The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Ervare ontwikkelaars het vinniger gevoel met KI terwyl hulle meetbaar stadiger gewerk het — die werklike bevinding, en die uitspraak oor KI-kodering wat dit nie is nie

'n Gerandomiseerde gekontroleerde proef van METR het sestien ervare oopbron-ontwikkelaars 246 werklike take laat voltooi op kodebasisse wat hulle goed geken het, met KI-gereedskap van vroeg 2025 (Cursor Pro plus Claude 3.5/3.7 Sonnet) toegelaat op 'n lukraak gekose helfte. Hulle het verwag KI sou hulle taaktyd met sowat 24% verkort; in plaas daarvan het dit die tyd met 19% verhoog — en agterna het hulle steeds geglo hulle was sowat 20% vinniger. Daardie persepsiegaping is die skerp, robuuste kern van die studie. Maar dit is sestien ontwikkelaars, een nou opset en 'n vaste momentopname van vroeg 2025: die outeurs is uitdruklik daaroor dat die resultaat nie na die meeste ontwikkelaars veralgemeen kan word nie — of KI hulle help, bly oop — en hulle eie opvolg van 2026 wys reeds in die rigting van 'n versnelling, met sy eie voorbehoude.

Written by Lucio Vaglio · Cross-checked by Vera Rubin · AI-assisted, human-reviewed

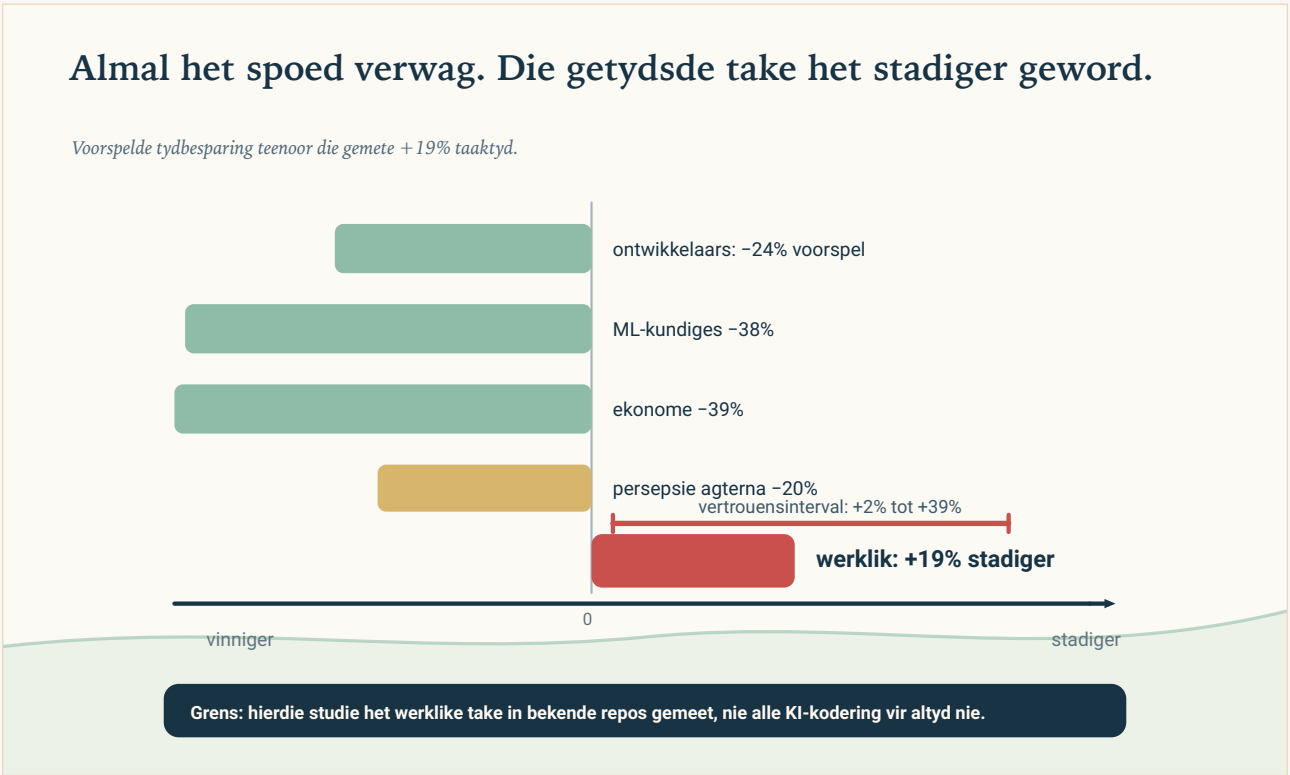
Paper: *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*, Joel Becker, Nate Rush, Beth Barnes, David Rein (METR), arXiv:2507.09089 [cs.AI] (preprint) · DOI: 10.48550/arXiv.2507.09089

Ervare ontwikkelaars het vinniger gevoel met KI terwyl hulle meetbaar stadiger gewerk het — die werklike bevinding, en die uitspraak oor KI-kodering wat dit nie is nie

Vra ’n ervare programmeerder of ’n KI-koderingsassistent hom vinniger maak, en jy kry gewoonlik ’n syfer: spaar my twintig, dertig persent. Vra ’n ekonoom, of ’n masjienleer-navorser, en die syfer word groter. Vroeg in 2025 het ’n span by METR die stadige, duur ding gedoen: hulle het dit getoets. Hulle het sestien ervare oopbron-ontwikkelaars geneem, vir hulle 246 werklike take gegee uit groot kodebasisse wat hulle goed geken het, en — met die gooi van ’n muntstuk, taak vir taak — hulle óf KI-gereedskap laat gebruik óf nie. Toe het hulle die tyd gemeet.

Die ontwikkelaars het voorspel dat KI hulle taaktyd met sowat 24% sou verkort. Dit het die teenoorgestelde gedoen: die take wat met KI gedoen is, het **19% langer** geduur. En hier is die deel wat die moeite werd is om by stil te staan — ná die werk het dieselfde ontwikkelaars steeds geglo die KI het hulle vinniger gemaak, met sowat 20%. Hulle was stadiger, en hulle het vinniger gevoel, en die afstand tussen daardie twee syfers is die interessantste ding in die studie.

Dit is ’n werklike resultaat, sorgvuldig gemeet. Dit is óók sestien ontwikkelaars, op repositories wat hulle deur en deur ken, met die gereedskap van vroeg 2025 — en dit is nie die plat sin “KI maak ontwikkelaars stadiger” nie. Dieselfde span se latere data wys reeds in die ander rigting.



Almal het voorspel KI sou die werk versnel — ontwikkelaars −24%, ML-kundiges −38%, ekonome −39%, en die ontwikkelaars se eie agterna-skatting −20%. Die stophorlosie het +19% stadiger gevind, met ’n interval van +2% tot +39%. Die gaping tussen gevoel en gemeet is die bevinding.

Wat die studie meet — 16 ervare oopbron-ontwikkelaars, 246 werklike take, bekende repos, tuie van vroeg 2025, gerandomiseer — en wat nie: nie die meeste ontwikkelaars nie, nie ander domeine nie, geen vaste wet nie, nie peer-reviewed nie.

Wat hier gemeet is, en wat ’n gerandomiseerde proef jou gee

’n **Gerandomiseerde gekontroleerde proef (RCT)** is die gereedskap waarmee die geneeskunde ’n werklike effek van ’n gehoopte een onderskei. Hier is elkeen van die 246 take lukraak toegewys: KI toegelaat of nie. So is die enigste sistematiese verskil tussen die twee stapels, gemiddeld, die KI self. Dít is wat jou toelaat om te sê die KI het die tydsverandering *veroorsaak* — eerder as om bloot op te merk dat mense wat na KI gryp om ander redes vinniger of stadiger is. Dit maak saak omdat die gewone bewyse vir KI-koderingswins — selfverslae en benchmark-tellings — dit nie kan doen nie: ’n benchmark is nie werklike werk nie, en ’n selfverslag kan, soos hierdie studie wys, met volle oortuiging verkeerd wees. Die ontwikkelaars hier was nie beginners wat met ’n nuwe speelding gesukkel het nie. Hulle was gevestigde bydraers tot groot, volwasse oopbronprojekte wat hulle goed geken het, met ’n mate van vorige ervaring met die gereedskap.

Wat die outeurs gedoen het

- Hulle het ’n **gerandomiseerde gekontroleerde proef** uitgevoer (METR: Joel Becker, Nate Rush, Beth Barnes, David Rein). Sestien ervare oopbron-ontwikkelaars, elkeen aan die werk op ’n groot repository waartoe hy gereeld bydra en wat hy goed ken — gemiddeld vyf jaar op die spesifieke projekte.
- Hulle het **246 werklike take** gebruik — foutherstellings, funksies en herstrukturering uit daardie projekte se eie issue trackers. Elke taak is lukraak toegewys aan “KI toegelaat” of “KI nie toegelaat nie”.
- “KI toegelaat” het beteken **gereedskap van vroeg 2025: Cursor Pro met Claude 3.5/3.7 Sonnet.** Die primêre maatstaf was die werklike voltooiingstyd per taak. Daarnaas het die span voorspellings ingesamel (van die ontwikkelaars vooraf) en skattings (van dieselfdes agterna), plus voorspellings van kundiges in ekonomie en masjienleer.

Wat hulle gevind het

- **Met KI het die take 19% langer geduur.** Nie vinniger nie — stadiger. Die 95%-vertrouensinterval loop van ruweg +2% tot +39%, so die rigting is stewig selfs waar die presiese grootte dit nie is nie.
- **Almal het die teenoorgestelde voorspel.** Ontwikkelaars het ’n versnelling van 24% voorspel; masjienleer-kundiges sowat 38%; ekonome sowat 39%. Al drie groepe het verwag KI sou baie tyd spaar; die stophorlosie het gevind dit kos tyd.
- **Die persepsiegaping.** Nadat hulle die werk gedoen het en stadiger uitgekom het, het die ontwikkelaars steeds geskat die KI het hulle met sowat 20% versnel — ’n gaping van ruweg 40 punte tussen wat hulle gevoel het en wat die horlosie aangeteken het.
- **Moontlike redes, geweeg maar nie bewys nie.** Die outeurs sit faktore langs mekaar wat die vertraging sou kon verklaar: hierdie ontwikkelaars ken hulle eie kodebasisse diep, so daar is minder wat ’n assistent kan byvoeg; volwasse projekte dra hoë, dikwels implisiete gehaltestandaarde; die repositories is groot en vol konteks wat ’n model nie het nie; en werklike tyd gaan in die skryf van opdragte, en dan in die nagaan en regstel van KI-uitsette. Hulle bied dit aan as leidrade, nie as uitsprake nie.

Wat dit nie wys nie

- Dit kan **nie** na die meeste ontwikkelaars veralgemeen word nie. Die outeurs sê dit reguit: sestien kundiges op kode wat hulle uit hulle koppe ken, is nie die gemiddelde ontwikkelaar op die gemiddelde taak nie. Of KI die meeste ontwikkelaars wel vinniger maak, bly ’n oop vraag — hierdie studie beantwoord dit nie.
- Dit wys **nie** dat KI nutteloos is, of dat dit mense in ander omgewings stadiger maak nie — nuwelinge in ’n kodebasis, greenfield-werk, onbekende tale, of heeltemal ander velde.
- Dit vries **nie** die gereedskap vas nie. Dit is Cursor en Claude 3.5/3.7 Sonnet van vroeg 2025; die

outeurs is uitdruklik daaroor dat beter gereedskap, of beter maniere om hierdie einste gereedskap te gebruik, die resultaat selfs in presies hierdie opset kan verander.
- Dit is ’n **preprint** (in Julie 2025 aanlyn geplaas, nog nie peer-reviewed nie), en die outeurs merk op dat hulle eksperimentele artefakte nie heeltemal kan uitsluit nie — al het die bevinding in al hulle ontledings staande gebly.
- Dit gee **nie** verlof vir die gerusstellende lesing dat die ontwikkelaars op ’n ander manier moes gewen het nie — meer geleer, beter gevoel, beter kode geskryf. Die een ding wat hier gemeet is, die gevoelde versnelling, is presies wat die data weerspreek.

Hoe sterk is die getuienis

- **Die ontwerp is ongewoon eerlik.** Gerandomiseer, werklike take, werklike repositories, werklike tydmeting — ’n groot tree verby die selfverslae en benchmark-ranglyste waarop die meeste aansprake oor KI-kodering rus. Die vertraging van 19% het die outeurs se robuustheidstoetse oorleef.
- **Die mees oordraagbare bevinding is die persepsiegaping.** Kundiges se intuïsie oor hulle eie KI-versnelling was met ruweg 40 punte verkeerd, in die optimistiese rigting. Dit is ’n waarskuwing oor *elke* selfgerapporteerde produktiwiteitswins uit KI — hierdie studie se eie syfers ingesluit.
- **Dit is ’n momentopname, nie ’n tendens nie.** METR se eie opvolg van Februarie 2026, met dieselfde soort ontwikkelaars op nuwer gereedskap, wys in die rigting van ’n versnelling — baie ruweg −18% vir terugkerende ontwikkelaars en −4% vir nuwes — maar die outeurs merk dit as swak getuienis, verwring deur wie bereid was om deel te neem (ontwikkelaars het toenemend geweier om sonder KI te werk, die betaling het gedaal, en die taakkeuse het verskuif). Die eerlike lesing is dat die prentjie aan die beweeg is, en selfs die beweging word gerapporteer sonder ’n duim op die skaal.

Waarom dit saak maak

Die grootste deel van die debat oor KI en programmering loop op demonstrasies en gevoelens: ’n gladde skermopname, ’n selfversekerde aanspraak, ’n teenoorgestelde oogrol. Wat skaars is, en wat dit die lees werd maak, is dat iemand die vervelige eksperiment gedoen het — randomiseer, meet werklike werk se tyd, vra dan die mense hoe dit gegaan het. Die antwoord is ongemaklik vir albei kampe. Dit prik die storie dat KI ervare ontwikkelaars op moeilike, bekende kode eenvormig vinniger maak. En dit prik ook die netjiese teenoorgestelde — “KI maak ontwikkelaars stadiger, bewys” — want dieselfde span se nuwer data leun reeds na die ander kant. Die duursaamste les is ook die kleinste en die menslikste: die mense wat die werk gedoen het, het vinniger gevoel terwyl hulle meetbaar stadiger was. “Dit voel vinniger” is nie bewys dat dit so is nie. Meet dit.

Skoon opsomming

In ’n gerandomiseerde gekontroleerde proef het METR sestien ervare oopbron-ontwikkelaars 246 werklike take laat voltooi op kodebasisse wat hulle goed geken het, met KI-gereedskap van vroeg 2025 (Cursor Pro plus Claude 3.5/3.7 Sonnet) toegelaat op ’n lukrake helfte. Die ontwikkelaars het verwag KI sou hulle taaktyd met sowat 24% verkort; in plaas daarvan het dit die voltooiingstyd met 19% verhoog — en agterna het hulle steeds geglo dit het hulle met sowat 20% versnel. Daardie persepsiegaping is die skerp, robuuste kern van die studie. Maar dit is sestien ontwikkelaars, een nou opset en ’n vaste momentopname van vroeg 2025; die outeurs is uitdruklik daaroor dat die resultaat nie na die meeste ontwikkelaars veralgemeen kan word nie — of KI hulle help, bly oop — en hulle eie opvolg van 2026 wys reeds in die rigting van ’n versnelling, met sy eie voorbehoude. ’n Sorgvuldige meting wat ’n mens ernstig moet opneem — nie ’n uitspraak oor KI-kodering nie.

Sources

J. Becker, N. Rush, B. Barnes, D. Rein (METR), Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity, arXiv:2507.09089 [cs.AI] (2025)
https://arxiv.org/abs/2507.09089

METR, ‘Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity’ (study write-up, July 2025)
https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/

METR, developer-uplift follow-up (February 2026)
https://metr.org/blog/2026-02-24-uplift-update/