



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

Erfahrene Entwickler fühlten sich mit KI schneller, während sie messbar langsamer arbeiteten — der wirkliche Befund, und warum er kein Urteil über KI-Coding ist

Eine randomisierte kontrollierte Studie von METR ließ sechzehn erfahrene Open-Source-Entwickler 246 echte Aufgaben auf Codebasen erledigen, die sie gut kannten — KI-Werkzeuge von Anfang 2025 (Cursor Pro plus Claude 3.5/3.7 Sonnet) waren auf einer zufällig gewählten Hälfte erlaubt. Sie erwarteten, die KI werde die Bearbeitungszeit um etwa 24 % senken; stattdessen erhöhte sie sie um 19 % — und hinterher glaubten sie immer noch, sie seien um etwa 20 % schneller geworden. Diese Wahrnehmungslücke ist der scharfe, robuste Kern der Studie. Aber es sind sechzehn Entwickler, ein enges Setting und eine fixe Momentaufnahme von Anfang 2025: Die Autoren sagen ausdrücklich, dies zeige nicht, dass KI den meisten Entwicklern nicht hilft — und ihr eigenes Follow-up von 2026 deutet bereits auf eine Beschleunigung hin, mit eigenen Vorbehalten.

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

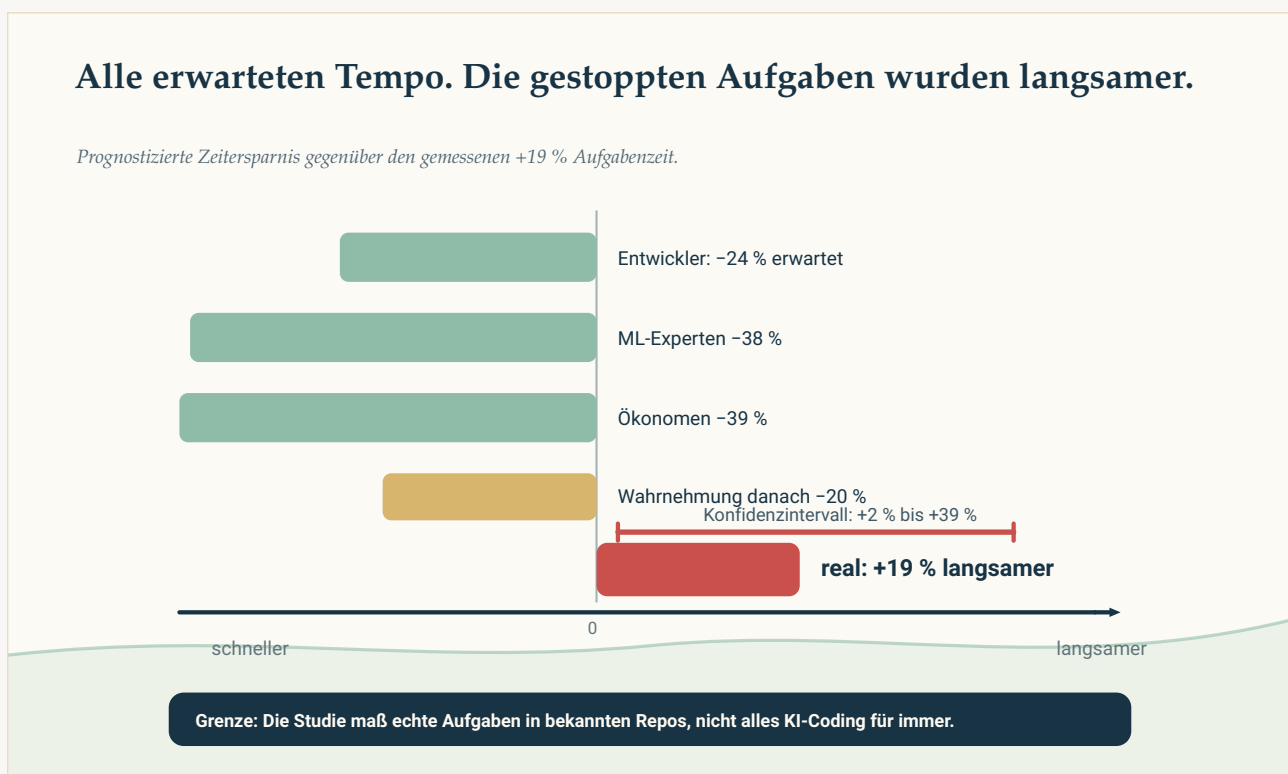
Paper: *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*, Joel Becker, Nate Rush, Beth Barnes, David Rein (METR), arXiv:2507.09089 [cs.AI] (preprint) · DOI: 10.48550/arXiv.2507.09089

Erfahrene Entwickler fühlten sich mit KI schneller, während sie messbar langsamer arbeiteten — der wirkliche Befund, und warum er kein Urteil über KI-Coding ist

Fragen Sie einen erfahrenen Programmierer, ob ein KI-Coding-Assistent ihn schneller macht, und Sie bekommen meist eine Zahl: spart mir zwanzig, dreißig Prozent. Fragen Sie einen Ökonomen oder eine Machine-Learning-Forscherin, und die Zahl wird größer. Anfang 2025 tat ein Team bei METR das Langsame, Teure: Es prüfte nach. Es nahm sechzehn erfahrene Open-Source-Entwickler, gab ihnen 246 echte Aufgaben aus großen Codebasen, die sie gut kannten, und ließ sie — per Münzwurf, Aufgabe für Aufgabe — entweder KI-Werkzeuge benutzen oder nicht. Dann stoppte es die Zeit.

Die Entwickler hatten vorhergesagt, die KI werde ihre Bearbeitungszeit um etwa 24 % senken. Sie tat das Gegenteil: Die mit KI erledigten Aufgaben dauerten **19 % länger**. Und hier ist der Teil, bei dem es sich zu verweilen lohnt — nach getaner Arbeit glaubten dieselben Entwickler immer noch, die KI habe sie schneller gemacht, um etwa 20 %. Sie waren langsamer, und sie fühlten sich schneller, und der Abstand zwischen diesen beiden Zahlen ist das Interessanteste an der Studie.

Das ist ein echtes, sorgfältig gemessenes Ergebnis. Es sind aber auch sechzehn Entwickler, auf Repositories, die sie in- und auswendig kennen, mit den Werkzeugen von Anfang 2025 — und es ist nicht der platte Satz „KI macht Entwickler langsamer“. Die späteren Daten desselben Teams deuten bereits in die andere Richtung.



Alle sagten voraus, die KI werde die Arbeit beschleunigen — Entwickler –24 %, ML-Experten –38 %, Ökonomen –39 % und die eigene nachträgliche Einschätzung der Entwickler –20 %. Die Stoppuhr fand +19 %

langsamer, mit einem Intervall von +2 % bis +39 %. Der Abstand zwischen gefühlt und gemessen ist der Befund.

Original diagram — The Clean Paper CC BY 4.0

Was diese KI-Produktivitätsstudie gemessen hat

DAS IST

- 16 erfahrene Open-Source-Entwickler
- 246 echte Aufgaben
- Repos, die sie kannten
- randomisiert und gestoppt
- KI-Tools von Anfang 2025

DAS IST NICHT

- nicht die meisten Entwickler
- nicht jede Domäne
- kein festes Gesetz
- nicht peer-reviewed
- kein Urteil über künftige Tools

Grenze: nützliche Evidenz über ein Setting, kein universelles Gesetz der KI-Arbeit.

Was die Studie misst — 16 erfahrene Open-Source-Entwickler, 246 echte Aufgaben, vertraute Repos, Tools von Anfang 2025, randomisiert — und was nicht: nicht die meisten Entwickler, nicht andere Domänen, kein festes Gesetz, nicht peer-reviewed.

Original diagram — The Clean Paper CC BY 4.0

Was hier gemessen wurde, und was eine randomisierte Studie einem bringt

Eine **randomisierte kontrollierte Studie (RCT)** ist das Werkzeug, mit dem die Medizin einen echten Effekt von einem erhofften unterscheidet. Hier wurde jede der 246 Aufgaben zufällig zugeteilt: KI erlaubt oder nicht. So ist im Mittel der einzige systematische Unterschied zwischen den beiden Stapeln die KI selbst. Genau das erlaubt die Aussage, die KI habe die Zeitveränderung *verursacht* — statt bloß zu bemerken, dass Leute, die zur KI greifen, aus anderen Gründen schneller oder langsamer sind. Das zählt, weil die üblichen Belege für KI-Coding-Gewinne — Selbstauskünfte und Benchmark-Ergebnisse — das nicht leisten können: Ein Benchmark ist keine echte Arbeit, und eine Selbstauskunft kann, wie diese Studie zeigt, mit voller Überzeugung falsch sein. Die Entwickler hier waren keine Anfänger, die an einem neuen Spielzeug herumfummelten. Sie waren etablierte Mitwirkende großer, gereifter Open-Source-Projekte, die sie gut kannten, mit einiger Vorerfahrung im Umgang mit den Werkzeugen.

Was die Autoren getan haben

- Sie führten eine **randomisierte kontrollierte Studie** durch (METR: Joel Becker, Nate Rush, Beth Barnes, David Rein). Sechzehn erfahrene Open-Source-Entwickler, jeder an einem großen Repository, zu dem er regelmäßig beiträgt und das er gut kennt — im Schnitt fünf Jahre auf den jeweiligen Projekten.
- Sie verwendeten **246 echte Aufgaben** — Bugfixes, Features und Refactorings aus den Issue-Trackern eben dieser Projekte. Jede Aufgabe wurde zufällig „KI erlaubt“ oder „KI nicht erlaubt“ zugeteilt.
- „KI erlaubt“ hieß **Tooling von Anfang 2025: Cursor Pro mit Claude 3.5/3.7 Sonnet**. Das primäre Maß war die tatsächliche Bearbeitungszeit pro Aufgabe. Daneben sammelte das Team Prognosen (von den Entwicklern vorher) und Schätzungen (von denselben hinterher), dazu Vorhersagen von Fachleuten aus Ökonomie und Machine Learning.

Was sie fanden

- **Mit KI dauerten die Aufgaben 19 % länger.** Nicht schneller — langsamer. Das 95-%-Konfidenzintervall reicht von rund +2 % bis +39 %; die Richtung ist also belastbar, auch wo die genaue Größe es nicht ist.
- **Alle hatten das Gegenteil vorhergesagt.** Die Entwickler prognostizierten eine Beschleunigung um 24 %, Machine-Learning-Fachleute etwa 38 %, Ökonomen etwa 39 %. Alle drei Gruppen erwarteten, die KI werde viel Zeit sparen; die Stoppuhr fand, dass sie welche kostete.
- **Die Wahrnehmungslücke.** Nachdem sie die Arbeit getan hatten und langsamer herausgekommen waren, schätzten die Entwickler immer noch, die KI habe sie um etwa 20 % beschleunigt — eine Lücke von rund 40 Punkten zwischen dem, was sie fühlten, und dem, was die Uhr aufzeichnete.
- **Mögliche Gründe — abgewogen, nicht bewiesen.** Die Autoren stellen Faktoren nebeneinander, die die Verlangsamung erklären könnten: Diese Entwickler kennen ihre Codebasen in der Tiefe, ein Assistent kann also weniger beisteuern; gereifte Projekte tragen hohe, oft implizite Qualitätsstandards; die Repositories sind groß und voller Kontext, den ein Modell nicht hat; und echte Zeit fließt ins Prompten und dann ins Prüfen und Korrigieren der KI-Ausgabe. Sie präsentieren das als Spuren, nicht als Urteile.

Was das nicht zeigt

- Es zeigt **nicht**, dass KI die meisten Entwickler nicht beschleunigt. Die Autoren sagen es direkt: Sechzehn Experten auf Code, den sie auswendig kennen, sind nicht der durchschnittliche Entwickler an der durchschnittlichen Aufgabe.
- Es zeigt **nicht**, dass KI nutzlos ist oder Menschen in anderen Umgebungen verlangsamt — Neulinge in einer Codebasis, Arbeit auf der grünen Wiese, unvertraute Sprachen oder ganz

andere Felder.

- Es friert die Werkzeuge **nicht** ein. Das ist Cursor und Claude 3.5/3.7 Sonnet von Anfang 2025; die Autoren sagen ausdrücklich, dass bessere Werkzeuge — oder bessere Arten, diese zu benutzen — das Ergebnis selbst in genau diesem Setting ändern könnten.
- Es ist ein **Preprint** (online gestellt im Juli 2025, noch nicht im Peer-Review begutachtet), und die Autoren merken an, dass sie experimentelle Artefakte nicht vollständig ausschließen können — auch wenn der Befund ihre Analysen überstand.
- Es erlaubt **nicht** die beruhigende Lesart, die Entwickler müssten auf andere Weise gewonnen haben — mehr gelernt, sich wohler gefühlt, besseren Code geschrieben. Das eine, was hier gemessen wurde, die gefühlte Beschleunigung, ist genau das, was die Daten widerlegen.

Wie belastbar sind die Belege?

- **Das Design ist ungewöhnlich ehrlich.** Randomisiert, echte Aufgaben, echte Repositories, tatsächliche Zeitmessung — ein großer Schritt über die Selbstauskünfte und Benchmark-Ranglisten hinaus, auf denen die meisten Behauptungen über KI-Coding ruhen. Die Verlangsamung um 19 % überstand die Robustheitsprüfungen der Autoren.
- **Der übertragbarste Befund ist die Wahrnehmungslücke.** Die Intuition von Experten über die eigene KI-Beschleunigung lag um rund 40 Punkte daneben, in die optimistische Richtung. Das ist eine Warnung vor *jedem* selbstberichteten Produktivitätsgewinn durch KI — die Zahlen dieser Studie eingeschlossen.
- **Es ist eine Momentaufnahme, kein Trend.** METRs eigenes Follow-up vom Februar 2026, mit derselben Art von Entwicklern auf neueren Werkzeugen, deutet auf eine Beschleunigung hin — sehr grob –18 % für wiederkehrende Entwickler und –4 % für neue —, aber die Autoren stufen es als schwache Evidenz ein, verzerrt dadurch, wer teilzunehmen bereit war (Entwickler lehnten es zunehmend ab, ohne KI zu arbeiten, die Vergütung sank, und die Aufgabenauswahl verschob sich). Die ehrliche Lesart: Das Bild bewegt sich, und selbst die Bewegung wird berichtet, ohne den Daumen auf die Waage zu legen.

Warum das wichtig ist

Der größte Teil der Debatte über KI und Programmierung läuft auf Demos und Bauchgefühl: eine glatte Bildschirmaufnahme, eine selbstbewusste Behauptung, ein Augenrollen von der Gegenseite. Selten ist — und lesenswert macht diese Arbeit —, dass jemand das langweilige Experiment durchgeführt hat: randomisieren, echte Arbeit mit der Stoppuhr messen, danach die Leute fragen, wie es lief. Die Antwort ist für beide Lager unbequem. Sie durchsticht die Geschichte, KI verleihe erfahrenen Entwicklern auf schwerem, vertrautem Code durchweg einen Turbo. Und sie durchsticht auch das ordentliche Gegenteil — „KI macht Entwickler langsamer, bewiesen“ —, denn die neueren Daten desselben Teams neigen bereits zur anderen Seite. Die haltbarste Lehre ist zugleich die kleinste und die menschlichste: Die Menschen, die die Arbeit taten, fühlten sich schneller, während sie messbar langsamer waren. „Es fühlt sich schneller an“ ist kein Beleg, dass es so ist. Messen Sie nach.

Kurz zusammengefasst

In einer randomisierten kontrollierten Studie ließ METR sechzehn erfahrene Open-Source-Entwickler 246 echte Aufgaben auf Codebasen erledigen, die sie gut kannten — KI-Werkzeuge von Anfang 2025 (Cursor Pro plus Claude 3.5/3.7 Sonnet) waren auf einer zufälligen Hälfte erlaubt. Die Entwickler erwarteten, die KI werde ihre Bearbeitungszeit um etwa 24 % senken; stattdessen erhöhte sie die Fertigstellungszeit um 19 % — und hinterher glaubten sie immer noch, sie habe sie um etwa 20 % beschleunigt. Diese Wahrnehmungslücke ist der scharfe, robuste Kern der Studie. Aber es sind sechzehn Entwickler, ein enges Setting und eine fixe Momentaufnahme von Anfang 2025; die Autoren sagen ausdrücklich, dass dies nicht zeigt, KI helfe den meisten Entwicklern nicht — und ihr eigenes Follow-up von 2026 deutet bereits auf eine Beschleunigung hin, mit eigenen Vorbehalten. Eine sorgfältige Messung, die man ernst nehmen sollte — kein Urteil über KI-Coding.

Quellen

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/>