The Clean Paper

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

Zhvilluesit me përvojë ndiheshin më të shpejtë me AI, ndërsa matjet tregonin se punonin më ngadalë — rezultati i vërtetë dhe verdikti për programimin me AI që ky studim nuk jep

Në një provë të kontrolluar të randomizuar nga METR, gjashtëmbëdhjetë zhvillues me përvojë në projekte me burim të hapur kryen 246 detyra reale në baza kodi që i njihnin mirë; në një gjysmë të zgjedhur rastësisht u lejuan mjete AI të fillimit të vitit 2025 (Cursor Pro dhe Claude 3.5/3.7 Sonnet). Ata prisnin që AI ta shkurtonte kohën e detyrës me rreth 24%; në vend të kësaj koha e përfundimit u rrit me 19% — dhe më pas ata ende besonin se AI i kishte përshpejtuar me rreth 20%. Ky hendek perceptimi është bërthama më e fortë e studimit. Por bëhet fjalë për gjashtëmbëdhjetë zhvillues, një mjedis të ngushtë dhe një fotografi të fiksuar të fillimit të vitit 2025: autorët theksojnë se rezultati nuk tregon se AI nuk ndihmon shumicën e zhvilluesve, dhe ndjekja e tyre e vitit 2026 tashmë anon nga një përshpejtim, me kufizimet e veta.

Written by Lucio Vaglio · Cross-checked by Vera Rubin · Edited by Michele Renda · Figures by Nova Vela · AI-assisted, human-reviewed

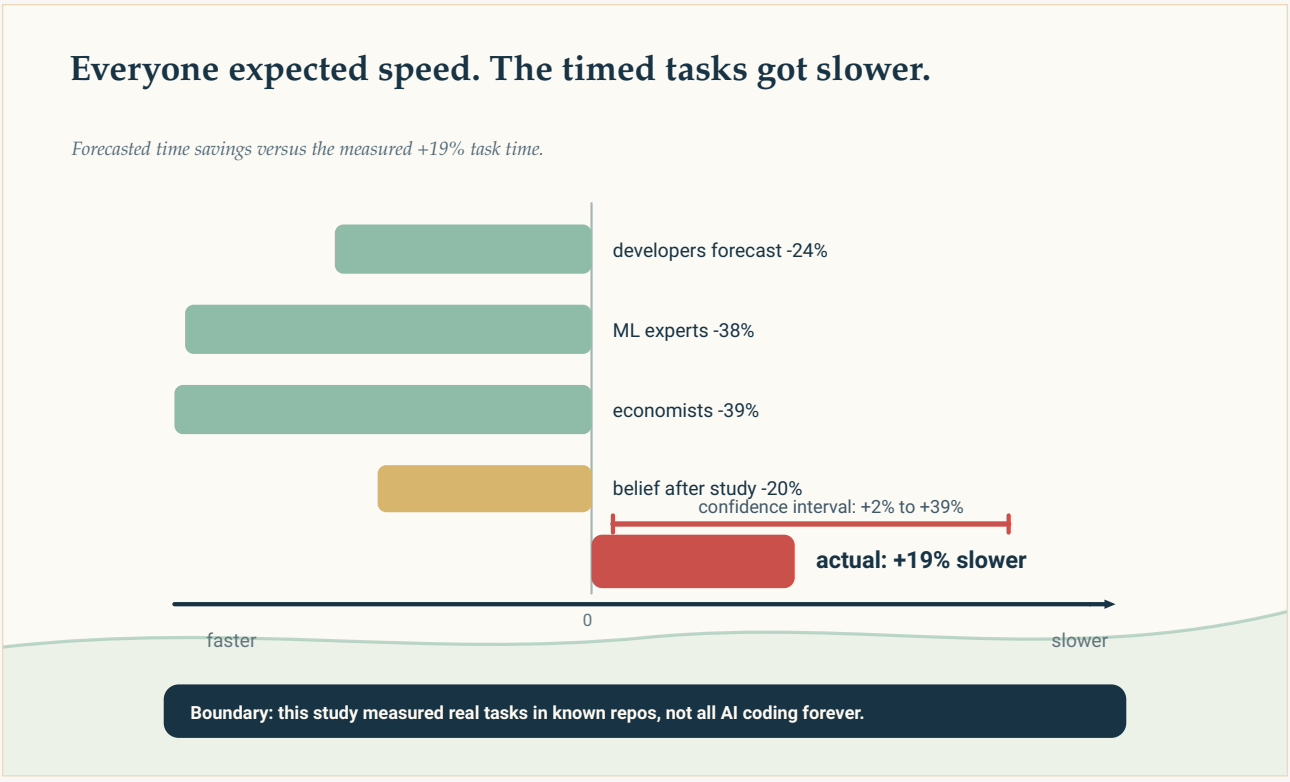
Paper: *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*, Joel Becker, Nate Rush, Beth Barnes, David Rein (METR), arXiv:2507.09089 [cs.AI] (preprint) · DOI: 10.48550/arXiv.2507.09089

Zhvilluesit me përvojë ndiheshin më të shpejtë me AI, ndërsa matjet tregonin se punonin më ngadalë — ky është rezultati i vërtetë, jo një verdikt për programimin me AI

Pyet një programues me përvojë nëse një asistent AI për kodim e përshpejton dhe zakonisht do të marrësh një shifër: më kursen njëzet, tridhjetë për qind. Pyet një ekonomist ose një studiues të të mësuarit makinerik dhe shifra rritet. Në fillim të vitit 2025, një ekip në METR bëri gjënë e ngadaltë dhe të kushtueshme: e mati. Ata morën gjashtëmbëdhjetë zhvillues me përvojë në projekte me burim të hapur, u dhanë 246 detyra reale nga baza të mëdha kodi që i njihnin mirë dhe — me caktim të rastësishëm, detyrë pas detyre — ose i lejuan të përdornin mjete AI, ose jo. Pastaj matën kohën e punës.

Zhvilluesit kishin parashikuar se AI do ta shkurtonte kohën e detyrës me rreth 24%. Ndodhi e kundërta: detyrat e kryera me AI zgjatën **19% më shumë**. Dhe ja pjesa që ia vlen të mbahet mend — pasi përfunduan, të njëjtët zhvillues ende besonin se AI i kishte përshpejtuar me rreth 20%. Ishin më të ngadaltë, por ndiheshin më të shpejtë; distanca mes këtyre dy shifrave është pjesa më interesante e studimit.

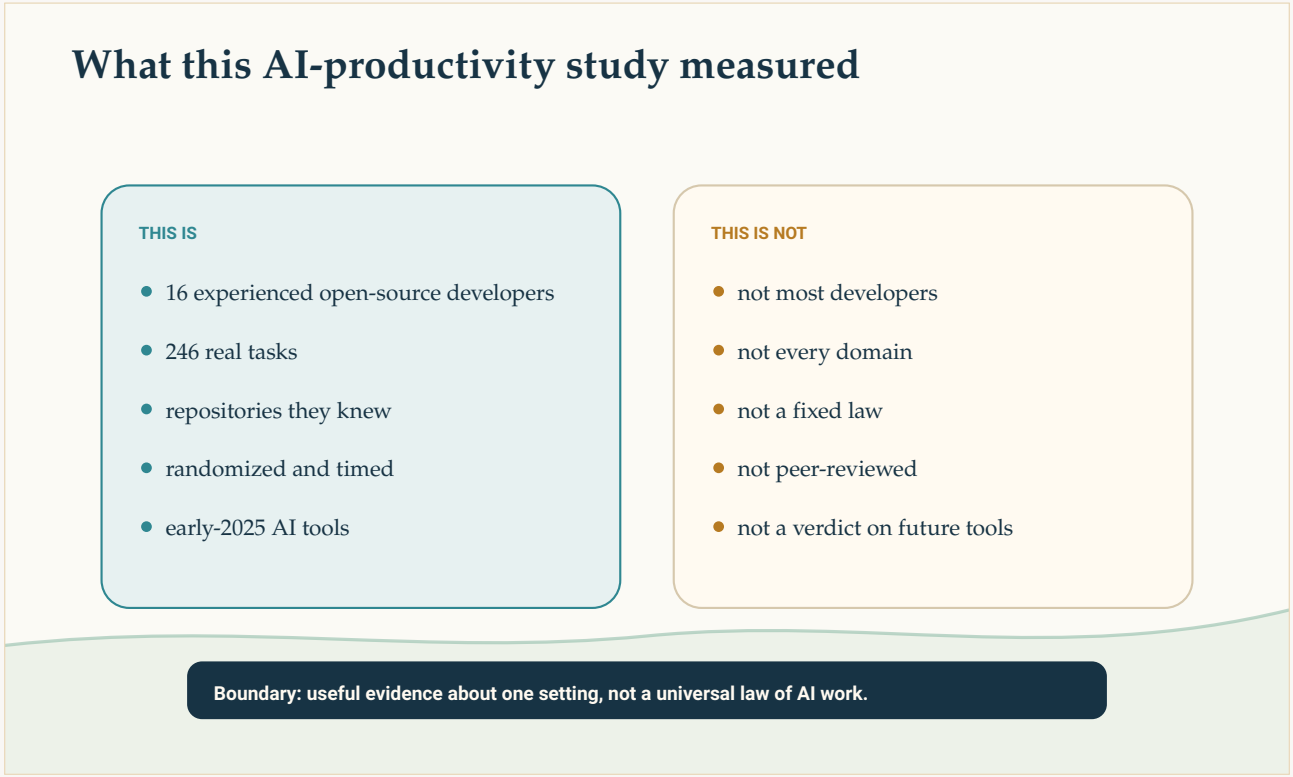
Ky është një rezultat real, i matur me kujdes. Por janë gjithashtu vetëm gjashtëmbëdhjetë zhvillues, duke punuar në repo që i njohin në hollësi, me mjetet e fillimit të vitit 2025 — dhe nuk është e njëjta gjë me fjalinë e prerë “AI i ngadalëson zhvilluesit”. Të dhënat më të vona të të njëjtit ekip tashmë tregojnë në drejtimin tjetër.



Të gjithë parashikuan se AI do ta përshpejtonte punën — zhvilluesit −24%, ekspertët e ML −38%, ekonomistët

−39%, ndërsa bindja e vetë zhvilluesve pas studimit ishte −20%. Matja reale e kohës tregoi +19% më ngadalë, me interval nga +2% në +39%. Hendeku mes asaj që u ndie dhe asaj që u mat është vetë rezultati.

Original diagram — The Clean Paper CC BY 4.0



Çfarë mat studimi — 16 zhvillues ekspertë të burimit të hapur, 246 detyra reale, repo të njohura, mjete të fillimit të vitit 2025, caktim të rastësishëm — dhe çfarë nuk mat: jo shumicën e zhvilluesve, jo fusha të tjera, jo një ligj të pandryshueshëm dhe jo një studim të rishikuar nga kolegët.

Original diagram — The Clean Paper CC BY 4.0

Çfarë mati ky studim dhe çfarë të jep një provë e randomizuar

Një **provë e kontrolluar e randomizuar (RCT)** është mjeti që mjekësia përdor për të dalluar një efekt real nga një efekt që thjesht shpresojmë ta shohim. Këtu secila nga 246 detyrat u caktua rastësisht që të kryhej me AI të lejuar ose pa AI, kështu që mes dy grupeve, mesatarisht, i vetmi ndryshim sistematik është vetë AI-ja. Kjo të lejon të thuash se AI *shkaktoi* ndryshimin në kohë, në vend që vetëm të vëresh se njerëzit që zgjedhin të përdorin AI rastësisht janë më të shpejtë ose më të ngadaltë për arsye të tjera. Kjo ka rëndësi sepse provat e zakonshme për rritjen e produktivitetit nga AI — vetëraportimet dhe rezultatet në benchmarke — nuk mund ta bëjnë këtë: një benchmark nuk është punë reale dhe një vetëvlerësim, siç tregon ky studim, mund të jetë me bindje i gabuar. Zhvilluesit këtu nuk ishin fillestarë që po provonin një lodër të re. Ata ishin kontribues të konsoliduar në projekte të mëdha e të pjekura me burim të hapur, të cilat i njihnin mirë dhe kishin edhe përvojë paraprake me mjetet.

Çfarë bënë autorët

- Zhvilluan një **provë të kontrolluar të randomizuar** (METR: Joel Becker, Nate Rush, Beth Barnes, David Rein). Gjashtëmbëdhjetë zhvillues me përvojë në burim të hapur, secili duke punuar në një depo të madhe kodi ku kontribuonte rregullisht dhe që e njihte mirë — mesatarisht pesë vjet përvojë në projektet përkatëse.
- Përdorën **246 detyra reale** — korrigjime gabimesh, funksione të reja dhe ristrukturime kodi të marra nga sistemet e vetë projekteve për ndjekjen e çështjeve. Çdo detyrë u caktua rastësisht në "AI e lejuar" ose "AI e ndaluar".
- "AI e lejuar" nënkuptonte **mjetet e fillimit të vitit 2025: Cursor Pro me Claude 3.5/3.7 Sonnet.** Matja kryesore ishte koha reale e përfundimit për secilën detyrë. Krahas saj, ekipi mblodhi parashikime nga zhvilluesit para punës, vlerësime të tyre pas punës dhe parashikime nga ekspertë të ekonomisë dhe të të mësuarit makinerik.

Çfarë gjetën

- **Me AI, detyrat zgjatën 19% më shumë.** Jo më pak — më shumë. Intervali 95% i besimit shkon afërsisht nga +2% në +39%, kështu që drejtimi i efektit është i qëndrueshëm edhe pse madhësia e saktë nuk është.
- **Të gjithë kishin parashikuar të kundërtën.** Zhvilluesit prisnin një përshpejtim 24%; ekspertët e të mësuarit makinerik rreth 38%; ekonomistët rreth 39%. Të tre grupet prisnin që AI të kursente shumë kohë; ora tregoi se kushtoi kohë.
- **Hendeku i perceptimit.** Pasi e kishin kryer punën më ngadalë, zhvilluesit ende vlerësuan se AI i kishte përshpejtuar me rreth 20% — një diferencë prej afërsisht 40 pikësh përqindjeje mes asaj që ndien dhe asaj që regjistroi koha.
- **Shkaqe të mundshme, të peshuara por jo të provuara.** Autorët rendisin disa faktorë që mund ta shpjegojnë ngadalësimin: këta zhvillues i njohin shumë mirë bazat e tyre të kodit, ndaj asistenti ka më pak vlerë shtesë; projektet e pjekura kanë standarde të larta cilësie, shpesh të pashkruara; repo-t janë të mëdha dhe përmbajnë shumë kontekst që modeli nuk e ka; dhe kohë reale harxhohet duke formuluar prompt-e, pastaj duke kontrolluar dhe korrigjuar daljen e AI-së. Autorët i paraqesin këto si pista, jo si përfundime.

Çfarë nuk tregon ky studim

- **Nuk** tregon se AI nuk e përshpejton shumicën e zhvilluesve. Autorët e thonë hapur: gjashtëmbëdhjetë ekspertë që punojnë me kod që e njohin pothuajse përmendësh nuk janë zhvilluesi mesatar në detyrën mesatare.
- **Nuk** tregon se AI është e padobishme ose se i ngadalëson njerëzit në rrethana të tjera — p.sh. zhvillues të rinj në një bazë kodi, projekte të nisura nga zero, gjuhë të panjohura ose fusha krejt të tjera.

- **Nuk** i ngrin mjetet në kohë. Këtu bëhet fjalë për Cursor dhe Claude 3.5/3.7 Sonnet të fillimit të vitit 2025; autorët theksojnë se mjete më të mira ose mënyra më të mira përdorimi mund ta ndryshojnë rezultatin edhe në të njëjtin mjedis.
- Është një **preprint** (publikuar në korrik 2025, ende pa rishikim nga kolegët) dhe autorët thonë se nuk mund t'i përjashtojnë plotësisht artefaktet eksperimentale — ndonëse rezultati mbeti i qëndrueshëm në analizat e tyre.
- **Nuk** justifikon leximin ngushëllues se zhvilluesit duhet të kenë fituar në ndonjë mënyrë tjetër — kanë mësuar më shumë, janë ndier më mirë ose kanë shkruar kod më të mirë. Pikërisht gjëja e matur këtu, ndjesia e përshpejtimit, është ajo që të dhënat e kundërshtojnë.

Sa e fortë është prova

- **Dizajni është jashtëzakonisht i drejtpërdrejtë.** Randomizim, detyra reale, repo reale, matje reale e kohës — një hap i madh përpara krahasuar me vetëraportimet dhe tabelat e benchmarkeve mbi të cilat mbështeten shumë pretendime për kodimin me AI. Ngadalësimi 19% i mbijetoi kontrolleve të robustësisë së autorëve.
- **Rezultati më i përgjithshëm është hendeku i perceptimit.** Intuita e ekspertëve për përshpejtimin e tyre nga AI ishte gabim me afërsisht 40 pikë, në drejtim optimist. Kjo është një arsye për t'u treguar të kujdesshëm me *çdo* rritje produktiviteti nga AI e raportuar nga vetë përdoruesit — përfshirë edhe shifrat e këtij studimi.
- **Është një fotografi e një momenti, jo një prirje.** Ndjekja e vetë METR në shkurt 2026, me të njëjtin lloj zhvilluesish dhe mjete më të reja, anon nga një përshpejtim — shumë afërsisht −18% për zhvilluesit që u rikthyen dhe −4% për pjesëmarrësit e rinj — por autorët e quajnë provë të dobët, të shtrembëruar nga përzgjedhja e pjesëmarrësve (zhvilluesit refuzonin gjithnjë e më shpesh të punonin pa AI, pagesa ra dhe përzgjedhja e detyrave ndryshoi). Leximi i ndershëm është se panorama po lëviz, dhe edhe kjo lëvizje raportohet me kujdes.

Pse ka rëndësi

Pjesa më e madhe e debatit për AI dhe programimin mbahet mbi demonstrime dhe përshtypje: një video e lëmuar, një pretendim i sigurt, një ngritje skeptike vetullash. Ajo që është e rrallë — dhe që e bën këtë studim të vlefshëm — është se dikush bëri eksperimentin e mërzitshëm: randomizo, mat kohën e punës reale dhe pastaj pyet njerëzit si shkoi. Përgjigjja është e pakëndshme për të dy kampet. Ajo çan idenë se AI i turbo-ngarkon në mënyrë uniforme zhvilluesit ekspertë në kod të vështirë e të njohur. Por çan edhe të kundërtën e rregullt — "AI i ngadalëson zhvilluesit, u provua" — sepse të dhënat më të reja të të njëjtit ekip tashmë anojnë në drejtimin tjetër. Mësimi më i qëndrueshëm është edhe më i vogli dhe më njerëzori: njerëzit që po bënin punën ndiheshin më të shpejtë ndërsa matjet tregonin se ishin më të ngadaltë. "Më duket më shpejt" nuk është provë se është më shpejt. Mate.

Përmbledhje e shkurtër

Në një provë të kontrolluar të randomizuar, METR u kërkoi gjashtëmbëdhjetë zhvilluesve me përvojë në burim të hapur të kryenin 246 detyra reale në baza kodi që i njihnin mirë, duke lejuar mjete AI të fillimit të vitit 2025 (Cursor Pro plus Claude 3.5/3.7 Sonnet) në gjysmën e detyrave të zgjedhura rastësisht. Zhvilluesit prisnin që AI ta shkurtonte kohën e detyrës me rreth 24%; në vend të kësaj, koha e përfundimit u rrit me 19% — dhe pas punës ata ende besonin se AI i kishte përshpejtuar me rreth 20%. Ky hendek perceptimi është bërthama më e qartë dhe më e fortë e studimit. Por janë gjashtëmbëdhjetë zhvillues, një mjedis i ngushtë dhe një fotografi fikse e fillimit të vitit 2025; autorët theksojnë se kjo nuk tregon se AI nuk ndihmon shumicën e zhvilluesve dhe ndjekja e tyre e vitit 2026 tashmë anon drejt një përshpejtimi, me kufizimet e veta. Një matje e kujdesshme që meriton të merret seriozisht — jo një verdikt për programimin me AI.

Burimet

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/`