



The CLEAN PAPER

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

Reyndum forriturum fannst þeir vinna hraðar með gervigreind þótt mælingar sýndu að þeir unnu hægar — þetta er raunverulega niðurstaðan, ekki endanlegur dómur um gervigreind við forritun

Í slembuðum samanburði METR unnu 16 reyndir opnir hugbúnaðarforritarar 246 raunverkefni í kóðagrunnum sem þeir þekktu. Þeir bjuggust við um 24% hraðari vinnu með gervigreind en mældur tími varð 19% lengri — og eftir á töldu þeir samt að hún hefði flýtt fyrir þeim. Rannsóknin er þröng mynd frá byrjun árs 2025, ekki almennur dómur um gervigreind við forritun.

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

Reyndum forriturum fannst þeir vinna hraðar með gervigreind þótt mælingar sýndu að þeir unnu hægar — þetta er raunverulega niðurstaðan, ekki endanlegur dómur um gervigreind við forritun

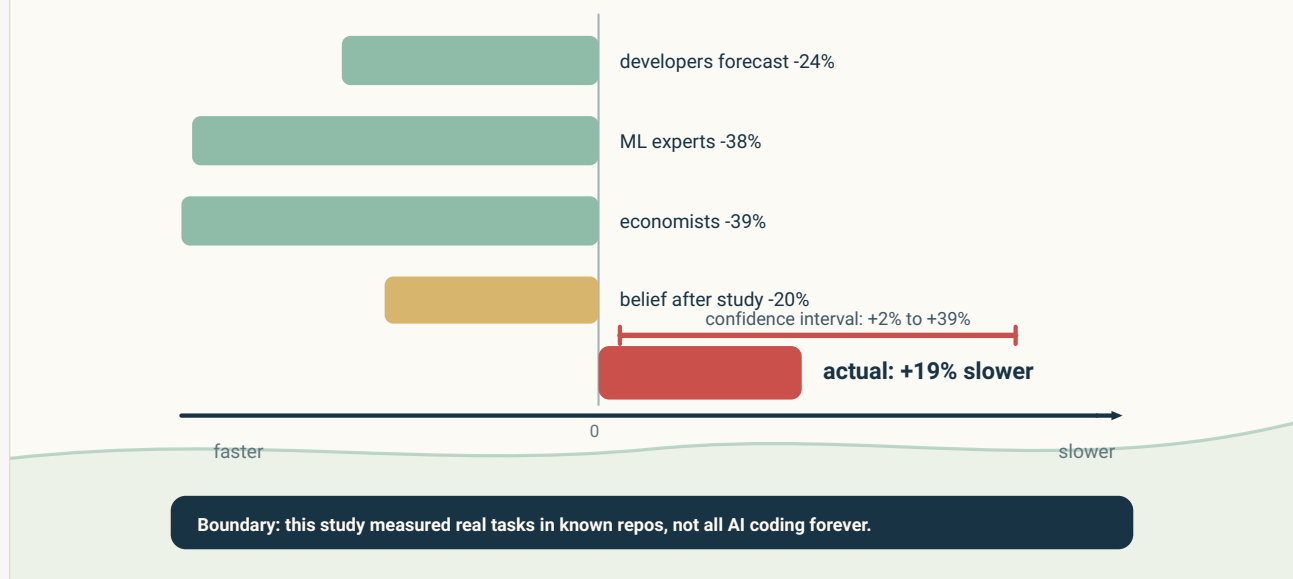
Spyrðu reyndan forritara hvort gervigreindaraðstoð við kóðun flýti fyrir vinnu hans og oft færðu tölu: hún sparar mér tuttugu, þrjátíu prósent. Spyrdum hagfræðing eða rannsakanda í vélanámi og talan verður gjarnan enn hærri. Snemma árs 2025 ákvað teymi hjá METR að gera hæga og dýra hlutinn: að mæla þetta. Sextán reyndir þátttakendur í opnum hugbúnaðarverkefnum fengu 246 raunveruleg verkefni úr stórum kóðagrunnum sem þeir þekktu vel. Fyrir hvert verkefni var með slembivali ákveðið hvort þeir mættu nota gervigreindarverkfæri eða ekki. Síðan var tíminn mældur.

Forritararnir höfðu spáð því að gervigreind myndi stytta verkefnatímann um um 24%. Niðurstaðan varð öfug: verkefnin sem unnin voru með gervigreind tóku **19% lengri tíma**. Og hér er sá hluti sem vert er að staldra við: eftir að vinnunni lauk töldu sömu forritarar enn að gervigreindin hefði flýtt fyrir þeim um um 20%. Þeir voru hægari en þeim fannst þeir vera hraðari, og bilið á milli þessara tveggja talna er athyglisverðasti hluti rannsóknarinnar.

Þetta er raunveruleg niðurstaða sem var mæld vandlega. En hún byggir líka á sextán forriturum sem unnu í kóðagrunnum sem þeir þekkja náið, með verkfærum frá byrjun árs 2025 — og hún jafngildir ekki einföldu fullyrðingunni „gervigreind gerir forritara hægari“. Nýrri gögn frá sama teymi benda þegar í hina áttina.

Everyone expected speed. The timed tasks got slower.

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



Allir spáðu því að gervigreind myndi flýta vinnunni — forritarar —24%, sérfræðingar í vélanámi —38%, hagfræðingar —39% og mat forritaranna eftir á —20%. Tímamælingin sýndi hins vegar +19% hægari vinnu, með öryggisbili frá +2% til +39%. Bilið milli upplifunar og mælingar er sjálf niðurstaðan.

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.

Það sem rannsóknin mælir — 16 reynda opna hugbúnaðarforritara, 246 raunveruleg verkefni, kunnug kóðasöfn, verkfæri frá byrjun árs 2025 og slembiröðun — og það sem hún mælir ekki: ekki flesta forritara, ekki önnur svið, ekki fast lögmál og ekki ritrýnda niðurstöðu.

Original diagram — The Clean Paper CC BY 4.0

Hvað var mælt og hvað slembuð samanburðarrannsókn gefur

Slembuð samanburðarrannsókn (RCT) er aðferðin sem meðal annars er notuð í læknisfræði til að greina raunveruleg áhrif frá vonum og tilviljunum. Hér var hverju af 246 verkefnum slembiraðað þannig að annaðhvort var gervigreind leyfð eða ekki. Því er að meðaltali aðeins einn kerfisbundinn munur á hópunum: gervigreindin sjálf. Það gerir kleift að segja að gervigreindin *olli* breytingunni á tímanum, í stað þess að sjá einungis fylgni milli þess að fólk velji að nota gervigreind og þess að það sé af öðrum ástæðum hraðara eða hægara. Þetta skiptir máli vegna þess að algengustu gögnin um ávinning af gervigreind við forritun — sjálfsmat og niðurstöður á stöðluðum prófum — geta ekki sýnt þetta. Staðlað próf er ekki raunveruleg vinna og sjálfsmat getur, eins og þessi rannsókn sýnir, verið sannfærandi en rangt. Þátttakendurnir voru heldur ekki byrjendur að fíkta með nýtt leikfang. Þeir voru rótgrónir þátttakendur í stórum og þroskuðum opnum hugbúnaðarverkefnum sem þeir þekktu vel og höfðu áður notað sambærileg verkfæri að einhverju leyti.

Hvað höfundarnir gerðu

- Þeir framkvæmdu **slembuðu samanburðarrannsóknina** (METR: Joel Becker, Nate Rush, Beth Barnes, David Rein). Sextán reyndir opnir hugbúnaðarforritarar unnu hver í stóru kóðasafni sem þeir leggja reglulega til og þekkja vel — að meðaltali höfðu þeir unnið við viðkomandi verkefni í fimm ár.
- Notuð voru **246 raunveruleg verkefni** — villuleiðréttingar, ný virkni og endurskipulagning kóða úr eigin verkbeiðnakerfum verkefnanna. Hverju verkefni var slembiraðað í „gervigreind leyfð“ eða „gervigreind óleyfð“.
- „Gervigreind leyfð“ þýddi **verkfæri frá byrjun árs 2025: Cursor Pro með Claude 3.5/3.7 Sonnet**. Aðalmælikvarðinn var raunverulegur tími sem tók að ljúka hverju verkefni. Að auki söfnuðu rannsakendur spám forritaranna fyrirfram og mati þeirra eftir á, auk spáa hagfræðinga og sérfræðinga í vélanámi.

Hvað þeir fundu

- **Með gervigreind tók vinnan 19% lengri tíma.** Ekki styttri — lengri. 95% öryggisbilið nær frá um +2% til +39%, þannig að stefnan er nokkuð skýr þótt nákvæm stærð áhrifanna sé óvissari.
- **Allir höfðu spáð öfugu.** Forritararnir bjuggust við 24% hraðaaukningu, sérfræðingar í vélanámi um 38% og hagfræðingar um 39%. Allir hóparnir töldu að gervigreind myndi spara mikinn tíma; tímamælingin sýndi kostnað í staðinn.
- **Skynjunarbilið.** Eftir að hafa unnið hægar mátu forritararnir samt að gervigreind hefði flýtt fyrir þeim um um 20% — um 40 prósentustiga bil milli upplifunar og mælds tíma.
- **Mögulegar skýringar, vegnar en ekki sannaðar.** Höfundarnir nefna þætti sem gætu skýrt hæginguna: þessir forritarar þekkja eigin kóðasöfn mjög vel og hafa því minna að græða á aðstoð; þroskuð verkefni hafa háar og oft óskráðar gæðakröfur; kóðasöfnin eru stór og full af samhengi sem líkanið hefur ekki; og raunverulegur tími fer í að semja fyrirmæli, yfirfara niðurstöður og leiðrétta gervigreindarkóða. Þeir setja þetta fram sem mögulegar vísbendingar, ekki lokaúrskurð.

Hvað rannsóknin sýnir ekki

- Hún sýnir **ekki** að gervigreind flýti ekki fyrir flestum forriturum. Höfundarnir segja það skýrt: sextán sérfræðingar að vinna í kóða sem þeir þekkja utan að eru ekki dæmigerður forritari við dæmigert verkefni.
- Hún sýnir **ekki** að gervigreind sé gagnslaus eða að hún hægi á fólki við aðrar aðstæður — til dæmis nýliðum í kóðasafni, við ný verkefni frá grunni, í ókunnugum forritunarmálum eða á allt öðrum sviðum.
- Hún frystir **ekki** tæknina á einum tímapunkti. Þetta eru Cursor og Claude 3.5/3.7 Sonnet frá byrjun árs 2025; höfundarnir taka fram að betri verkfæri eða betri notkun þeirra gætu breytt niðurstöðunni jafnvel við sömu aðstæður.

- Þetta er **forprent** (birt í júlí 2025 og ekki enn ritrynt), og höfundarnir segja að ekki sé hægt að útiloka alla tilraunaskekkju — þótt niðurstaðan hafi haldist í ýmsum næmnigreiningum þeirra.
- Hún réttlætir **ekki** þægilegu túlkunina að forritararnir hljóti að hafa unnið eitthvað annað upp — lært meira, liðið betur eða skrifað betri kóða. Það sem var beinlínis mælt hér, upplifaður hraðaávinningur, er einmitt það sem gögnin stangast á við.

Hversu sterk eru gögnin?

- **Rannsóknarsniðið er óvenju hreint.** Slembiröðun, raunveruleg verkefni, raunveruleg kóðasöfn og raunveruleg tímamæling eru stórt skref fram yfir sjálfsmat og stigatöflur prófa sem margar fullyrðingar um gervigreindarkóðun byggja á. 19% hægingin stóðst næmnipróf höfundanna.
- **Almennasta niðurstaðan er líklega skynjunarbilið.** Mat sérfræðinganna á eigin hraðaaukningu með gervigreind var um 40 prósentustig frá mælingunni, og skekkjan var bjartsýn. Það er áminning um að fara varlega með öll sjálfskráð framleiðniaukningartölur vegna gervigreindar — þar með talið tölurnar í þessari rannsókn.
- **Þetta er ljósmynd af augnabliki, ekki þróunarlína.** Eftirfylgni METR frá febrúar 2026, með samþærulegum forriturum og nýrri verkfærum, bendir frekar til hraðaaukningar — mjög gróflega — 18% hjá þeim sem sneru aftur og –4% hjá nýjum þátttakendum — en höfundarnir kalla gögnin veik og benda á valskekkju: fleiri vildu ekki lengur vinna án gervigreindar, greiðslur lækkuðu og verkefnaval breyttist. Hreinskilna túlkunin er að myndin er á hreyfingu og jafnvel sú hreyfing er óviss.

Af hverju skiptir þetta máli?

Stór hluti umræðunnar um gervigreind og forritun byggist á sýnidæmum og tilfinningu: glæsileg skjáupptaka, sjálfsörugg fullyrðing eða jafn sjálfsörugg augnarúlla. Það sem er sjaldgæft — og gerir þessa rannsókn lestrarverða — er að einhver gerði leiðinlegu tilraunina: slembiraðaði, mældi raunverulega vinnu og spurði síðan fólk hvernig því fannst ganga. Niðurstaðan er óþægileg fyrir báðar fylkingar. Hún stingur gat á söguna um að gervigreind hraði alltaf mjög reyndum forriturum í erfiðum, kunnugum kóða. En hún stingur líka gat á snyrtilegu gagnsöguna „gervigreind gerir forritara hægari, sannað“, því nýrri gögn sama teymis halla nú þegar í hina áttina. Þórnasta lærdóminn er jafnframt sá minnsti og mannlegasti: fólkinu sem vann verkið fannst það hraðara á sama tíma og mælingar sýndu að það var hægara. „Mér finnst þetta hraðara“ er ekki sönnun fyrir því að svo sé. Mældu það.

Í stuttu máli

Í slembuðu samanburðarrannsókn lét METR sextán reynda opna hugbúnaðarforritara leysa 246 raunveruleg verkefni í kóðasöfnum sem þeir þekktu vel. Í slembivöldum helmingi verkefnanna máttu þeir nota gervigreindarverkfæri frá byrjun árs 2025 (Cursor Pro með Claude 3.5/3.7 Sonnet). Forritararnir bjuggust við að gervigreind myndi stytta verkefnatímann um um 24%; í staðinn jókst hann um 19% — og eftir á töldu þeir samt að hún hefði flýtt fyrir þeim um um 20%. Þetta skynjunarbil er skarpasti og traustasti kjarni rannsóknarinnar. En þetta voru sextán forritarar, mjög afmarkaðar aðstæður og ákveðið tæknilegt augnablik í byrjun 2025. Höfundarnir taka skýrt fram að rannsóknin sýnir ekki að gervigreind hjálpi ekki flestum forriturum, og eigin eftirfylgni þeirra frá 2026 bendir þegar til hraðaaukningar, þó með sínum eigin fyrirvörum. Vönduð mæling sem ber að taka alvarlega — ekki endanlegur dómur um gervigreind við forritun.

Heimildir

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/>