Дасведчаным распрацоўшчыкам здавалася, што з ШІ яны працуюць хутчэй, хоць вымярэнні паказалі адваротнае — вось сапраўдны вынік, а не прысуд праграмаванню з ШІ

У рандомізаваным кантраляваным даследаванні METR шаснаццаць дасведчаных распрацоўшчыкаў адкрытага ПЗ выканалі 246 рэальных задач у добра знаёмых ім кодавых базах; для выпадковай паловы задач дазваляліся інструменты ШІ пачатку 2025 года (Cursor Pro разам з Claude 3.5/3.7 Sonnet). Яны чакалі, што ШІ скароціць час выканання прыкладна на 24%; замест гэтага ён павялічыў яго на 19%, а пасля працы распрацоўшчыкі ўсё роўна лічылі, што ШІ паскорыў іх прыкладна на 20%. Менавіта гэты разрыў ва ўспрыманні — самае яснае і ўстойлівае ядро даследавання. Але гэта шаснаццаць распрацоўшчыкаў, адно вузкае асяроддзе і фіксаваны здымак пачатку 2025 года: аўтары наўпрост кажуць, што вынік не паказвае, нібы ШІ не дапамагае большасці распрацоўшчыкаў, а іх уласны працяг 2026 года ўжо схіляецца ў бок паскарэння — са сваімі агаворкамі.

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

Дасведчаным распрацоўшчыкам здавалася, што з ШІ яны працуюць хутчэй, хоць вымярэнні паказалі адваротнае — вось сапраўдны вынік, а не прысуд праграмаванню з ШІ

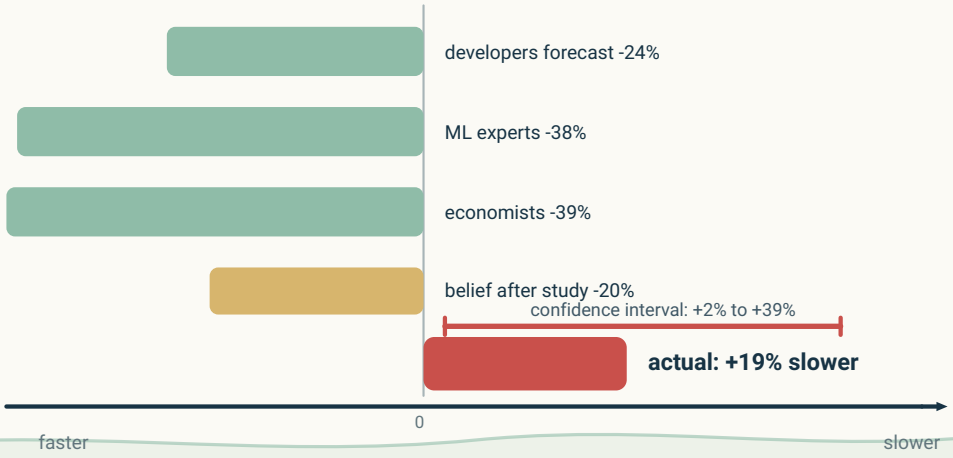
Спытайце ў дасведчанага праграміста, ці паскарае яго працу ШІ-памочнік для напісання кода, — і звычайна пачуеце канкрэтную лічбу: эканоміць мне дваццаць, трыццаць працэнтаў часу. Спытайце ў эканаміста або даследчыка машыннага навучання — і лічба стане яшчэ большай. У пачатку 2025 года каманда METR зрабіла павольную і дарагую рэч: праверыла гэта. Даследчыкі набралі шаснаццаць дасведчаных распрацоўшчыкаў адкрытага ПЗ, далі ім 246 рэальных задач з буйных кодавых баз, якія тыя добра ведалі, і — выпадковым чынам для кожнай задачы — або дазвалялі выкарыстоўваць інструменты ШІ, або не. А потым засікалі час.

Распрацоўшчыкі прагназавалі, што ШІ скароціць час выканання задач прыкладна на 24%. Адбылося адваротнае: задачы, якія выконваліся з ШІ, патрабавалі **на 19% больш часу**. І вось на чым варта затрымацца: нават закончыўшы працу, тыя ж распрацоўшчыкі па-ранейшаму лічылі, што ШІ паскорыў іх прыкладна на 20%. Яны працавалі павольней, але адчувалі, што хутчэй, — і менавіта разрыў паміж гэтымі дзвюма лічбамі з’яўляецца самай цікавай часткай даследавання.

Гэта рэальны і старанна вымераны вынік. Але гэта таксама ўсяго шаснаццаць распрацоўшчыкаў, якія працавалі з рэпазіторыямі, вядомымі ім да дробязяў, і карысталіся інструментамі пачатку 2025 года. Таму выснову нельга звесці да простага «ШІ запавольвае распрацоўшчыкаў». Пазнейшыя даныя той жа каманды ўжо паказваюць у супрацьлеглы бок.

Everyone expected speed. The timed tasks got slower.

Forecasted time savings versus the measured +19% task time.



Boundary: this study measured real tasks in known repos, not all AI coding forever.

Усе чакалі, што ШІ паскорыць працу: распрацоўшчыкі — на 24%, спецыялісты па машынным навучанні — на 38%, эканамісты — на 39%, а самі распрацоўшчыкі пасля эксперыменту лічылі, што сталі хутчэйшымі на 20%. Секундамер паказаў адваротнае: +19% да часу выканання, з інтэрвалам ад +2% да +39%. Разрыў паміж адчуваннем і вымярэннем — і ёсць ключавы вынік.

What this AI-productivity study measured

THIS IS

- 16 experienced open-source developers
- 246 real tasks
- repositories they knew
- randomized and timed
- early-2025 AI tools

THIS IS NOT

- not most developers
- not every domain
- not a fixed law
- not peer-reviewed
- not a verdict on future tools

Boundary: useful evidence about one setting, not a universal law of AI work.

Што менавіта вымярала даследаванне: 16 дасведчаных распрацоўшчыкаў адкрытага ПЗ, 246 рэальных задач, знаёмыя рэпазіторыі, інструменты пачатку 2025 года і рандомізацыя. І чаго яно не вымярала: не большасць распрацоўшчыкаў, не іншыя галіны, не нязменны закон; даследаванне таксама яшчэ не прайшло рэцэнзаванне.

Што менавіта тут вымяралі і навошта патрэбнае рандомізаванае даследаванне

Рандомізаванае кантраляванае даследаванне (RCT) — інструмент, якім, напрыклад, медыцына аддзяляе рэальны эфект ад жаданага. Тут кожную з 246 задач выпадкова адносілі або да ўмовы, дзе ШІ дазволены, або да ўмовы, дзе ён забаронены. Таму ў сярэднім адзінае сістэматычнае адрозненне паміж двума наборамі задач — менавіта наяўнасць ШІ. Гэта дазваляе сказаць, што ШІ *выклікаў* змяненне часу, а не проста заўважыць, што людзі, якія цягнуцца да ШІ, па нейкіх іншых прычынах і без таго працуюць хутчэй або павольней. Гэта важна, бо звычайныя доказы росту прадукцыйнасці ад ШІ ў праграмаванні — самаацэнкі і вынікі бенчмаркаў — такой высновы не дазваляюць: бенчмарк не з'яўляецца рэальнай працай, а самаацэнка, як паказвае гэта даследаванне, можа быць упэўнена памылковай. Распрацоўшчыкі тут не былі навічкамі, якія няўклюдна асвойваюць новую цацку. Гэта былі дасведчаныя ўдзельнікі буйных сталых праектаў адкрытага ПЗ, якія добра ведалі свае кодавыя базы і ўжо мелі пэўны досвед працы з такімі інструментамі.

Што зрабілі аўтары

- Правялі **рандомізаванае кантраляванае даследаванне** (METR: Joel Becker, Nate Rush, Beth Barnes, David Rein). Шаснаццаць дасведчаных распрацоўшчыкаў адкрытага ПЗ кожны працаваў з буйным рэпазіторыем, у які рэгулярна ўносіў уклад і які добра ведаў; у сярэднім яны працавалі над канкрэтнымі праектамі пяць гадоў.
- Выкарысталі **246 рэальных задач** — выпраўленні памылак, новыя функцыі і рэфактарынг з уласных трэкераў задач гэтых праектаў. Кожная задача выпадкова трапляла ва ўмову «ШІ дазволены» або «ШІ забаронены».
- «ШІ дазволены» азначала **інструменты пачатку 2025 года: Cursor Pro з Claude 3.5/3.7 Sonnet.** Асноўным паказчыкам быў рэальны час выканання кожнай задачы. Акрамя яго каманда сабрала прагнозы распрацоўшчыкаў да працы і іх ацэнкі пасля яе, а таксама прагнозы экспертаў па эканоміцы і машынным навучанні.

Што яны выявілі

- **З ШІ задачы выконваліся на 19% даўжэй.** Не хутчэй, а павольней. 95%-ы даверны інтэрвал цягнецца прыкладна ад +2% да +39%, таму кірунак эфекту даволі надзейны, нават калі яго дакладная велічыня менш пэўная.
- **Усе прадказалі адваротнае.** Распрацоўшчыкі прагназавалі паскарэнне на 24%, спецыялісты па машынным навучанні — прыкладна на 38%, эканамісты — прыкладна на 39%. Усе тры групы чакалі істотнай эканоміі часу; секундамер паказаў пэўную яго страту.
- **Разрыў ва ўспрыманні.** Пасля таго як распрацоўшчыкі выканалі працу павольней, яны ўсё роўна ацэньвалі, што ШІ паскорыў іх прыкладна на 20%. Гэта разрыў прыкладна ў 40 працэнтных пунктаў паміж адчуваннем і тым, што зафіксаваў гадзіннік.
- **Магчымыя прычыны ўзважаныя, але не даказаныя.** Аўтары пералічваюць фактары, якія маглі выклікаць запаволенне: гэтыя распрацоўшчыкі вельмі добра ведаюць уласныя кодавыя базы, таму памочніку застаецца менш што дадаць; сталыя праекты маюць высокія і часта неяўныя стандарты якасці; рэпазіторыі вялікія і ўтрымліваюць шмат кантэксту, якога мадэль не бачыць; рэальны час ідзе на фармуляванне запытаў, а потым на праверку і выпраўленне вынікаў ШІ. Аўтары падаюць гэта як магчымыя тлумачэнні, а не як устаноўленыя прычыны.

Чаго гэта не паказвае

- Гэта **не** паказвае, што ШІ не здольны паскараць большасць распрацоўшчыкаў. Аўтары кажуць пра гэта наўпрост: шаснаццаць экспертаў, якія ведаюць свой код амаль напамяць, — не тое самае, што сярэдні распрацоўшчык з сярэдняй задачай.
- Гэта **не** паказвае, што ШІ бескарысны або што ён запавольвае людзей у іншых умовах — напрыклад, навічкоў у кодавой базе, працу над новым праектам, незнаёмыя мовы праграмавання або зусім іншыя галіны.
- Яно **не** замарожвае інструменты назаўжды ў адным стане. Тут гаворка ідзе пра Cursor і Claude 3.5/3.7 Sonnet пачатку 2025 года; аўтары наўпрост адзначаюць, што лепшыя інструменты або лепшыя спосабы выкарыстання тых жа інструментаў могуць змяніць вынік нават у дакладна такім жа асяроддзі.
- Гэта **прэпрынт** (апублікаваны ў ліпені 2025 года, яшчэ не рэцэнзаваны), і аўтары прызнаюць, што не могуць цалкам выключыць эксперыментальныя артэфакты, хоць вынік захоўваўся ў розных варыянтах іх аналізу.
- Яно **не** дае падстаў для суцяшальнага тлумачэння, нібы распрацоўшчыкі абавязкова выйгралі ў чымсьці іншым — больш навучыліся, былі больш задаволеныя або напісалі больш якасны код. Адзіная рэч такога кшталту, якую тут вымяралі, — адчуванае паскарэнне — якраз супярэчыць даным.

Наколькі моцныя доказы

- **Дызайн незвычайна сумленны.** Рандомізацыя, рэальныя задачы, рэальныя рэпазіторыі і фактычнае вымярэнне часу — вялікі крок наперад у параўнанні з самаацэнкамі і табліцамі лідараў бенчмаркаў, на якіх грунтуецца большасць заяў пра ШІ-праграмаванне. Запаволенне на 19% захоўвалася ў праверках устойлівасці выніку, праведзеных аўтарамі.
- **Найбольш пераносны вынік — разрыў ва ўспрыманні.** Экспертная інтуіцыя наконт уласнага паскарэння дзякуючы ШІ памылілася прыкладна на 40 працэнтных пунктаў — у аптымістычны бок. Гэта папярэджанне адносна *любых* самаацэнак росту прадукцыйнасці ад ШІ, уключаючы лічбы з гэтага ж даследавання.
- **Гэта здымак моманту, а не трэнд.** Уласны працяг METR у лютым 2026 года, з падобнымі распрацоўшчыкамі, але навейшымі інструментамі, ужо паказвае на паскарэнне — вельмі прыблізна −18% часу для распрацоўшчыкаў, якія вярнуліся ў даследаванне, і −4% для новых удзельнікаў. Аднак аўтары называюць гэтыя даныя слабымі: на іх паўплывала тое, хто згаджаўся ўдзельнічаць (распрацоўшчыкі ўсё часцей адмаўляліся працаваць без ШІ, аплата знізілася, а выбар задач зрушыўся). Асцярожнае прачытанне такое: карціна змяняецца, і нават гэты рух аўтары апісваюць без спробы нахіліць шалі ў патрэбны бок.

Чаму гэта важна

Большасць спрэчак пра ШІ і праграмаванне жыве на дэманстрацыях і адчуваннях: эфектны запіс экрана, упэўненая заява, у адказ — скептычна закочаныя вочы. Рэдкасць тут у тым, што нехта правёў сумны эксперымент: выпадкова размеркаваў умовы, вымераў час рэальнай працы, а потым спытаў людзей, як, на іх думку, усё прайшло. Адказ нязручны для абодвух лагераў. Ён прабівае дзірку ў гісторыі пра тое, што ШІ аднолькава моцна турбапаскарае дасведчаных распрацоўшчыкаў на складаным і добра знаёмым кодзе. І гэтак жа прабівае дзірку ў акуратнай супрацьлегласці — «даказана, што ШІ запавольвае распрацоўшчыкаў», — бо навейшыя даныя той жа каманды ўжо схіляюцца ў іншы бок. Найбольш устойлівы ўрок адначасова самы малы і самы чалавечы: людзі, якія выконвалі працу, адчувалі сябе хутчэйшымі, калі аб’ектыўна працавалі павольней. «Так здаецца хутчэй» — не доказ таго, што гэта сапраўды хутчэй. Трэба вымяраць.

Кароткі вынік

У рандомізаваным кантраляваным даследаванні METR шаснаццаць дасведчаных распрацоўшчыкаў адкрытага ПЗ выканалі 246 рэальных задач у добра знаёмых ім кодавых базах; для выпадковай паловы задач былі дазволены інструменты ШІ пачатку

2025 года — Cursor Pro разам з Claude 3.5/3.7 Sonnet. Распрацоўшчыкі чакалі, што ШІ скароціць час выканання прыкладна на 24%; замест гэтага ён павялічыў яго на 19%, а пасля працы яны ўсё роўна лічылі, што ШІ паскорыў іх прыкладна на 20%. Менавіта гэты разрыў ва ўспрыманні — самае яснае і ўстойлівае ядро даследавання. Але гэта шаснаццаць распрацоўшчыкаў, адно вузкае асяроддзе і фіксаваны здымак пачатку 2025 года; аўтары наўпрост кажуць, што вынік не паказвае, нібы ШІ не дапамагае большасці распрацоўшчыкаў, а іх уласны працяг 2026 года ўжо паказвае на паскарэнне — са сваімі агаворкамі. Стараннае вымярэнне, якое варта ўспрымаць сур’ёзна, але не прысуд праграмаванню з ШІ.

Крыніцы

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/