseerd voor Claude, Codex voor GPT. Maar tests met hetzelfde model in beide vonden geen harde lock-in. De ingebouwde gedrag-instructies (plannen, controleren) maken het verschil, niet het merk.
 - **Praktijk van ervaren bouwers:** Claude Code en Codex naast elkaar. Claude voor architectuur en lange sessies, Codex voor betrouwbaar afleveren en delegeren.
-

1. Hermes

Hermes Agent is een open-source agent-framework van Nous Research (MIT-licentie). Dezelfde kern draait als CLI, TUI, een native desktop-app, web-dashboard en op Telegram, Discord, Slack, WhatsApp en iMessage. Dat past precies bij jouw opzet.

Het is model-onafhankelijk. De site noemt toegang tot ruim 300 modellen en werkt met elke provider. Er staat nergens dat het op één model is afgestemd.

Eerlijk: er is geen goede publieke test gevonden van iemand die meerdere modellen in Hermes vergeleek en een winnaar aanwees. Dat wordt nog nauwelijks openbaar besproken. Het framework is juist bedoeld om elk model te draaien.

Bron: <https://hermes-agent.nousresearch.com/>

2. Codex CLI (OpenAI)

Codex is duidelijk gebouwd op OpenAI-modellen (de GPT-codex-familie is de standaard).

Codex is open source en kan ook lokale of niet-OpenAI-modellen draaien via een oss-vlag en eigen providers in de config (bijvoorbeeld Ollama, LM Studio, OpenRouter). Let op: dit komt uit guides die de onderzoeker niet zelf opende, dus behandel het als waarschijnlijk waar, niet hard bewezen.

Een tester (Prathit Joshi, Composio, augustus 2026) draaide hetzelfde niet-OpenAI-model in zowel Codex als Claude Code en kreeg een gelijke score. Zijn conclusie: de harnessen zijn geen dichtgetimmerde koker voor hun eigen model. Codex won bij hem net op betrouwbaarheid en delegeren. Claude was "beter op een goede dag, maar niet betrouwbaar op commando".

Bron: <https://composio.dev/content/claude-code-vs-openai-codex>

3. De bredere vergelijking

Composio, "8 Best AI Agent Harnesses" (Sunil Kumar Dash, augustus 2026). Zij haalden één model door 8 harnessen. Zelfde model, wild verschillende resultaten: 68% tot 88% succes. Les: het harnas doet er enorm veel toe, los van het model.

Bron: <https://composio.dev/content/best-ai-agent-harnesses>

Latch Bio, "Why An Open Source Harness Outperforms Claude Code" (Joel Simonoff, juli 2026). Zij testten hetzelfde model in twee harnessen (een open-source harnas "Pi" tegen Claude Code). De winnaar wisselde per taak. Conclusie: de ingebouwde gedrag-instructies van het harnas verklaren het verschil, niet het model. Citaat: "twee regels gedrag-prompting kunnen de prestatie flink beïnvloeden".

Bron: <https://blog.latch.bio/p/why-an-open-source-harness-outperforms>

AIMultiple (Şevval Alper, augustus 2026). Draaide elke agent-CLI op hetzelfde model om puur het harnas te meten. Noemt Claude Code "strak geïntegreerd met de Claude-modelfamilie" en koploper op de SWE-bench-test.

Bron: <https://aimultiple.com/agent-harness>

4. Is een harnas afgestemd op één model?

Het bewijs wijst twee kanten op.

Vóór het idee: Claude Code is aantoonbaar strak verweven met Claude. Codex heeft GPT-codex als standaard. Composio zegt letterlijk dat Claude Code anders kan presteren met Claude, en Codex anders met een OpenAI-model.

Tegen een harde koppeling: de twee tests die écht hetzelfde model door meerdere harnassen haalden vonden geen bewijs dat een harnas alleen goed werkt met zijn eigen model. Het harnas-ontwerp verklaart meer dan het merk van het model.

Kort: niemand bewees hard dat een model slechter draait in het "verkeerde" harnas puur om het merk. De grote verschillen komen van hoe het harnas is opgebouwd.

Wat dit voor jouw vloot betekent

- Kies je harnas net zo bewust als je model. Het harnas kan 20 punten verschil maken.
 - Hermes mag je vrij met elk model voeden. Test per taak welk model wint, er is geen vaste "beste".
 - In Codex houd je voor topwerk GPT-codex. Wil je goedkoop of lokaal, dan kan het, maar test het eerst.
 - Twee harnassen naast elkaar draaien is normaal en slim: elk zijn sterke kant.
-

Onzeker of niet gevonden

- Een directe "beste model per harnas"-uitspraak voor Hermes specifiek: niet gevonden.
- De Codex-config voor lokale modellen: alleen via guides gezien, niet zelf geopend.

- Waarschuwing bij dit onderwerp: er staat veel SEO-spam en onbetrouwbare pagina's tussen de zoekresultaten. Dit verslag gebruikt alleen echt gelezen bronnen.