



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

Pembangun berpengalaman berasa lebih pantas dengan AI walaupun secara terukur bekerja lebih perlahan — penemuan sebenar, dan mengapa ia bukan keputusan muktamad tentang pengekodan AI

Dalam satu percubaan terkawal rawak oleh METR, enam belas pembangun sumber terbuka berpengalaman menyelesaikan 246 tugas sebenar pada pangkalan kod yang mereka kenali dengan baik, dengan alat AI awal 2025 (Cursor Pro bersama Claude 3.5/3.7 Sonnet) dibenarkan secara rawak pada separuh tugas. Mereka menjangka AI akan mengurangkan masa tugas kira-kira 24%; sebaliknya, masa siap meningkat 19% — dan selepas itu mereka masih percaya AI telah mempercepatkan mereka kira-kira 20%. Jurang persepsi itu ialah teras kajian yang paling tajam dan kukuh. Namun ia hanya melibatkan enam belas pembangun, satu keadaan yang sempit dan gambaran beku awal 2025: para penulis menegaskan bahawa hasil ini tidak menunjukkan AI gagal membantu kebanyakan pembangun, dan susulan mereka sendiri pada 2026 sudah cenderung ke arah peningkatan kelajuan, dengan batasannya tersendiri.

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

Pembangun berpengalaman berasa lebih pantas dengan AI walaupun secara terukur bekerja lebih perlahan — pene- muan sebenar, dan mengapa ia bukan keputusan muktamad tentang pengekodan AI

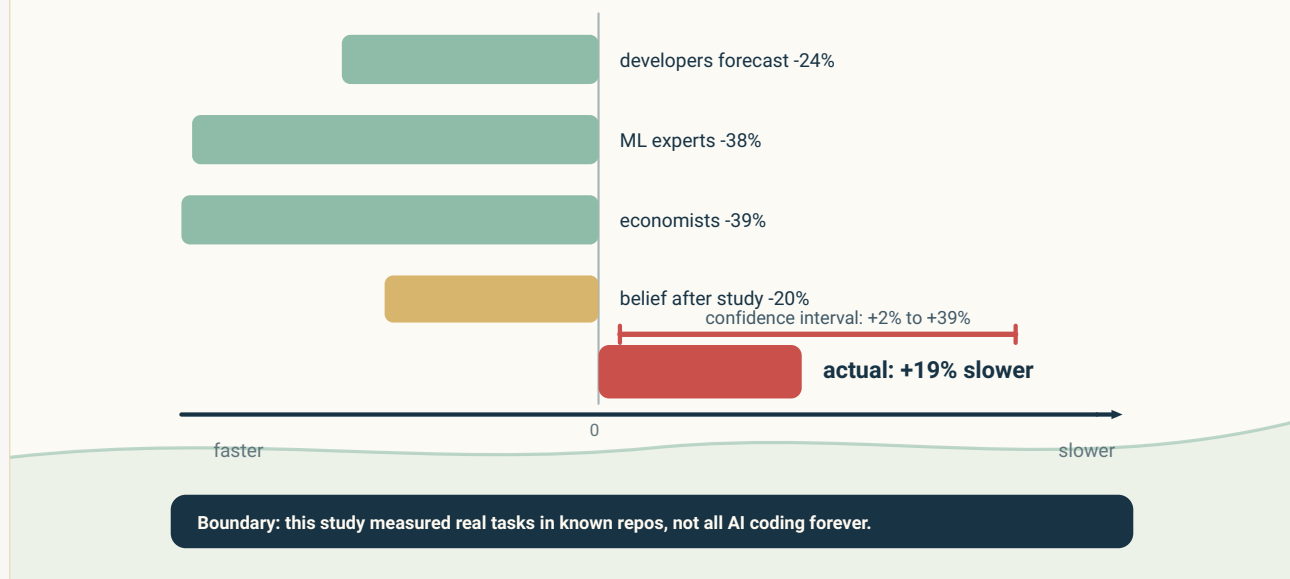
Tanya seorang pengatur cara berpengalaman sama ada pembantu pengekodan AI menjadikan mereka lebih pantas, dan anda biasanya akan mendapat satu angka: ia menjimatkan masa saya dua puluh, tiga puluh peratus. Tanya ahli ekonomi atau penyelidik pembelajaran mesin, dan angkanya menjadi lebih besar. Pada awal 2025, satu pasukan di METR melakukan perkara yang lambat dan mahal: mereka mengujinya. Mereka mengambil enam belas pembangun sumber terbuka berpengalaman, memberikan 246 tugas sebenar daripada pangkalan kod besar yang mereka kenali dengan baik, lalu — seperti melambung syiling bagi setiap tugas — sama ada membenarkan alat AI atau tidak. Kemudian mereka mengukur masa kerja itu.

Para pembangun meramalkan bahawa AI akan mengurangkan masa tugas mereka kira-kira 24%. Yang berlaku ialah sebaliknya: tugas yang dilakukan dengan AI mengambil masa **19% lebih lama**. Dan inilah bahagian yang patut direnungkan — selepas selesai, pembangun yang sama masih percaya AI telah mempercepatkan mereka, kira-kira 20%. Mereka sebenarnya lebih perlahan, tetapi berasa lebih pantas, dan jarak antara dua angka itu ialah perkara paling menarik dalam kajian ini.

Ini ialah hasil sebenar yang diukur dengan teliti. Tetapi ia juga hanya enam belas pembangun, bekerja pada repositori yang mereka kenali secara mendalam, menggunakan alat dari awal 2025 — dan ia bukan ayat mutlak bahawa “AI menjadikan pembangun lebih perlahan.” Data susulan daripada pasukan yang sama sudah menunjuk ke arah sebaliknya.

Everyone expected speed. The timed tasks got slower.

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



Semua orang meramalkan AI akan mempercepatkan kerja — pembangun —24%, pakar ML —38%, ahli ekonomi —39%, dan kepercayaan pembangun sendiri selepas kajian —20%. Jam randik pula mendapati kerja +19% lebih perlahan, dengan selang +2% hingga +39%. Jurang antara apa yang dirasakan dan apa yang diukur itulah penemuannya.

Original diagram — The Clean Paper CC BY 4.0

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.

Apa yang diukur oleh kajian — 16 pembangun sumber terbuka pakar, 246 tugas sebenar, repositori yang dikenali, alat awal 2025, penugasan rawak — dan apa yang tidak diukurnya: bukan kebanyakan pembangun, bukan bidang lain, bukan hukum tetap, dan belum melalui semakan rakan sebaya.

Original diagram — The Clean Paper CC BY 4.0

Apa yang diukur, dan apa kelebihan percubaan rawak

Percubaan terkawal rawak (RCT) ialah alat yang digunakan dalam perubatan untuk membezakan kesan sebenar daripada kesan yang sekadar dijangka. Di sini, setiap satu daripada 246 tugas ditugaskan secara rawak untuk dilakukan sama ada dengan AI dibenarkan atau tidak, supaya secara purata satu-satunya perbezaan sistematik antara dua kumpulan itu ialah AI itu sendiri. Itulah yang membolehkan kita mengatakan AI *menyebabkan* perubahan masa, bukannya sekadar memerhatikan bahawa orang yang memilih menggunakan AI kebetulan lebih pantas atau lebih perlahan kerana sebab lain. Ini penting kerana bukti biasa bagi peningkatan pengekodan dengan AI — laporan sendiri dan skor penanda aras — tidak boleh melakukan perkara ini: penanda aras bukan kerja sebenar, dan laporan sendiri, seperti ditunjukkan kajian ini, boleh salah dengan penuh keyakinan. Pembangun di sini bukan orang baharu yang terkial-kial dengan alat baharu. Mereka ialah penyumbang mapan kepada projek sumber terbuka besar dan matang yang mereka kenali dengan baik, dan mempunyai sedikit pengalaman terdahulu menggunakan alat tersebut.

Apa yang dilakukan oleh para penulis

- Menjalankan **percubaan terkawal rawak** (METR: Joel Becker, Nate Rush, Beth Barnes, David Rein). Enam belas pembangun sumber terbuka berpengalaman, masing-masing bekerja pada repositori besar yang mereka sumbangkan secara tetap dan kenali dengan baik — purata lima tahun pada projek tertentu itu.
- Menggunakan **246 tugas sebenar** — pembaikan pepijat, ciri baharu dan refaktor yang diambil daripada penjejak isu projek tersebut sendiri. Setiap tugas ditugaskan secara rawak kepada “AI dibenarkan” atau “AI tidak dibenarkan.”
- “AI dibenarkan” bermaksud **alat awal 2025: Cursor Pro dengan Claude 3.5/3.7 Sonnet**. Ukuran utama ialah masa sebenar yang diambil untuk menyiapkan setiap tugas. Selain itu, pasukan mengumpulkan ramalan (daripada pembangun sebelum tugas) dan anggaran (daripada mereka selepas tugas), serta ramalan daripada pakar ekonomi dan pembelajaran mesin.

Apa yang mereka temui

- **Dengan AI, tugas mengambil masa 19% lebih lama.** Bukan lebih pantas — lebih perlahan. Selang keyakinan 95% ialah kira-kira +2% hingga +39%, jadi arah kesannya kukuh walaupun saiz tepatnya kurang pasti.

- **Semua orang telah meramalkan sebaliknya.** Pembangun meramalkan peningkatan kelajuan 24%; pakar pembelajaran mesin kira-kira 38%; ahli ekonomi kira-kira 39%. Ketiga-tiga kumpulan menjangka AI menjimatkan banyak masa; jam randik mendapati ia sebaliknya memakan masa.
- **Jurang persepsi.** Selepas melakukan kerja dan ternyata lebih perlahan, para pembangun masih menganggap AI telah mempercepatkan mereka kira-kira 20% — jurang kira-kira 40 mata antara apa yang mereka rasakan dan apa yang direkodkan oleh masa sebenar.
- **Sebab yang mungkin, dinilai tetapi tidak dibuktikan.** Para penulis menyenaraikan faktor yang boleh menjelaskan kelembapan: pembangun ini mengenali pangkalan kod mereka sendiri secara mendalam, jadi kurang ruang untuk pembantu menambah nilai; projek matang mempunyai piawaian kualiti tinggi yang sering tidak dinyatakan secara eksplisit; repositori besar dan penuh konteks yang model tidak miliki; dan masa sebenar digunakan untuk menulis arahan, kemudian menyemak serta membetulkan output AI. Mereka mengemukakannya sebagai petunjuk, bukan keputusan muktamad.

Apa yang tidak ditunjukkan oleh kajian ini

- Ia **tidak** menunjukkan bahawa AI gagal mempercepatkan kebanyakan pembangun. Para penulis menyatakannya secara langsung: enam belas pakar yang bekerja pada kod yang mereka hafal luar dan dalam bukanlah pembangun purata pada tugas purata.
- Ia **tidak** menunjukkan AI tidak berguna, atau bahawa AI memperlahankan orang dalam keadaan lain — orang baharu pada sesuatu pangkalan kod, projek yang dibina dari kosong, bahasa pengaturcaraan yang tidak dikenali, atau bidang lain sama sekali.
- Ia **tidak** membekukan teknologi pada satu tahap. Ini ialah Cursor dan Claude 3.5/3.7 Sonnet pada awal 2025; para penulis secara jelas mengatakan alat yang lebih baik, atau cara yang lebih baik untuk menggunakan alat yang sama, boleh mengubah hasil walaupun dalam keadaan yang tepat ini.
- Ia ialah **pracetak** (disiarkan Julai 2025, belum melalui semakan rakan sebaya), dan para penulis mengakui mereka tidak dapat menolak sepenuhnya artifak eksperimen — walaupun hasil itu kekal dalam pelbagai analisis mereka.
- Ia **tidak** membenarkan tafsiran yang menenangkan bahawa pembangun pasti mendapat faedah lain — belajar lebih banyak, berasa lebih gembira, menulis kod lebih baik. Satu perkara yang benar-benar diukur di sini, iaitu rasa bahawa mereka lebih pantas, ialah perkara yang secara langsung disangkal oleh data.

Sejauh mana kukuh buktinya

- **Reka bentuknya luar biasa jujur.** Rawak, tugas sebenar, repositori sebenar, masa sebenar — satu peningkatan besar berbanding laporan sendiri dan papan kedudukan penanda aras yang menjadi asas kebanyakan dakwaan tentang pengekodan AI. Kelembapan 19% kekal selepas semakan kekukuhan oleh para penulis.
- **Penemuan yang paling mudah dibawa ke konteks lain ialah jurang persepsi.** Intuisi pakar ten-

tang peningkatan kelajuan mereka sendiri dengan AI tersasar kira-kira 40 mata, ke arah yang terlalu optimistik. Itu ialah amaran terhadap *setiap* dakwaan peningkatan produktiviti AI yang berdasarkan laporan sendiri — termasuk angka dalam kajian ini sendiri.

- **Ia ialah gambaran pada satu masa, bukan suatu trend.** Susulan METR sendiri pada Februari 2026, dengan jenis pembangun yang sama tetapi alat lebih baharu, menunjuk ke arah peningkatan kelajuan — secara sangat kasar —18% bagi pembangun yang kembali dan —4% bagi pembangun baharu — tetapi para penulis menandainya sebagai bukti lemah yang terjejas oleh siapa yang sanggup mengambil bahagian (pembangun semakin enggan bekerja tanpa AI, kadar bayaran menurun, dan pemilihan tugas menjadi berat sebelah). Tafsiran yang jujur ialah gambaran itu sedang berubah, dan perubahan itu sendiri dilaporkan tanpa cuba memihak kepada satu keputusan.

Mengapa ia penting

Sebahagian besar perdebatan tentang AI dan pengaturcaraan bergerak atas demo dan perasaan: rakaman skrin yang kemas, dakwaan penuh keyakinan, dan jelingan skeptikal sebagai balasan. Yang jarang berlaku — dan yang menjadikan kajian ini patut dibaca — ialah seseorang benar-benar menjalankan eksperimen yang membosankan: rawakkan tugas, ukur masa kerja sebenar, kemudian tanya orang bagaimana rasanya. Jawapannya tidak selesa bagi kedua-dua pihak. Ia mencucuk lubang pada cerita bahawa AI sentiasa memberi turbo kepada pembangun pakar pada kod sukar yang mereka kenali. Dan ia juga mencucuk lubang pada cerita bertentangan yang terlalu kemas — “AI menjadikan pembangun lebih perlahan, sudah terbukti” — kerana data lebih baharu daripada pasukan yang sama sudah cenderung ke arah sebaliknya. Pelajaran yang paling tahan lama juga yang paling kecil dan paling manusiawi: orang yang melakukan kerja itu berasa lebih pantas walaupun secara terukur lebih perlahan. “Rasanya lebih pantas” bukan bukti bahawa ia memang lebih pantas. Ukurlah.

Ringkasan utama

Dalam satu percubaan terkawal rawak, METR meminta enam belas pembangun sumber terbuka berpengalaman menyiapkan 246 tugas sebenar pada pangkalan kod yang mereka kenali dengan baik, dengan alat AI awal 2025 (Cursor Pro bersama Claude 3.5/3.7 Sonnet) dibenarkan secara rawak pada separuh tugas. Para pembangun menjangka AI akan mengurangkan masa tugas kira-kira 24%; sebaliknya, masa siap meningkat 19% — dan selepas itu mereka masih percaya AI telah mempercepatkan mereka kira-kira 20%. Jurang persepsi itu ialah teras kajian yang paling jelas dan kukuh. Tetapi kajian ini hanya melibatkan enam belas pembangun, satu keadaan yang sempit, dan gambaran beku awal 2025; para penulis secara jelas menyatakan bahawa ia tidak menunjukkan AI gagal membantu kebanyakan pembangun, dan susulan mereka sendiri pada 2026 sudah menunjuk ke arah peningkatan kelajuan, dengan batasannya tersendiri. Satu pengukuran teliti yang wajar dipandang serius — bukan keputusan muktamad tentang pengekodan AI.

Sumber

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/>