



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

Developer berpengalaman merasa lebih cepat dengan AI saat secara terukur justru lebih lambat — itulah temuan sebenarnya, bukan vonis atas coding dengan AI

Dalam randomized controlled trial, METR meminta enam belas developer open-source berpengalaman mengerjakan 246 tugas nyata pada codebase yang mereka kenal baik, dengan alat AI awal 2025 (Cursor Pro plus Claude 3.5/3.7 Sonnet) diizinkan pada separuh tugas secara acak. Mereka memperkirakan AI akan memangkas waktu tugas sekitar 24%; sebaliknya, waktu penyelesaian naik 19% — dan setelahnya mereka masih percaya AI telah mempercepat mereka sekitar 20%. Kesenjangan persepsi itulah inti studi yang tajam dan cukup kokoh. Namun ini enam belas developer, satu konteks sempit, dan potret alat awal 2025: penulis tegas bahwa hasil ini tidak menunjukkan AI gagal membantu sebagian besar developer, dan tindak lanjut mereka sendiri pada 2026 sudah condong ke arah percepatan, dengan keterbatasannya sendiri.

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

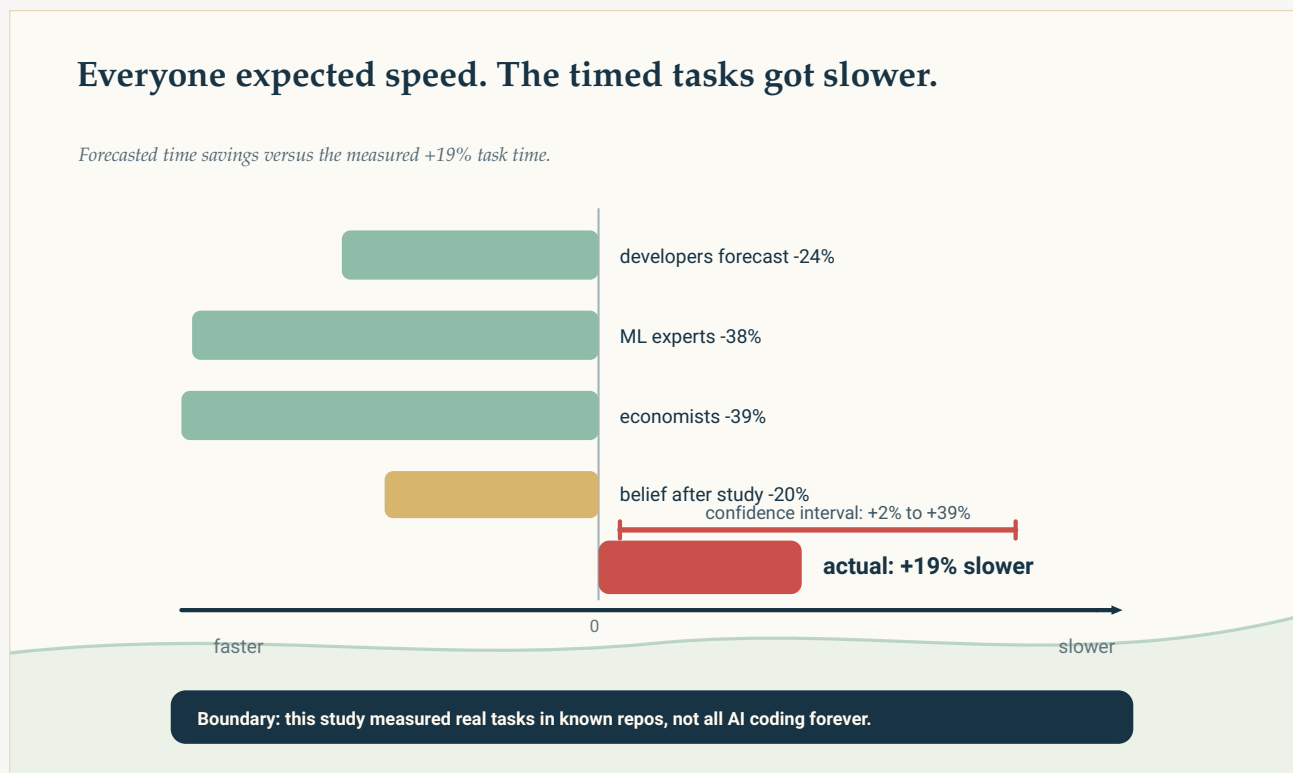
Paper: *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*, Joel Becker, Nate Rush, Beth Barnes, David Rein (METR), arXiv:2507.09089 [cs.AI] (preprint) · DOI: 10.48550/arXiv.2507.09089

Developer berpengalaman merasa lebih cepat dengan AI saat secara terukur justru lebih lambat — itulah temuan sebenarnya, bukan vonis atas coding dengan AI

Tanyakan kepada programmer berpengalaman apakah asisten coding AI mempercepat pekerjaannya dan biasanya Anda akan mendapat angka: menghemat dua puluh, tiga puluh persen. Tanyakan kepada ekonom atau peneliti machine learning, angkanya biasanya lebih besar. Pada awal 2025, sebuah tim di METR melakukan hal yang lambat dan mahal: mereka benar-benar mengukurnya. Mereka merekrut enam belas developer open-source berpengalaman, memberi mereka 246 tugas nyata dari codebase besar yang sangat mereka kenal, lalu — dengan pengacakan per tugas — mengizinkan penggunaan alat AI pada sebagian tugas dan melarangnya pada yang lain. Setelah itu mereka mengukur waktunya.

Para developer sebelumnya memperkirakan AI akan memangkas waktu pengerjaan sekitar 24%. Yang terjadi justru sebaliknya: tugas yang dikerjakan dengan AI memakan waktu **19% lebih lama**. Dan bagian yang layak direnungkan adalah ini — setelah selesai, developer yang sama masih percaya AI telah mempercepat mereka sekitar 20%. Mereka lebih lambat, tetapi merasa lebih cepat, dan jarak antara dua angka itu adalah bagian paling menarik dari penelitian ini.

Ini adalah hasil nyata yang diukur dengan hati-hati. Tetapi ini juga hanya enam belas developer, pada repositori yang mereka kenal sangat baik, menggunakan alat dari awal 2025 — dan bukan kalimat datar “AI membuat developer lebih lambat.” Data lanjutan dari tim yang sama sudah mengarah ke arah sebaliknya.



Semua pihak memperkirakan AI akan mempercepat pekerjaan — developer —24%, pakar ML —38%, ekonom —39%, dan keyakinan developer sendiri setelah studi —20%. Stopwatch justru menemukan +19% lebih lambat, dengan interval +2% hingga +39%. Jarak antara yang dirasakan dan yang terukur adalah temuannya.

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 studi — 16 developer open-source ahli, 246 tugas nyata, repositori yang dikenal, alat awal 2025, dan pengacakan — serta apa yang tidak diukur: bukan kebanyakan developer, bukan bidang lain, bukan hukum tetap, dan belum ditinjau seawal.

Original diagram — The Clean Paper CC BY 4.0

Apa yang diukur, dan apa keuntungan uji acak

Sebuah **randomized controlled trial (RCT)** adalah alat yang biasa dipakai kedokteran untuk membedakan efek nyata dari efek yang hanya diharapkan. Di sini, setiap dari 246 tugas secara acak ditetapkan untuk dikerjakan dengan AI diizinkan atau tidak, sehingga secara rata-rata satu-satunya perbedaan sistematis antara kedua kelompok adalah AI itu sendiri. Itulah yang memungkinkan kita mengatakan AI *menyebabkan* perubahan waktu, bukan sekadar mengamati bahwa orang yang memilih AI kebetulan lebih cepat atau lambat karena alasan lain. Ini penting karena bukti yang biasanya dipakai untuk klaim peningkatan coding dengan AI — laporan diri dan skor tolok ukur — tidak dapat melakukan hal tersebut: tolok ukur bukan pekerjaan nyata, dan laporan diri, seperti ditunjukkan studi ini, dapat salah dengan penuh keyakinan. Para developer di sini bukan pemula yang kikuk dengan mainan baru. Mereka adalah kontributor mapan pada proyek open-source besar dan matang yang sangat mereka kenal, serta sudah punya sebagian pengalaman menggunakan alat tersebut.

Apa yang dilakukan para penulis

- Menjalankan **randomized controlled trial** (METR: Joel Becker, Nate Rush, Beth Barnes, David Rein). Enam belas developer open-source berpengalaman masing-masing bekerja pada repositori besar yang rutin mereka kontribusikan dan kenal baik — rata-rata lima tahun pada proyek tertentu itu.
- Menggunakan **246 tugas nyata** — perbaikan bug, fitur, dan refaktor yang diambil dari issue tracker proyek mereka sendiri. Setiap tugas secara acak ditetapkan ke kondisi “AI diizinkan” atau “AI tidak diizinkan.”
- “AI diizinkan” berarti **alat awal 2025: Cursor Pro dengan Claude 3.5/3.7 Sonnet**. Ukuran utama adalah waktu penyelesaian nyata per tugas. Selain itu, tim mengumpulkan prediksi dari developer sebelum studi, estimasi mereka setelah studi, serta prediksi dari pakar ekonomi dan machine learning.

Apa yang mereka temukan

- **Dengan AI, tugas memakan waktu 19% lebih lama.** Bukan lebih cepat — lebih lambat. Interval keyakinan 95% berkisar dari kira-kira +2% hingga +39%, sehingga arahnya cukup jelas walau besar tepatnya masih tidak pasti.
- **Semua pihak telah memprediksi sebaliknya.** Developer memperkirakan percepatan 24%; pakar machine learning sekitar 38%; ekonom sekitar 39%. Ketiga kelompok berharap AI menghemat banyak waktu; stopwatch menemukan AI justru menambah waktu.
- **Kesenjangan persepsi.** Setelah menyelesaikan pekerjaan dan ternyata lebih lambat, developer masih memperkirakan AI telah mempercepat mereka sekitar 20% — jarak sekitar 40 poin antara yang mereka rasakan dan yang dicatat waktu.
- **Kemungkinan alasan, dipertimbangkan tetapi tidak dibuktikan.** Para penulis menyusun sejumlah faktor yang dapat menjelaskan perlambatan: developer ini sangat mengenal codebase mereka sehingga ruang bagi asisten untuk menambah nilai lebih kecil; proyek matang memiliki standar kualitas tinggi yang sering tidak tertulis; repositori besar menyimpan banyak konteks yang tidak dimiliki model; dan waktu nyata habis untuk memberi prompt, lalu meninjau serta memperbaiki keluaran AI. Semua ini diajukan sebagai petunjuk, bukan putusan.

Apa yang tidak ditunjukkan

- Ini **tidak** menunjukkan bahwa AI gagal mempercepat sebagian besar developer. Para penulis mengatakannya secara langsung: enam belas ahli yang bekerja pada kode yang mereka hafal bukan developer rata-rata pada tugas rata-rata.
- Ini **tidak** menunjukkan AI tidak berguna atau memperlambat orang di konteks lain — pendatang baru ke sebuah codebase, proyek greenfield, bahasa yang tidak dikenal, atau bidang lain sama sekali.

- Ini **tidak** membekukan kemampuan alat. Yang diuji adalah Cursor dan Claude 3.5/3.7 Sonnet awal 2025; penulis secara eksplisit mengatakan alat yang lebih baik, atau cara penggunaan yang lebih baik, dapat mengubah hasil bahkan dalam konteks yang sama.
- Ini adalah **pracetak** (diposting Juli 2025, belum ditinjau sejawat), dan penulis mencatat bahwa mereka tidak dapat sepenuhnya menyingkirkan artefak eksperimental — walau hasilnya bertahan dalam berbagai analisis mereka.
- Ini **tidak** mengizinkan pembacaan yang menenangkan bahwa developer pasti mendapat manfaat lain — belajar lebih banyak, lebih bahagia, atau menulis kode lebih baik. Hal yang diukur di sini, yaitu perasaan bahwa pekerjaan lebih cepat, justru bertentangan dengan data.

Seberapa kuat buktinya?

- **Desainnya luar biasa jujur.** Acak, tugas nyata, repositori nyata, pengukuran waktu nyata — langkah besar dibanding laporan diri dan papan peringkat tolok ukur yang menjadi dasar banyak klaim tentang coding dengan AI. Perlambatan 19% bertahan dalam pemeriksaan robustness penulis.
- **Temuan yang paling mudah dibawa ke konteks lain adalah kesenjangan persepsi.** Intuisi para ahli tentang percepatan diri sendiri dengan AI salah sekitar 40 poin, ke arah yang optimistis. Itu adalah peringatan untuk *setiap* klaim produktivitas AI yang dilaporkan sendiri — termasuk angka dari studi ini.
- **Ini potret sesaat, bukan tren.** Tindak lanjut METR pada Februari 2026, dengan tipe developer yang sama dan alat yang lebih baru, mengarah ke percepatan — sangat kasar sekitar —18% untuk peserta yang kembali dan -4% untuk peserta baru — tetapi para penulis sendiri menandainya sebagai bukti lemah, terdistorsi oleh siapa yang bersedia ikut (developer semakin enggan bekerja tanpa AI, tarif bayaran turun, dan pemilihan tugas bergeser). Pembacaan jujur: gambarnya sedang berubah, dan bahkan perubahan itu dilaporkan dengan hati-hati.

Mengapa ini penting

Sebagian besar perdebatan tentang AI dan pemrograman berjalan dengan demo dan perasaan: rekaman layar yang mulus, klaim penuh percaya diri, lalu cibiran dari pihak lain. Yang langka, dan membuat studi ini layak dibaca, adalah seseorang melakukan eksperimen yang membosankan: acak tugasnya, ukur pekerjaan nyata, lalu tanya orang bagaimana rasanya. Jawabannya tidak nyaman bagi kedua kubu. Ia menusuk cerita bahwa AI secara seragam membuat developer ahli melesat pada kode sulit yang sudah mereka kenal. Namun ia juga menusuk cerita kebalikannya — “AI terbukti membuat developer lebih lambat” — karena data terbaru dari tim yang sama sudah mengarah ke arah sebaliknya. Pelajaran paling tahan lama juga yang paling kecil dan paling manusiawi: orang yang mengerjakan tugas merasa lebih cepat ketika secara terukur mereka lebih lambat. “Rasanya lebih cepat” bukan bukti bahwa memang lebih cepat. Ukurlah.

Singkatnya

Dalam sebuah randomized controlled trial, METR meminta enam belas developer open-source berpengalaman menyelesaikan 246 tugas nyata pada codebase yang mereka kenal baik, dengan alat AI awal 2025 (Cursor Pro plus Claude 3.5/3.7 Sonnet) diizinkan pada separuh tugas secara acak. Developer memperkirakan AI akan memangkas waktu tugas sekitar 24%; sebaliknya, AI menaikkan waktu penyelesaian 19% — dan setelahnya mereka masih percaya AI mempercepat mereka sekitar 20%. Kesenjangan persepsi itulah inti studi yang tajam dan cukup kokoh. Tetapi ini enam belas developer, satu konteks sempit, dan potret alat awal 2025; para penulis tegas bahwa hasil ini tidak menunjukkan AI gagal membantu sebagian besar developer, dan tindak lanjut mereka sendiri pada 2026 sudah mengarah ke percepatan, dengan keterbatasannya sendiri. Sebuah pengukuran hati-hati yang layak dianggap serius — bukan vonis atas coding dengan 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/>