

Discussion Paper Series

IZA DP No. 18984

September 2026

Can AI Improve Household Welfare in Poorer Countries? Evidence and Measurement Challenges

Hai-Anh H. Dang

World Bank, IZA@LISER, GLO,
Indiana University, LSE and UEH

Minh N. N. Do

Vietnam National University

The IZA Discussion Paper Series (ISSN: 2365-9793) ("Series") is the primary platform for disseminating research produced within the framework of the IZA@LISER Network, an unincorporated international network of labour economists coordinated by the Luxembourg Institute of Socio-Economic Research (LISER). The Series is operated by LISER, a Luxembourg public establishment (établissement public) registered with the Luxembourg Business Registers under number J57, with its registered office at 11, Porte des Sciences, 4366 Esch-sur-Alzette, Grand Duchy of Luxembourg.

Any opinions expressed in this Series are solely those of the author(s). LISER accepts no responsibility or liability for the content of the contributions published herein. LISER adheres to the European Code of Conduct for Research Integrity. Contributions published in this Series present preliminary work intended to foster academic debate. They may be revised, are not definitive, and should be cited accordingly. Copyright remains with the author(s) unless otherwise indicated.

Can AI Improve Household Welfare in Poorer Countries? Evidence and Measurement Challenges*

Abstract

Can artificial intelligence (AI) make scarce expertise more widely available in poorer countries and improve household welfare? Can it help measure welfare where surveys are infrequent? This structured critical review examines both questions and finds that the answers depend on the application and setting. Across nine selected evaluations, AI assistance raised output per unit of time by 29% on average, and the average remains positive in every reported sensitivity check. Eight tutoring comparisons give an average gain of 0.21 standard deviations on tests taken without the tool, including in poorer-country schools, but this result is less robust when higher-risk trials are excluded. Business and labor-market effects vary, and our search found no qualifying causal study of a generative-AI tool's effect on household consumption poverty in a poorer country. Measured use rises steeply with national income. AI-based models predict wealth levels more accurately than changes, and their value for targeting depends on the alternative and its cost. The review identifies promising applications for evaluated expansion, proposes a survey module, and sets out conditions for validating AI-based statistics and for assessing costs, persistence, and equity.

JEL classification

O33, I32, D63, J24, C81

Keywords

artificial intelligence, poverty, inequality, developing countries, measurement, critical review

Corresponding author

Hai-Anh H. Dang

hdang@worldbank.org

* We would like to thank Gero Carletto, Gabriel Demombynes, Jed Friedman, Peter Lanjouw, and various colleagues at and outside the World Bank for helpful discussions. We are grateful to the UK Foreign, Commonwealth and Development Office (FCDO) for funding assistance through the Data and Evidence for Tackling Extreme Poverty (DEEP) Research Program.

1. Introduction

Poorer (low- and middle-income) countries face shortages of expertise across many fields, including education, health, agriculture, and small-business support. Artificial intelligence (AI) could make some of this expertise cheaper and more widely available. Yet, better answers may not translate into better learning, higher earnings, or more effective services. Access is also highly uneven. Microsoft estimates that, in the first quarter of 2026, about one in three working-age adults had used a generative-AI tool in the average high-income economy, compared with fewer than one in ten in the average low-income economy. This gap widened over the preceding year (Figure 1; Table A1). AI may also reduce demand for services exported by poorer countries.

Policymakers therefore face two related questions: Which AI applications are likely to improve household welfare?¹ How can their effects be measured where reliable welfare data are scarce? Improvements in one dimension do not necessarily produce gains in the others.

Measurement is part of the problem. Many poorer countries field a household consumption survey only once every several years, so the outcomes that AI might change are measured least often where AI use is lowest.² The review therefore looks at AI through two lenses. Under the intervention lens, AI is a treatment, and its effects on earnings, jobs, learning, service quality, poverty, and inequality are judged against what people would otherwise receive, such as usual instruction. Under the measurement lens, AI is a tool for estimating poverty and welfare, and its estimates are judged against observed data (the “ground truth”) and against the best feasible alternative.

¹ We define household welfare broadly to include material living standards, learning and health outcomes, time savings, and access to services.

² Poorer countries have fewer household surveys and weaker analytical capacity, particularly for panel surveys that can track poverty mobility over time (Jerven 2013; Serajuddin *et al.* 2015; World Bank 2021; Dang and Lanjouw 2026).

The intervention lens concentrates on generative AI, systems that produce text, code, images, or speech on request. Older predictive or forecasting applications are included as background evidence from earlier AI. The measurement lens concerns mostly predictive machine learning, which estimates an outcome or classifies an observation. Different questions need different evidence: what a model can do does not show who uses it or whether households benefit, and a prediction is not a measurement until it has been validated.

These questions are pressing because important decisions are being made now. Policy reports present AI as a path to national competitiveness (OECD 2025; World Economic Forum and Kearney 2026) and, for poorer economies, to prosperity (World Bank 2026). Governments are already choosing AI systems for schools, clinics, and social programs and deciding how to obtain computing capacity (often called “compute”). Scores on standardized technical tests of AI models track rapid progress but offer limited guidance for these decisions, because they show how models perform on set tasks, not how their applications work in schools, clinics, or offices.

This review organizes the evidence around three findings. First, AI can raise productivity on selected tasks and improve learning in some tutoring programs, but it can also cause losses. Across nine evaluations that measure output per unit of time, including studies of workers in the Philippines, China, and South Asia, generative-AI tools raised output by more than 25% on average. The average remains positive in every reported sensitivity check, although gains differ across tasks and settings. Eight comparisons of AI tutoring with human support, or with guardrails that make the tool tutor rather than simply answer, raised scores on tests taken without the tool by about a fifth of a standard deviation (SD), including in Ghana, Nigeria, and Sierra Leone. The effects differ substantially, however. The average becomes less precise when trials judged at high risk of bias are excluded, and there is little evidence on whether the gains persist after the programs

end. In Türkiye, a hint-only tutor had no detectable effect on exam scores, and an unrestricted chatbot lowered them.

Second, effects on businesses and workers vary across settings. The one business experiment in a poorer country, in Kenya, found no statistically detectable average gain, and the weakest firms lost. Danish registers rule out earnings effects larger than 2%, Finnish data show no significant change in wages or employment, and US data show higher earnings in exposed occupations. But on online platforms, demand for exposed freelance work fell by a fifth or more after ChatGPT's release. Our search found no causal study of a generative-AI tool's effect on household consumption poverty in a poorer country, although studies there do measure learning, health, and time-use outcomes. On climate shocks, it found one causal study of machine-learning flood alerts in Bihar, India, with positive AI effects on adaptation and health (Section 3.4).

Third, compared to the generative-AI tools discussed above, AI-based measurement has advanced considerably. Models trained on satellite and phone data predict where wealth is concentrated with good accuracy, and phone-data targeting beat most feasible alternatives in poorer countries like Togo. These models predict changes in wealth far less well, however, as comparisons within the same studies show most clearly, and phone-data targeting lost to a survey-based method in Bangladesh.

Taken together, these findings indicate some positive AI effects under certain conditions. They also suggest that whether poorer households benefit depends not only on what the technology can do but also on access, skills, language, competition, and institutions. The policy advice follows: evaluate applications against stated outcomes, costs, and feasible alternatives before expanding them at scale.

This review makes several contributions with direct policy relevance, connecting the evidence on AI with its applicability, its measurement, and the decisions it should inform. First, it brings together two sets of distinct but related evidence on AI applications (intervention lens) and AI-based welfare measurement (measurement lens). The first set covers productivity and earnings, inequality, learning, public services, and the environment; the second set covers poverty measurement and targeting. Previous studies typically review them separately, but they meet whenever predicted welfare is used to evaluate a program. Second, it distinguishes performance with AI assistance from outcomes for households, records where each study was conducted, and asks what the findings imply for poorer countries. Third, it offers a guide to AI indicators, a test for AI-based poverty statistics, a summary of what the evidence shows and how confident we can be in each finding, and a table of policy decisions with tentative suggestions. It also offers a short survey module, which translates the assessment into questions that governments can test.

These contributions build on recent syntheses: flagship reports on AI and development by international organizations (UNCTAD 2025; UNDP 2025; World Bank 2026), an academic review of technology, labor markets, and inequality (Winkler 2026), and two systematic reviews prepared for the World Development Report (WDR) 2026, on digital learning tools and on AI support for frontline health workers (Burneo *et al.* 2026; Sautmann and Weinert 2026).

The rest of the article is organized as follows. Section 2 sets out the framework and the measures of AI use. Section 3 reviews the studies qualitatively, domain by domain, weighing each study's design, comparator (what the tool is compared with), and setting. Section 4 pools the few comparable studies statistically and shows how much uncertainty remains. Section 5 draws out the implications for policy and research.

2. Conceptual framework and measurement of AI use

2.1 How AI reaches household welfare

Figure 2 traces the economic pathway from AI to household welfare in five stages: capability, access and adoption, task performance, market and organizational response, and household welfare. The move from each stage to the next depends on conditions listed in the figure. A second pathway concerns measurement and evaluation. Validated predictions can help measure the outcomes along the economic pathway, and when predictions decide who receives a transfer or a service, the resulting allocation can itself affect welfare. Even so, AI-based welfare estimates are a measurement tool to be validated, not a stage through which adoption turns into welfare.

Many AI models are available across countries, but access and performance depend on whether providers serve the country and on connectivity, cost, and language. Model quality can fall sharply for languages with little digital text, the languages of many of the world's poor (Joshi *et al.* 2020). Adoption, in turn, improves task performance only where skills and tool design allow. Gains depend on the user's skills, on whether a task lies inside the “jagged frontier” of what the tool does well, and on the support around it (Noy and Zhang 2023; Brynjolfsson *et al.* 2025; Dell'Acqua *et al.* 2026). A tool's design can also decide whether students learn or let the tool do the work (Bastani *et al.* 2025).

Market and organizational responses decide much of what AI does to earnings and poverty, yet the evidence on them is thinnest. Four channels matter. First, AI can raise workers' output by helping them perform existing tasks, but whether their earnings rise then depends on demand and on how the gains are shared.³ Second, AI can perform tasks outright, reducing demand for some kinds of labor. Economies that export services such as customer support may be especially

³ The studies reviewed here do not yet observe the creation of new tasks, the channel that can offset the displacement of labor by automation (Acemoglu and Restrepo 2018).

vulnerable to job losses, because their comparative advantage rested on labor costs that AI can undercut (Korinek and Stiglitz 2021). Third, AI can make expertise cheaper. Households can then gain from cheaper advice, tutoring, and diagnosis without using AI themselves, and less trained workers can perform tasks once reserved for experts (Autor 2024; World Bank 2026). Fourth, better information can change public decisions, such as who receives a transfer (Section 3.5), although a more accurate prediction is not itself evidence of a welfare gain.

Households can gain as workers and business owners, as consumers of cheaper or better services, and as recipients of public programs, but these gains need not move together. A worker whose tasks become easier may also face lower demand for those tasks, so better performance on a task done with AI is only an intermediate result. A household that never uses AI can still be affected. Subsistence farmers rarely use the tools, but the prices of their inputs and outputs, the demand for their labor, and the price of advice can all change without any adoption in the household. Earnings, prices, service quality, poverty, and access therefore need to be assessed separately, and households' share of any gains depends on prices, wages, taxes, transfers, and the costs of adjustment.

At the country level, the pathway in Figure 2 has two feedbacks, and both are conditional. Catch-up can occur if broader access and cheaper expertise let poorer countries build capability and local data advantages. Divergence can occur if computing capacity, models, data, and skills concentrate in richer economies, if the gains accrue to the owners of the technology, and if demand for tradable services falls (Alonso *et al.* 2020). Even in rich economies, the size of the aggregate gains is uncertain: projections range from a productivity gain of well under 1% over a decade to much larger gains if AI speeds the discovery of ideas (Aghion *et al.* 2019; Brynjolfsson *et al.* 2021; Acemoglu 2025).

Two caveats are in order. First, results from one setting need not carry over to another. Their relevance depends on the tasks, on workers' skills and the devices, connectivity, and organizational support available to them, and on language: performance on English-language business tasks says little about local problems in Hausa or Khmer. A tool's strengths and weaknesses must therefore be assessed where it will be used.

Second, AI risks are not negligible. Table A2 organizes the risks by the strength of the evidence behind them. They range from measured harms, such as learning losses under unrestricted use and falling demand for exposed freelance work (Bastani *et al.* 2025; Demirci *et al.* 2025), to demonstrated capabilities without measured aggregate harm and high-impact scenarios whose probabilities remain highly uncertain (International AI Safety Report 2026). Most recently, the United Nations' Independent International Scientific Panel on AI and some US industry leaders have called for stronger safeguards as AI agents advance (Capoot 2026; UN News 2026). Such agents are systems that act with little human direction. Where institutions for detecting and addressing harms are weak, these risks may be harder to manage, and poorer countries have little voice in the global rules.

2.2 How much is AI used, and by whom?

Before asking what AI does, we need to know how much it is used, and by whom. Generative AI has been adopted quickly: within two years of the release of ChatGPT, nearly 40% of working-age Americans had used a generative-AI tool (Bick *et al.* 2025), and by July 2025 about one adult in ten worldwide had used ChatGPT, with use growing fastest in poorer countries (Chatterji *et al.* 2025). The most advanced models, known as frontier models, remain concentrated in a handful of economies, however, as do the computing capacity behind them and the rules that govern them (UNCTAD 2025).

We rely on two measures of AI use. The first, diffusion, is the share of working-age people who used a generative-AI tool in a quarter, as modeled by Microsoft from its telemetry, the usage data that software providers record. The second is usage per person on Anthropic’s platform. Both measures rise steeply with national income, which accounts for about 70% of the cross-country variation in each, and the two are correlated at 0.74 (Figure 1, panels A and B; Table A1). The share of working-age people using AI rises by about five percentage points with each doubling of GDP per capita.

The absolute gap between income groups widened. Diffusion rose most, in percentage points, in richer economies (Figure 1, panel C): the gap between the high- and low-income averages grew from about 21 to 25 points, although their ratio stayed near 3.8. Low-income and lower-middle-income countries account for more than two-fifths of the working-age population but less than a third of estimated users in the covered economies (panel D). None of the four panels, however, measures a causal effect.⁴

How AI use is measured matters. A regional industry survey reports daily AI use among 81% of internet users in Viet Nam, while Microsoft’s telemetry-based estimate puts generative-AI use at 26.5% of the working-age population (Google, Temasek, and Bain 2025; Microsoft AI Economy Institute 2026). The two estimates differ in population, definition, tools, and reference period, so they are not directly comparable.

Table 1 lists the main cross-country sources on AI use and readiness. Microsoft’s telemetry covers about 147 economies, including low-income ones, with several updates a year (Microsoft

⁴A regression of growth on AI diffusion across countries produces a negative coefficient with one growth window, a coefficient near zero with another, and a smaller negative coefficient in a placebo window that closes in 2019, before the public release of ChatGPT; none of the three is an estimate of AI’s effect (Table A1). Diffusion is highest in economies that grew slowly before AI, and adoption measured in 2026 postdates the growth measured in 2023 and 2024, so these correlations cannot separate causal effects from reverse causality and shared determinants.

AI Economy Institute 2026). Provider data describe use by task, occupation, topic, and country (Anthropic 2025; Chatterji *et al.* 2025). Readiness indices cover 170 to 190 countries, exposure indices map the occupations that AI could affect, and connectivity series track internet and electricity access. Official firm surveys now include questions on AI in the European Union, the United States, and Brazil, as does Colombia’s household survey (DANE 2025; IBGE 2025).

These sources leave important gaps. Estimates modeled from telemetry can miss shared devices, a single provider’s usage data can overrepresent early adopters and English-language use, readiness is not use, and official surveys mainly cover richer countries and formal firms. Most sources cannot link use to the household characteristics needed for distributional analysis.⁵

As an attempt to fill data gaps, we propose in Table A4 a short module of questions, ready for piloting, for household and labor-force surveys in poorer countries. After a fixed definition of AI with local examples, it asks about deliberate AI use in the past twelve months and seven days. All respondents are asked about access and barriers, and users also about tools, language, tasks, and how they check the results. Linked to the rest of a survey, the module would show who uses AI by income, gender, location, and informality, which provider data cannot. It measures use rather than capability or impact, relies on self-reports, and refers to periods that differ from the quarter over which providers report use, so a comparison with telemetry needs an added item or a stated assumption.

⁵ Because the sources in Table 1 measure different things, Table A3 further sets out the questions that measures of capability, exposure, access, use, and impact answer, and where the interpretation of each stops. The first four do not establish a causal effect on welfare: exposure indices, for example, estimate the share of work tasks that AI could in principle perform or assist, which is not adoption or displacement (Gmyrek *et al.* 2023; Cazzaniga *et al.* 2024; Eloundou *et al.* 2024). The measurement agenda also extends beyond AI use: AI is increasingly used to measure poverty itself, from satellite images, phone records, and survey text, in the countries whose surveys are least frequent (Section 3.5; Table 2).

3. Existing evidence: AI as intervention and as measurement tool

This section reviews the evidence on AI as an intervention (Sections 3.1 to 3.4) and as a measurement tool (Section 3.5). For each study, we assess the setting, intervention and human support, comparator, design, outcomes, follow-up, costs, uncertainty, and provider involvement. We distinguish randomized trials and credible quasi-experiments from descriptive studies, simulations, and commentary, and review both published and working papers.

Starting from the studies known to us and from the study lists of the reviews cited earlier (UNCTAD 2025; UNDP 2025; Burneo *et al.* 2026; Sautmann and Weinert 2026; Winkler 2026; World Bank 2026), we searched for the studies citing or cited by them and for related studies. We then ran two rounds of web searches through September 2026: 55 logged queries returned 481 links, and 22 of the queries were restricted to working-paper series and repositories.⁶ Eligible interventions involve human use of a generative-AI tool released in 2020 or later, random or credible quasi-random assignment, a comparison group without the tool, and a quantitative outcome. Pooling also requires an effect size that can be converted to a common scale.

We logged each screening decision and its reason, read every candidate study in full and coded it, and verified every pooled input against its source. Tables A5 and A6 record the decisions, and Appendix A.1 details the procedure. The 77-study list comprises 56 intervention⁷ and 21 measurement studies, including 12 earlier-AI interventions retained as benchmarks or background. It includes every eligible generative-AI study in the two WDR reviews.

Table 3 counts the studies by domain and by the income group of the study population. Of the 77 studies, 7 have low-income samples and 18 high-income samples. Of the 37 studies placed

⁶The main sources were NBER, SSRN, IZA, CEPR, the World Bank, arXiv, RePEc, J-PAL and IPA, 3ie, and Stanford SCALE.

⁷ Of these 56 studies, 43 report estimates recorded in the effect-size table of Section 4, and 17 are pooled.

in a single low- or middle-income group, the largest block, 14, concerns poverty measurement rather than any intervention, and 7 are the earlier-AI health trials. Studies spanning several countries, counted separately, can also contain substantial poorer-country evidence.

3.1 Does AI raise productivity and earnings?

Two questions are easily confused. Does AI raise the output of the people who use it? And what happens to their jobs and earnings once markets respond to AI? Experiments that measure output while people use AI answer the first question. The evidence on the second is more limited and less consistently positive. This section takes the two questions in turn.

Work with AI assistance

Studies show large average gains with important exceptions (Table A7, row 1a). For professionals, assignment to ChatGPT reduced writing time by an estimated 11 minutes, about 40% of the control mean, and raised assessed quality by about 0.4 SD (Noy and Zhang 2023). In a larger preregistered experiment with UK adults who already had ChatGPT accounts, encouragement to use it cut the time spent on three work tasks by about a third (Haslberger *et al.* 2025). Online freelance programmers, most of them in India and Pakistan, completed a coding task in less than half the control group's time (Peng *et al.* 2023).⁸ In three field experiments at software firms, covering nearly 5,000 developers, completed tasks rose by about a quarter (Cui *et al.* 2026). This instrumental-variables (IV) estimate applies to developers induced to use the tool, rather than to everyone assigned access. Product-development professionals at a consumer-goods firm who were assigned the tool spent about 30% less time on a new-product task by the study's

⁸ This comparison covers only the 35 participants in each arm who finished, not all 95 assigned.

log-time estimate, or 16% less in average minutes. Individuals with AI also matched the performance of two-person teams without it (Dell’Acqua *et al.* 2025).

The gains have limits. In a consultancy experiment, consultants took about a quarter less time and produced higher-quality work on tasks within the tool’s competence, but on a task chosen to lie just outside it they were 19 percentage points less likely to reach the right answer (Dell’Acqua *et al.* 2026). Sixteen experienced open-source developers took 19% longer with early-2025 tools while believing they had been faster (Becker *et al.* 2025; METR 2026). These results suggest that the gains depend on whether a task suits the tool and whether users can tell when to rely on it.

Three features of these results may matter for poorer countries. First, in several studies the gains are largest for the least experienced workers. In the largest workplace study, of more than 5,000 customer-support agents, resolutions per hour rose by about 15% on average, most for agents in their first month and hardly at all for those with over a year’s experience (Brynjolfsson *et al.* 2025). Second, some of the evidence comes from poorer countries. About 89% of the agents in that study worked outside the United States, mainly in the Philippines, and programmers at a Chinese technology firm who were given an in-house coding assistant completed about a quarter more tasks than matched colleagues (Gambacorta *et al.* 2026). Both studies are non-randomized rollouts, however, and provide less secure causal evidence than the experiments (Section 4.1). Third, in the consultancy experiment the tool helped on some tasks and hurt on others within one firm, so how it performs on the tasks of poorer-country workers must be tested, not assumed.⁹

The one experiment that follows AI assistance through to business outcomes in a poorer country is sobering. About 640 small businesses in Kenya were offered a GPT-4 advisor over

⁹ Section 4 pools these nine estimates and returns to the uncertainty around their average (Figure 3, panel A); what each of the nine comparisons estimates, from assignment effects to effects among users after a rollout, is recorded in Table A8.

WhatsApp. The average effect on profits and revenues was not statistically distinguishable from zero, but businesses already performing well gained about 15%, and those performing poorly lost about 8% (Otis *et al.* 2026).

Jobs and earnings

The evidence on jobs and earnings comes from richer countries and global platforms, and the results differ across settings rather than pointing to a common average. In Denmark, where linked surveys and registers cover about 25,000 workers in the eleven occupations most exposed to chatbots, earnings and hours two years after the launch of ChatGPT were within 2% of what they would otherwise have been (Humlum and Vestergaard 2025a). Finnish registers show no significant change in wages or employment in exposed occupations over the same period (Kauhanen and Rouvinen 2025), and US survey data show higher earnings and a small rise in unemployment in exposed occupations (Chen *et al.* 2025).

Platform studies show a different pattern. On a global freelancing platform, demand for writing and coding work fell by about a fifth after ChatGPT, and demand for image work by about 17% after image generators (Demirci *et al.* 2025). In another platform study, postings for substitutable skills such as writing and translation fell by about a quarter relative to unaffected skills (Teutloff *et al.* 2025). Participation in such work fell even on a Russian platform with constrained access to the leading tools (Tekic *et al.* 2026). Exposed freelancers' monthly earnings fell by about 5% and their number of jobs by 2% (Hui *et al.* 2024), and translators' earnings fell by about 30% while web developers' earnings rose, possibly more so in the United States (Qiao *et al.* 2025). In US payroll data, employment of early-career workers in the most exposed occupations

has fallen relative to other workers, although the cause is not isolated (Brynjolfsson *et al.* 2026). So far, the clearest losses appear in freelance work that the tools can substitute for.

Earlier digital technologies offer reminders rather than forecasts. Phones and broadband raised growth and lowered price dispersion where they were used intensively, mobile money in Kenya lifted about 2% of households out of poverty over a decade, and fast internet raised employment and reduced poverty in Africa (Jensen 2007; Aker and Mbiti 2010; Czernich *et al.* 2011; Suri and Jack 2016; Hjort and Poulsen 2019; Bahia *et al.* 2024). These gains took years and depended on complements such as regulation, networks of mobile-money agents, and electricity (Aron 2018).

For poverty, the most important questions remain open. The studies reviewed here do not yet follow productivity gains through to wages in a poorer country, or measure whether cheaper expertise lowers the prices that poor households pay for advice, technical support, or diagnosis. The Kenyan experiment is the only study among informal firms, few labor-market studies observe outcomes two years after ChatGPT's release, and most other effects are measured within months of adoption.

3.2 Does AI narrow or widen inequality?

Within-country inequality

Two processes matter for inequality: differences in how much users gain, and differences in who adopts AI. Their combined effect remains uncertain. AI has narrowed performance gaps among users in several studies. Less experienced support agents and weaker writers gained most (Noy and Zhang 2023; Brynjolfsson *et al.* 2025). In an online experiment with 1,174 participants, the gap in task performance between more and less educated workers fell from 0.55 SD without AI to 0.14 SD with it (Cruces *et al.* 2026).

The narrowing is conditional rather than general, however, and the counterexamples are as informative (Table A9). Kenyan entrepreneurs who were already performing well gained from an AI advisor while weaker performers lost (Otis *et al.* 2026). Students with stronger baseline mathematics gained more from the Sierra Leone tutor (LearnLM Team and Fab AI 2026). A preregistered experiment in the United Kingdom found no narrowing of gaps by gender, age, education, or occupation (Haslberger *et al.* 2025). Language divides users too: AI-generated drafts improved the actionability of business emails, that is, how easily the recipient could act on them, but the gain was about half a standard deviation smaller in Arabic and Chinese than in English (Koo 2025). Where a task requires the user to direct the tool, gains concentrate among those who already have that competence, although built-in guidance narrows the spread (Idan and Anand 2026).

Three further points matter for inequality. First, these studies are consistent with one reading, although none identifies the mechanism: built-in support helps novices on tasks whose output they can check, whereas open-ended advice requires judgment and the capacity to act on it (Section 4.2 returns to this pattern). Because the studies measure different characteristics and outcomes, however, they do not establish one general relationship between skill and benefit.¹⁰

Second, unequal adoption may offset part of the narrowing of performance gaps. In Denmark, where adoption can be observed for the same occupations in register data, women were markedly less likely to use ChatGPT than men in the same jobs, and younger, higher-achieving workers adopted first (Humlum and Vestergaard 2025b). The offset operates only if those who could gain most are less likely to use AI, and it may be strongest where richer elites capture more

¹⁰Table A9 records both subgroup estimates and the contrast for each study, and its note gives the rule for pooling contrasts; a difference between a significant and an insignificant subgroup estimate is not itself evidence of a difference.

of the benefits from AI than poorer workers (Acemoglu 2024). Adoption also rises steeply with firm size: in 2025, 55% of large enterprises in the European Union used AI, against 17% of small ones (Eurostat 2025). The informal microenterprises that employ most of the poor, by contrast, appear in no adoption survey and in only one experiment (Section 3.1). The income gradient in Figure 1 is the cross-country counterpart of this unequal adoption.

Finally, exposure indices show who could be affected. Clerical work is the occupation most exposed to generative AI, and because clerical jobs are disproportionately held by women, measured exposure is higher for women in every income group (Gmyrek *et al.* 2023). Exposure is lower in poorer countries because more work is manual, agricultural, or informal, and lower still once actual digital access is taken into account (Gmyrek *et al.* 2024, 2025; Demombynes *et al.* 2025). In Viet Nam, 20.8% of employment is potentially exposed: 24.1% of women's jobs against 17.8% of men's, and 28.7% of formal against 16.9% of informal jobs (Weidenkaff and Huynh 2026). But economies that rely on exposed service exports may face job losses even if domestic adoption remains limited (Gmyrek *et al.* 2026). Occupation-based indices, moreover, capture the tasks of informal workers poorly.

Between-country inequality

The evidence cannot yet tell catch-up from divergence. The case for catch-up is that expertise becomes cheaper everywhere at once and that some applications can be reached through a phone, although their effective use still requires connectivity, skills, and support. The case for divergence is that the complements are concentrated. Only 32 countries host specialized AI computing facilities, most of them richer countries (Lehdonvirta *et al.* 2024). Vendor concentration, switching costs, and model quality in local languages favor the countries and firms

that already lead. History favors caution: earlier technologies reached poorer countries with shorter and shorter lags, yet incomes diverged because the intensity of use diverged (Comin and Mestieri 2018). In Figure 1, panel C, Microsoft’s estimates rise most, in percentage points, where use is already highest.

Regressions confirm the pattern in Figure 1: measured AI use is strongly associated with income per person. Once income is controlled for, however, internet use and electricity access add little, and the language indicators, although jointly significant, are sensitive to the specification (Table A10). These regressions cannot identify which constraints hold back adoption, because income, infrastructure, and language are closely related.

Within and between countries, the main unknowns are distributional. The studies reviewed here do not measure the wage effects of AI adoption by skill group in a poorer country, and the within-firm narrowing of gaps and unequal adoption have not been observed in the same population, so their net effect on inequality is unknown even in rich countries. A first step is to collect better data on who uses the tools, for example with the survey module proposed in Section 2.2 (Table A4).

3.3 Can AI improve learning and public services?

Learning

Education provides some of the clearest poorer-country evidence on how design affects outcomes. The promise is tutoring at scale. Better-off households spend more on private tutoring, including in poorer countries (Dang and Rogers 2008; Park *et al.* 2016), and AI could make tutoring more widely available because software can be copied at little cost. Delivery still requires devices, connectivity, and teacher time, however, and unequal access to these could instead widen

the tutoring gap. The central question is whether an AI tutor teaches the student or simply does the homework.

An experiment in Türkiye separates practice with AI from later performance without it. About a thousand high-school students practiced mathematics in four class sessions with one of two versions of GPT-4 or with none. During practice, students with an unrestricted chatbot performed 48% better, and students with a version built to give hints rather than answers, a hint-only tutor, 127% better. On the exam that closed each session, taken without AI, the unrestricted-chatbot group scored 17% below students who had practiced without AI, or 0.19 SD. The hint-only-tutor group's estimate was 0.01 SD below the control group, not statistically distinguishable from zero (Bastani *et al.* 2025). Guardrails avoided the learning loss but produced no detectable exam gain.

Three African trials of tutoring programs with human support found gains on tests taken without the tool against business-as-usual comparators, but with different degrees of confidence (Table A8). The strongest evidence comes from a trial in Sierra Leone, a low-income country. Across 48 classrooms, eight weeks with the tutor raised mathematics scores by 0.26 SD, more for students with stronger baseline scores; the trial was preregistered and run by the tool's provider (LearnLM Team and Fab AI 2026). Two other trials offer promising results for positive AI effects but carry a high risk of bias (Section 4.1).¹¹ In Ghana, students in grades 3 to 8 who received a WhatsApp mathematics tutor for eight months gained 0.36 SD (Henkel *et al.* 2024). In Nigeria, a six-week facilitated after-school program raised a composite of English, AI knowledge, and digital

¹¹ For Ghana, only 11 schools were randomized, however, and the source calculates uncertainty at the student level even though assignment was by school (Henkel *et al.* 2024). For Nigeria, the gain may reflect the whole program rather than the AI model alone, and 43% of those assigned did not take the final test (De Simone *et al.* 2025).

skills by 0.31 SD (De Simone *et al.* 2025). Four studies from richer countries complete the group pooled in Section 4 (Figure 3, panel B).¹²

Studies outside the pooled group qualify the picture. Unrestricted assistance is not always harmful: among US undergraduates who wrote an essay with it, it raised scores on an unaided test one week later by 0.27 SD (Contractor and Reyes 2026). Guidance on learning-oriented use raised course grades in Chile by 0.22 SD over access alone, while encouragement did not (Gallegos 2026), and pedagogical prompts added nothing detectable over standard ChatGPT in a German school trial (Fütterer *et al.* 2026). In a large experiment that also randomized a mastery requirement, AI tutoring's effect on a delayed test without the requirement was small and imprecise. Because the source does not report the outcome's standard deviation, the study enters only a sensitivity analysis (Oreopoulos *et al.* 2026).¹³

Several cautions apply. First, the studies test learning without the tool but say little about whether gains persist well beyond the program. None of the eight pooled outcomes measures retention after the program ends. Second, established alternatives set a demanding cost benchmark: phone-based tutoring in Botswana raised learning by 0.12 SD at very low cost (Angrist *et al.* 2022), and personalized computer-assisted learning in India, before generative AI, raised mathematics by 0.37 SD (Muralidharan *et al.* 2019). A review of education technology in developing countries

¹² These include a two-lesson university physics crossover in the United States (Kestin *et al.* 2025); a German laboratory study that tested students after a 25-minute session with a tutor grounded in the course text, that is, one that drew its answers from that text (Fischer *et al.* 2025); eight weeks of GPT-4 homework sessions in four Italian classes, which raised English scores by an imprecise 0.25 SD (Vanzo *et al.* 2025); and a two-year trial across 18 Tennessee middle schools, in which Khan Academy practice with the Khanmigo tutor raised mathematics by an estimated 0.07 SD per school year over usual remediation, not statistically distinguishable from zero, with low median use (Oreopoulos and Low 2026).

¹³ The pooled learning estimate is above the 0.125 SD that the World Development Report 2026 background review finds for digital learning tools as a whole, and above the median effect of education interventions (Kraft 2020; Burneo *et al.* 2026), but the two pools differ in populations, outcomes, comparators, durations, inclusion rules, and treatment packages. That review excluded the Nigerian trial for attrition above 30%, and within its own sample it finds no detectable advantage of generative over earlier tools (Table A11, whose panel B matches the two study lists). Without the Nigerian trial, our pool yields 0.18 SD, closer to that benchmark (Table A12). Without the physics crossover and the Italian study, which do not report whether AI was unavailable during testing, it yields 0.17 SD.

locates such gains in software adapted to the level of the student rather than in hardware (Rodriguez-Segura 2022). Third, adaptivity alone is no guarantee. In 315 rural Chinese classes, an adaptive practice system built on pre-generative machine learning did no better than a non-adaptive version, or than no computer-assisted learning at all (-0.04 SD after 11 weeks; Ma *et al.* 2026).

In short, selected AI tutoring programs with human support or guardrails can improve short-run learning measured without the tool, and unrestricted assistance can lower it. But the trials do not show that generative AI generally improves learning or that it outperforms earlier educational technologies. The remaining policy questions are whether gains persist in routine schooling and whether full program costs compare favorably with feasible alternatives (Section 5).

Public services

We cover health and other public services only as interventions, and in health the evidence on patient outcomes remains thin. The most rigorous study is a cluster-randomized trial across 16 primary care facilities in Kenya, in which 103 clinical officers were randomized to a decision-support tool built on a language model or to usual care. The tool improved the quality of clinical documentation. Treatment failure within 14 days was rare in both groups, occurring in 2.2% of visits with the tool and 2.0% without it. In the trial's primary analysis, which allows for differences between clinical officers and facilities, the odds ratio was 0.77, with a confidence interval from 0.55 to 1.08. This estimate neither establishes a patient benefit nor rules one out (Agweyu *et al.* 2026). An earlier before-and-after comparison in the same clinic chain reported 16% fewer diagnostic errors, a different outcome (Korom *et al.* 2025).

In Pakistan, physicians trained to use GPT-4o scored 71% on diagnostic reasoning against 43% for physicians using conventional online resources (Qazi *et al.* 2026). A randomized comparison across settings adds a rare within-study contrast: on the same simulated cases, a general-purpose model raised physicians’ performance by 18 percentage points in Kenya, 11 in Indonesia, and 7 in the Netherlands, against controls without internet or guideline access (Rounding *et al.* 2026). The gain was largest in Kenya, the poorest of the three settings. Yet these trials show better reasoning on cases with AI assistance rather than better patient outcomes.

Building on the list in the systematic review for the WDR 2026 (Sautmann and Weinert 2026), we count its seven trials in poorer countries as benchmarks from earlier AI, as we do in education (Table 3; Table A5, rows 53 to 59).¹⁴ Most of these trials show encouraging gains in screening, referral, detection, or throughput. AI-based eye screening raised clinic throughput in Bangladesh and referral uptake in Rwanda, an AI-assisted slit lamp raised village doctors’ cataract referrals in China, and AI-guided screening doubled the detection of cardiomyopathy among pregnant women in Nigeria (Mathenge *et al.* 2022; Abramoff *et al.* 2023; Adedinsewo *et al.* 2024; Yu *et al.* 2025). Two other Chinese trials cut diagnosis time, at some cost in accuracy, and reduced mis-triage in an emergency department (Lin *et al.* 2019; Liu *et al.* 2021). In Malawi, computer-aided X-ray screening for tuberculosis cut the time to treatment but, at US\$4,620 per quality-adjusted life year, was not cost-effective; it is the only cost-effectiveness analysis among these trials (MacPherson *et al.* 2021).

Outside health, the largest randomized evaluation of a generative tool in a poorer-country public service is in Pakistan’s courts (Mehmood *et al.* 2026). The study randomized 1,559 judges, about half of the country’s trial judges, to a GPT-4 assistant grounded in Pakistani case law, with

¹⁴ That review found 17 causal studies of AI support for frontline health workers in low-resource settings, but only one using a generative AI model (Sautmann and Weinert 2026).

targeted or generic training, or to generic training alone. Districts with more treated judges resolved more cases, 6.3% more at the median share of treated judges, and appeals per resolved case, a rough check on the quality of decisions, did not rise. The study puts the tool's annual infrastructure cost for eight judges, excluding development and maintenance, at about one-twentieth of one judge's salary.

Evaluations in health and other public services should measure patient or service outcomes, assess exclusion and harms such as barriers for people without phones and risks to privacy, and check whether people can obtain an explanation and appeal a decision. Section 5 and Table 4 return to these safeguards.

3.4 AI and the environment

AI may help poorer countries respond to climate-related hazards, which fall hardest on the poor and widen inequality (Dang *et al.* 2026a; Triyana *et al.* 2026), but it also uses electricity and water. The clearest application is forecasting, which relies on predictive machine learning rather than generative AI and so enters as background evidence. A global machine-learning model now produces flood forecasts for river basins without stream gauges, which include most basins in poorer countries, and its five-day forecasts are about as reliable as the same-day forecasts of the best earlier system (Nearing *et al.* 2024).

Better forecasts help households only if warnings reach them in time. Across 319 communities in Bihar, delivering the model's forecasts through trained local volunteers raised the number of flood alerts that households received several-fold. In the communities that flooded severely, the alerts improved indices of adaptation and health by 0.18 SD and cut medical spending by about 30%, with no detectable effect on economic well-being (Jagnani and Pande 2024).

Beyond forecasting, most climate applications of machine learning were built for richer countries' problems, and their use where sensors are sparse and crops differ has to be evaluated rather than assumed (Rolnick *et al.* 2022). Text-message extension offers a comparison: advice sent to East African farmers modestly raised the take-up of recommended inputs at a few dollars per farmer (Fabregas *et al.* 2025). Our search found no published randomized evaluation of a generative-AI adviser for farmers in local languages (Section 5).

Data centers are a small but fast-growing share of global electricity demand (de Vries 2023; IEA 2026). In a poorer country they compete with households and firms for an often unreliable grid and for scarce water, and the public spending that attracts them has an opportunity cost in schools and clinics, so proposals should be assessed by their full lifecycle costs rather than treated as a sign of readiness.

3.5 AI and poverty measurement in data-scarce contexts

Poverty measurement

The second lens asks how well AI measures poverty. The motivation is the survey gap described in the introduction: poverty is least observed where it is most prevalent (World Bank 2024). AI-based estimates could help update statistics between surveys, describe areas smaller than a survey can reliably represent, and reduce the cost of selecting transfer recipients. Table 2 states, for each use, what must be predicted, the comparison it must beat, and the main risk to check.

AI-based poverty measurement, mostly in small-area estimation contexts, predicts welfare rather than observing it. Models learn relationships between survey outcomes, such as consumption or wealth, and household, satellite, geospatial, or phone-data predictors. Applying these relationships to the same predictors elsewhere estimates outcomes in areas the survey did

not reach, or at finer geographic levels where it was not representative. Accuracy is often judged on held-out units, that is, units not used to fit the model, and is reported as the share of the variation in the survey measure that the predictions reproduce (R^2). The product is a map or a statistic (Table 2). Targeting adds a decision rule: units are ranked by predicted welfare, the budget fixes how many can be served, and the test becomes the share of the eligible poor missed and of the non-poor included, against a survey benchmark and against feasible alternatives at the same budget and cost.

The strongest validation evidence comes from satellite imagery and phone records, and almost all of it comes from poorer countries. Daytime satellite images explained 55% to 75% of the local variation in asset wealth across five African countries, and a later model explained 70% of village-level variation across the continent (Jean *et al.* 2016; Yeh *et al.* 2020). Phone metadata predicted wealth in Rwanda, phone and satellite data together mapped poverty in Bangladesh and Senegal, and high-resolution imagery predicted small-area consumption in Sri Lanka (Blumenstock *et al.* 2015; Pokhriyal and Jacques 2017; Steele *et al.* 2017; Engstrom *et al.* 2022).

Accuracy depends on what is being predicted. In Bangladesh, phone-and-satellite models explained 76% of the variation in an asset-wealth index but only about a quarter of the variation in income or in a poverty-probability index (Steele *et al.* 2017). Figure 4 shows this and other results for 10 of the 17 validation studies listed in Table A6, selected by a stated rule. Figure 4 also shows where progress is still needed for monitoring. The reviewed models generally predict wealth levels more accurately than changes, most clearly in paired comparisons within studies that hold the model, unit, and validation design fixed. In a benchmark across 59 countries, features that explained more than half of the variation in wealth levels explained almost none of the variation in changes between survey rounds (Marty and Duhaut 2024), and the continental model that

explained 70% of the variation in village wealth explained 35% of the variation in village-level changes (Yeh *et al.* 2020).

Three caveats apply. First, an area-level asset wealth index is not household consumption poverty. Second, predicting area averages is not ranking households. Third, predicting levels in an unseen year is not predicting changes between years.

A further question is whether machine learning improves on established “poverty-mapping” methods (i.e., imputing from a survey into a census; see Elbers *et al.* 2003), including in measuring uncertainty. The evidence is mixed. In Malawi, a conventional spatial regression on survey and remote-sensing data reproduced a census-based poverty map closely but diverged in some areas (van der Weide *et al.* 2024). In Mexico, machine learning on remotely sensed data did worse than conventional small-area estimation and was more biased in the poorest areas (Corral *et al.* 2025). Across seven countries with census ground truth, by contrast, machine learning predicted small-area asset wealth from satellite indicators somewhat more accurately than the standard small-area model (correlations of 0.87 to 0.89 against 0.83); with improved residual bootstrap procedures, it gave 95% intervals with better coverage, although such intervals are rarely reported with a map (Merfeld *et al.* 2025).

However, when poverty is imputed from one survey to another survey, machine learning does not generally beat well-specified existing imputation methods and may do worse once uncertainty is accounted for (Lucchetti *et al.* 2025; Verme 2025; Dang *et al.* 2026b). An AI-based estimator should therefore be tested against such a method on the relevant data, with its uncertainty reported.

Language models can also support statistical work. They can code open-ended survey responses, process administrative text, and translate questionnaires into languages with little digital

text, where their quality is weakest and checking matters most (Joshi *et al.* 2020; Korinek 2023). They cannot replace respondents: synthetic answers from a language model are not data about a population (Bisbee *et al.* 2024).

Poverty targeting

Targeting offers the clearest test of whether AI improves a policy decision, because there is always a feasible alternative to beat: can phone-data targeting reach more poor people than that alternative at the same budget? In Togo, the rural expansion of the Novissi cash transfers during the COVID-19 pandemic was evaluated with a 2020 phone survey of mobile subscribers in the 100 poorest cantons. Welfare was measured with a proxy means test, a score predicted from household characteristics that a short survey can observe, and the poorest 29% were eligible. At the program's budget, phone-data targeting left 53% of the eligible poor unreached, against 59% to 78% for the feasible geographic and occupational alternatives.

A separate nationwide simulation used the 2018 to 2019 household survey and measured welfare by consumption. Phone-based targeting missed 50% of the eligible poor, compared with 52% to 76% under most feasible alternatives. An optimized rule allocating all transfers to agricultural workers did slightly better, missing 48%, and methods requiring a comprehensive social registry, which did not exist, would have done better still (Aiken *et al.* 2022; Figure 4, panel B).

Results elsewhere reinforce the importance of the population and the alternative. In Afghanistan, the value of phone data depended on who was missing from the phone network and on what the operator would share (Aiken *et al.* 2023). In Bangladesh, a conventional proxy means test built from a survey beat both phone-based and community-based targeting outright, reaching

52% of the poor at the program budget against 32% for the phone-based method (Aiken *et al.* 2025b).

Accuracy alone does not determine the best targeting method, and the benchmark is the best feasible alternative, not the absence of data. A survey-based proxy means test costs more to collect and update than phone data. At a fixed total budget, those costs leave less for transfers and can offset its advantage in identifying poor households, a trade-off the Bangladesh study reports. Comparisons should therefore count data and administrative costs and define each method's errors over the same population, whether eligible households, phone owners, or applicants.¹⁵

Overall, the validation record for satellite and phone data is encouraging. The main open questions are whether accuracy established for one population, outcome, and period carries over to others, especially when the aim is to measure changes in consumption, and how models drift, that is, how their accuracy declines as landscapes and phone markets change. In Sri Lanka, a model built on phone and satellite data identified 83% of the poorest quarter of local divisions, as ranked by a census asset index. When each province was held out in turn, however, this share ranged from 46% in the Northern Province to 94% in the Western Province (Algama *et al.* 2026). These questions matter most when predictions serve as evaluation outcomes. Survey observations are then needed to test whether the predictions capture the relevant change and to support valid inference (Table A13). Evaluations that measure consumption directly do not require predicted outcomes.¹⁶

¹⁵ Even careful validation faces data limitations: the surveys used as ground truth are infrequent, survey coordinates are displaced for confidentiality, and uncertainty is rarely reported alongside a map. An analogy from health care shows the risk of predicting the wrong concept. An algorithm that predicted health spending in order to allocate care systematically underestimated the needs of Black patients, because spending measured access rather than need (Obermeyer *et al.* 2019). A model that predicts consumption from phone use could likewise mismeasure the welfare of people who use phones differently; whether it does is a question to test, not a known bias.

¹⁶ Some evaluations already use predicted welfare as outcome data: satellite imagery for electrification in Uganda and for cash transfers in Kenya, and phone data for transfers in Togo (Huang *et al.* 2021; Ratledge *et al.* 2022; Aiken *et al.* 2025a). In Haiti, however, phone-based predictions were too imprecise to detect a transfer's effect that surveys

Predictions can also be combined with a representative sample of observed outcomes. One recent method, prediction-powered inference, uses the observed data to correct the average prediction error. Its precision gains depend on the sample sizes, the prediction errors, and the design, not on R^2 alone (Angelopoulos *et al.* 2023). Box 1 applies both lenses to one program, in Togo; neither of its two impact estimates shows the welfare gain from choosing AI-based targeting over another method.

4. Evidence synthesis

4.1 Which studies are pooled, and how

Section 3 examined individual applications. This section asks whether the findings support general conclusions. We pool two outcomes, task performance with AI assistance and learning measured without the tool, and compare the business, labor-market, and validation findings without pooling them (Table A7).

The exercise is exploratory. It rests on the study list described in Section 3 and recorded in Tables A5 and A6.¹⁷ It has four parts, detailed in Appendix B.1. First, we pool the estimates in each domain with a random-effects model, which allows the true effect to differ across studies, and report a small-sample (modified Hartung–Knapp) interval. Second, we compare the estimates descriptively by user skill, task type, and whether the tool came with support or guardrails, and

detected (Barriga-Cabanillas *et al.* 2025). Because prediction errors can differ between program and comparison groups, such evaluations need a representative survey sample of both groups to test and correct the predictions (Angelopoulos *et al.* 2023).

¹⁷Two systematic reviews prepared for the World Development Report 2026 were among the starting points of the search (Section 3) and cover the sectors closest to this one. Burneo *et al.* (2026) pool 14 randomized trials of digital learning tools from ten economies, none low-income; the education studies are re-pooled here on this review’s criteria (Section 4.2), and the two lists are reconciled study by study (Table A11, panel B). Sautmann and Weinert (2026) synthesize the frontline-health studies narratively; their seven poorer-country trials are counted here as benchmarks from earlier AI (Section 3.3; Table 3), not re-pooled. Other meta-analyses pool human–AI collaboration and programming studies without distinguishing countries by income, or ChatGPT-in-education studies with weak screening on design (Table A11).

arrange the individual estimates by sample geography (Figure A1). Third, we report within-study differences by experience, education, baseline performance, and language (Table A9). Fourth, we run sensitivity analyses (Tables A12 and A14).

Risk-of-bias assessment offers a standard practice that systematic reviews employ to grade the quality of the studies they pool. We apply the Cochrane tools: RoB 2 to the randomized trials and ROBINS-I to the two non-randomized workplace rollouts (Sterne *et al.* 2016, 2019). Each study is judged on five sources of bias: randomization or the control of confounding, deviations from the intended intervention, missing outcome data, outcome measurement, and selection of the reported result. Each then receives an overall rating (Table A14).¹⁸ Following standard guidance, the ratings are not used as weights (Higgins *et al.* 2024). Instead, they identify the studies on which each result rests, and we recompute the pooled estimates without those at high or serious risk of bias.

Business and labor-market findings are compared rather than pooled, for three reasons (Appendix A.1). They concern different outcomes and interventions; none of their subgroups meets the review’s rule of at least five independent estimates of the same outcome and comparison; and several use ChatGPT’s release as a common source of variation.

4.2 What the pooled estimates show

Across nine evaluations, AI assistance raised output per unit of time by an average of 29% (small-sample 95% interval: 10% to 51%; Figure 3, panel A), equivalent to doing the same work in about 22% less time. The average stays positive, with an interval above zero, when any one study is dropped. Randomized comparisons give 33% (6% to 67%), and excluding studies at high or serious risk of bias gives 28% (5% to 56%) (Table A12). Eight evaluations found gains; the one

¹⁸ One author made the ratings with AI assistance, and the other author reviewed them independently.

loss was among experienced developers working on their own software projects. The size varies widely ($I^2 = 90\%$), so the pooled estimate supports substantial gains in the evaluated applications without establishing a return that other workplaces should expect.¹⁹

On average, AI tutoring with human support or guardrails also improved learning measured without the tool. The eight comparisons give a gain of 0.21 SD (95% interval: 0.05 to 0.36; Figure 3, panel B); seven estimates are positive, and four of them have 95% intervals above zero. This is encouraging evidence that some supported programs improve learning, including in poorer-country schools. The learning result is less robust than the task result, however. Excluding the Ghana and Nigeria trials rated at high risk of bias gives 0.16 SD (-0.04 to 0.37), and the interval also includes zero without Nigeria alone, for school trials alone, and with reported variances only (Table A12). Effects differ substantially ($I^2 = 76\%$). The comparisons cannot identify a generally superior design, because populations, delivery, assessments, and duration vary together. The unrestricted Türkiye arm is kept outside this pool (Section 3.3).

The ten business and labor-market results, compared rather than pooled, include falls of a fifth or more in demand for exposed platform work. They also show no detectable average gain for Kenyan firms, no significant change in Finland, and earnings effects within 2% in Denmark, but they show higher earnings in exposed US occupations (Table A8). None measures effects on household consumption poverty in a poorer country.

Validation results are compared, not averaged (Table 2; Figure 4). Across five multi-country studies of asset wealth, predictions explain roughly half to four-fifths of the variation in levels out of sample, an encouraging record for mapping, but at most half of the variation in changes and about 4% for cluster-level changes in a 59-country benchmark.

¹⁹ The prediction interval for a comparable new setting runs from -7% to $+79\%$, but it applies only if the included studies describe a population of such settings, which is unlikely for task performance (Appendix B).

Poorer-country evidence is concentrated in poverty measurement, health, and education, and in Africa (Table 3; Figure A2); in health it comes largely from the earlier-AI screening trials. Of the 37 studies whose population is in a single poorer country, 17 are in Sub-Saharan Africa, 9 in South Asia, 6 in East Asia and the Pacific, 4 in Latin America and the Caribbean, and 1 in Türkiye. Three task-performance estimates and four of the eight learning comparisons come mainly from poorer countries. The trial in Sierra Leone, run by the tool’s developer, is the only causal study of a generative-AI tool in a low-income country under the intervention lens.

Some caution is needed in interpreting these results, for three reasons. First, this is a structured critical review, not a systematic review of a defined literature, so its averages describe the selected studies. Second, three pooled variances rest on stated assumptions; Table A12 varies them and omits the affected studies. Third, the search is not exhaustive, and there appear to be too few studies to assess publication bias or the sources of heterogeneity reliably.

Subject to these caveats, the results in Sections 3 and 4 identify promising AI uses: assistance on tasks where performance can be checked, and tutoring programs evaluated on learning without the tool. The studies suggest that task, user, and program design matter, but do not establish which features explain the differences.²⁰ Policy should build on the demonstrated gains while testing whether they carry over to other settings, whether they last, and what they cost to achieve. Table 5 summarizes the evidence and our confidence in each finding.

²⁰ Table A9 reports ten within-study contrasts with differing subgroup definitions and statistical support. Some favor less experienced or less able users; others favor more skilled users or vary by language. Gambacorta *et al.* (2026) report significance for juniors only, without a junior–senior difference test. These comparisons suggest hypotheses about design and user support rather than establish a general mechanism.

5. Policy implications and conclusions

This section draws out the implications for five decisions that governments and employers in poorer countries face (Table 4). Although the evidence base is thin, our early assessment offers policy relevance because policy is being made ahead of the evidence.²¹ The evidence identifies promising applications and the conditions for expanding them. Agencies can prioritize evaluated approaches to work with AI assistance, tutoring, and small-area poverty measurement, while judging each against stated outcomes, full costs, equity objectives, and the best feasible alternative. The task gains are the most robust pooled finding. The learning gains justify further evaluated use, while there is limited evidence on effects on earnings and poverty reduction.

The first decision is whether to deploy AI in schools. Among public-service applications, education offers some of the clearest direct welfare evidence. The tutoring trials in Ghana, Nigeria, and Sierra Leone suggest that well-supported programs can raise learning measured without the tool, which justifies further pilots. Because the Türkiye trial shows that performance with AI assistance can mislead, and no single design has been shown to work best, ministries should test designs that support students' own problem solving against feasible alternatives (Section 3.3). A program should be expanded only if learning measured without the tool persists, at the end of the year and later, and the program meets its learning, cost, and equity objectives better than the best feasible alternative, whether human tutoring, structured pedagogy, or existing adaptive software. Costs should cover the whole program, from devices and teacher time to subscriptions and maintenance, with research costs kept separate. Wider reach among underserved students can

²¹ Forty of the 44 generative-AI intervention studies reviewed here were published or last revised in 2025 or 2026, 20 of them in the first nine months of 2026. Section 4.2 separates what is established from what is exploratory and identifies the further evidence needed, and the study list and code are built to be updated as new trials report, as in a “living” review (Elliott *et al.* 2017).

justify a higher cost per unit of learning if the trade-off is explicit. Plans should also address teacher preparation and local-language content.

The second decision belongs to statistics offices and social-protection agencies: when to adopt AI-based poverty estimates and targeting. These methods have already mapped poverty in many poorer countries and, in Togo, helped select transfer recipients in an emergency. The two uses face different tests (Table 2). Statistics offices need estimates of the intended welfare quantity for the stated population and period, with credible uncertainty and, for monitoring, validated changes. Social-protection agencies should compare methods against stated coverage and targeting objectives at the same total budget, including data and administrative costs, report both inclusion and exclusion errors, and provide a route to review and appeal. Because existing methods predict levels more reliably than changes, their performance should be rechecked with each new survey round. But further research on how AI methods compare against traditional survey-to-survey imputation or poverty-mapping methods is necessary.

The third decision concerns AI at work and worker transitions. Employers and public agencies can test assistance on tasks where completion time and quality can be checked, assessing both against full costs before expansion. Productivity gains need not become earnings gains: exposed platform work has contracted, while national studies report differing earnings results (Section 3.1). Our search found no evaluated transition program for AI-displaced workers, so lasting job placements, measured against a credible alternative, remain a proposed criterion. A panel in a service-exporting economy could track adoption, tasks, hours, earnings, and job entry and exit over two years; causal interpretation would require credible variation in adoption or exposure.

The fourth decision, whether to deploy AI in other public services, calls for pilots rather than programs. AI tools have improved physicians' diagnostic reasoning on simulated cases in several countries. But the only generative-AI trial with patient outcomes, in Kenyan private clinics, neither shows nor rules out a benefit. Our search found no eligible impact study of a generative tool in a poorer-country public health system, and the earlier-AI screening trials improved detection and referral, rarely patient outcomes (Section 3.3). Outside health and education, the one randomized evaluation we know of, in Pakistan's courts, found more cases resolved where trained judges used a generative assistant (Section 3.3). A pilot should compare AI with the best feasible non-AI service model, measure patient or service outcomes and exclusion rather than documentation alone, and scale up only if net gains persist at an affordable cost and exclusion stays below a threshold fixed in advance.

The fifth decision, on data governance, competition, and infrastructure, concerns conditions that cut across applications. The evidence does not establish that domestic capacity to build frontier AI models is needed to obtain the gains reviewed here, or that purchasing applications is always best. Governments should compare purchasing applications, adapting existing models, sharing computing capacity, and training models domestically by purpose, scale, expected use, and full lifecycle cost. Contracts can require data portability and audit access, and regional procurement can lower fixed costs.

Three principles apply to all five decisions. First, deployment designs and evaluation standards should be adapted to local conditions and context-specific nuances rather than imported with the tool. Second, every high-stakes use needs independent outcome evaluation, human review and appeal, and explicit measures of who benefits and who is excluded. Third, review dates should

be set in advance, with follow-up long enough for slower labor-market and poverty effects, which for mobile money took years to appear (Suri and Jack 2016; Aron 2018).

The research agenda follows from these gaps. Promising directions for research include randomized evaluation of AI program effects in local languages and public services. Measurement ideas include a reconciliation of adoption estimates across telemetry and provider data, which can be validated or supplemented with a multi-country household survey.

In summary, the evidence reviewed here shows that selected AI applications improve learning, an important component of household welfare, and that AI raises output on a range of tested tasks performed with its assistance. Evidence that these gains lead to higher earnings or lower consumption poverty remains limited, and whether they do depends on costs, access, and the responses of workers, firms, and public services. AI is spreading faster than the evidence on its effects, and faster still than the data that would let poorer countries see those effects in their own populations. Better data and carefully evaluated pilots can proceed together, starting with more data collection efforts (e.g., piloting the proposed survey modules in Table A4). Governments can begin with clear objectives and feasible alternatives, and expand the applications that meet those objectives in the settings where they will be used.

Table 1. Cross-country sources on AI adoption, use, and readiness

Source	What it measures	Coverage	Some limitations for poorer countries
1. Microsoft AI Economy Institute, AI Diffusion Report	Share of working-age people using AI tools (telemetry, modeled)	About 147 economies; periodic releases	A modeled estimate, not a census; misses shared devices
2. Anthropic Economic Index	Use by task and occupation, one provider	Global where the product is available	One provider; early adopters; English
3. OpenAI usage data	Use by topic, geography, and demographics	Consumer accounts, one provider	Excludes enterprise and API use
4. Oxford Insights, Government AI Readiness Index	Government readiness scores	About 190 countries; annual	Readiness is not adoption
5. IMF AI Preparedness Index	Economy-wide preparedness	About 170 economies	Preparedness is not use
6. Eurostat ICT survey; US Census BTOS; IBGE PINTEC (Brazil); DANE ENTIC (Colombia)	Use of AI in official firm surveys, and in one household ICT survey	European Union; United States; Brazil (industrial firms with 100 or more workers); Colombia (internet users aged five and older)	Mostly richer countries; no informal firms; templates for the module in Table A4
7. e-Conomy SEA report	Self-reported use among internet users	Southeast Asia; annual	Different population and definitions from telemetry
8. ILO and World Bank exposure indices	Task-based exposure of occupations	Global; poorer-country variants	Exposure is not adoption
9. ITU and WDI connectivity series	Internet, broadband, and electricity access	Global; annual	Access is not use

Source: Row 1, Microsoft AI Economy Institute (2026) and Misra *et al.* (2025); row 2, Anthropic (2025, 2026); row 3, Chatterji *et al.* (2025); row 4, Oxford Insights (2025); row 5, Cazzaniga *et al.* (2024); row 6, DANE (2025), Eurostat (2025), IBGE (2025), and US Census Bureau survey documentation; row 7, Google, Temasek, and Bain (2025); row 8, Gmyrek *et al.* (2023, 2025, 2026) and Demombynes *et al.* (2025); row 9, ITU and World Development Indicators.

Table 2. Some validation tests for AI-based poverty estimates

Use	What must be predicted	Compare against	Test	Main risk to check
1. Small-area estimation (“poverty mapping”) for geographic prioritization	Relative wealth of areas or households	Existing census- or survey-based maps	Validation in areas excluded from model fitting; error by subgroup	Drift over time; validation data not independent of model construction
2. Program targeting	Ranking of households at a fixed budget	Feasible registries and geographic targeting	Exclusion and inclusion errors at the budget	Households without phones cannot be scored or paid; people may change phone use to qualify
3. Welfare statistics, poverty monitoring between surveys	Consumption or income levels, poverty indicators	Direct survey estimates or survey-to-survey imputation	Out-of-sample by region and time; uncertainty intervals; validation of predicted changes against later survey rounds; tracking known shocks	The predicted quantity does not match the intended welfare outcome; declining model accuracy mistaken for a poverty trend

Source: Authors’ elaboration based on studies discussed in the paper.

Note: Accurate predictions can support these uses, but predictive accuracy is not welfare impact. Each use is judged against the feasible alternative rather than against no data; for targeting, the test compares accuracy and cost over the same population and budget definition (Section 3.5). Row numbers are used in Figure 4 and Table A6.

Table 3. Number of distinct studies by domain and income group of the study population

Domain	Low income	Lower middle	Upper middle	High income	Multiple countries or geography unspecified	Total
1. Work: tasks with AI assistance	0	0	3	3	8	14
2. Work: jobs, earnings, demand	0	1	0	5	5	11
3. Education	1	3	3	10	0	17
4. Health and public services	2	6	3	0	1	12
5. Agriculture and environment	0	1	0	0	1	2
6. Poverty measurement and targeting	4	8	2	0	7	21
Total	7	19	11	18	22	77

Source: Authors' compilation from the lists of candidate studies (Tables A5 and A6) and the flood-forecasting study of Section 3.4; Figure A2 shows the same counts as an evidence map.

Note: Income groups follow the World Bank FY2027 classification of the study population, not of the authors' institutions; Togo, for example, moved from low to lower-middle income in that classification. Each study was read in full and is counted once, by outcome domain and by the geography of its participants. Multi-country studies, global platforms, and samples whose participants' countries are not stated are in the fifth column. Studies are distinct but not independent: several labor-market studies use the same technological shock (Section 4.1). Row 3 includes three studies of pre-generative AI tools retained as benchmarks (Muralidharan *et al.* 2019; Ferman *et al.* 2021; Ma *et al.* 2026), and row 4 the seven earlier-AI trials in poorer countries discussed in Section 3.3; a scripted-chatbot study compared with a human reading partner is out of scope (Table A5, row 44).

Table 4. Policy decisions, conditions, and criteria for piloting and expansion

Decision	Who decides	What the evidence says	Conditions	Criteria for expansion or further action (fixed before evaluation)
1. Deploying AI in schools	Education ministries	- Selected tutoring programs, not the model alone, raised learning measured without the tool by about 0.2 SD on average; persistence is unknown - The three poorer-country gains differ in precision and attrition, and an unrestricted chatbot lowered scores in Türkiye	Devices; teacher training; local-language content	- Meets the agency's learning, cost, and equity objectives against the best feasible alternative - Counts the cost of the whole program - Learning is measured without the tool, on a comparable outcome and horizon
2. Adopting AI-based poverty statistics and targeting	Statistics offices; social-protection agencies	- Models predict wealth levels well, changes less reliably - Phone-data targeting beat most feasible alternatives in Togo but lost to a survey-based proxy means test in Bangladesh	Statistics: recent ground truth and documented uncertainty. Targeting: reported inclusion and exclusion errors; review and appeal	- Statistics: validated for the stated population, outcome, and period, with credible uncertainty; changes validated before use for monitoring - Targeting: meets stated coverage and targeting objectives better than the feasible alternative at the same total budget, data and administration included - Both: revalidated on each survey round
3. AI at work and worker transitions	Labor ministries; employers	- AI raises output on tested tasks; earnings effects differ across platforms and countries - No evaluated transition program for AI-displaced workers	Verifiable task quality and time; full costs; employer demand; a worker panel	- Task gains justify full costs; transition support improves lasting job placements against a credible alternative - Monitoring names its source of variation in exposure in advance
4. Deploying AI in other public services	Service ministries	- One Kenyan private-clinic trial improved clinical documentation; its patient-outcome estimate neither shows nor rules out a benefit - A randomized rollout in Pakistan's courts raised cases resolved; seven earlier-AI screening trials improved detection and referral, rarely patient outcomes; no eligible study of a generative tool in a poorer-country public health system	Safeguards intended to reduce and monitor risks: trained staff; secure data; review and appeal	- Net service gains over the best feasible non-AI service model persist - The cost stays affordable - Exclusion stays below a threshold set in advance
5. Data governance, competition, and infrastructure	Finance and sector ministries	- Supply is concentrated - The reviewed evidence establishes neither a need for domestic capacity to build frontier AI models nor that procurement is best everywhere	Portability and audit access in procurement; energy and water planning at full lifecycle cost	- Procurement, adaptation, shared compute, and domestic investment are compared by purpose, scale, expected use, and full lifecycle cost - No universal rule

Source: Sections 3 and 4; the studies in Tables A5, A6, and A8; the pooled results in Table A12.

Note: Agencies should fix thresholds and review dates before evaluation. Rows 1 and 2 draw on evaluated applications; rows 3 and 4 propose pilots and monitoring; row 5 states institutional principles. Row 2 concerns predictive machine learning; row 4 includes generative and earlier AI.

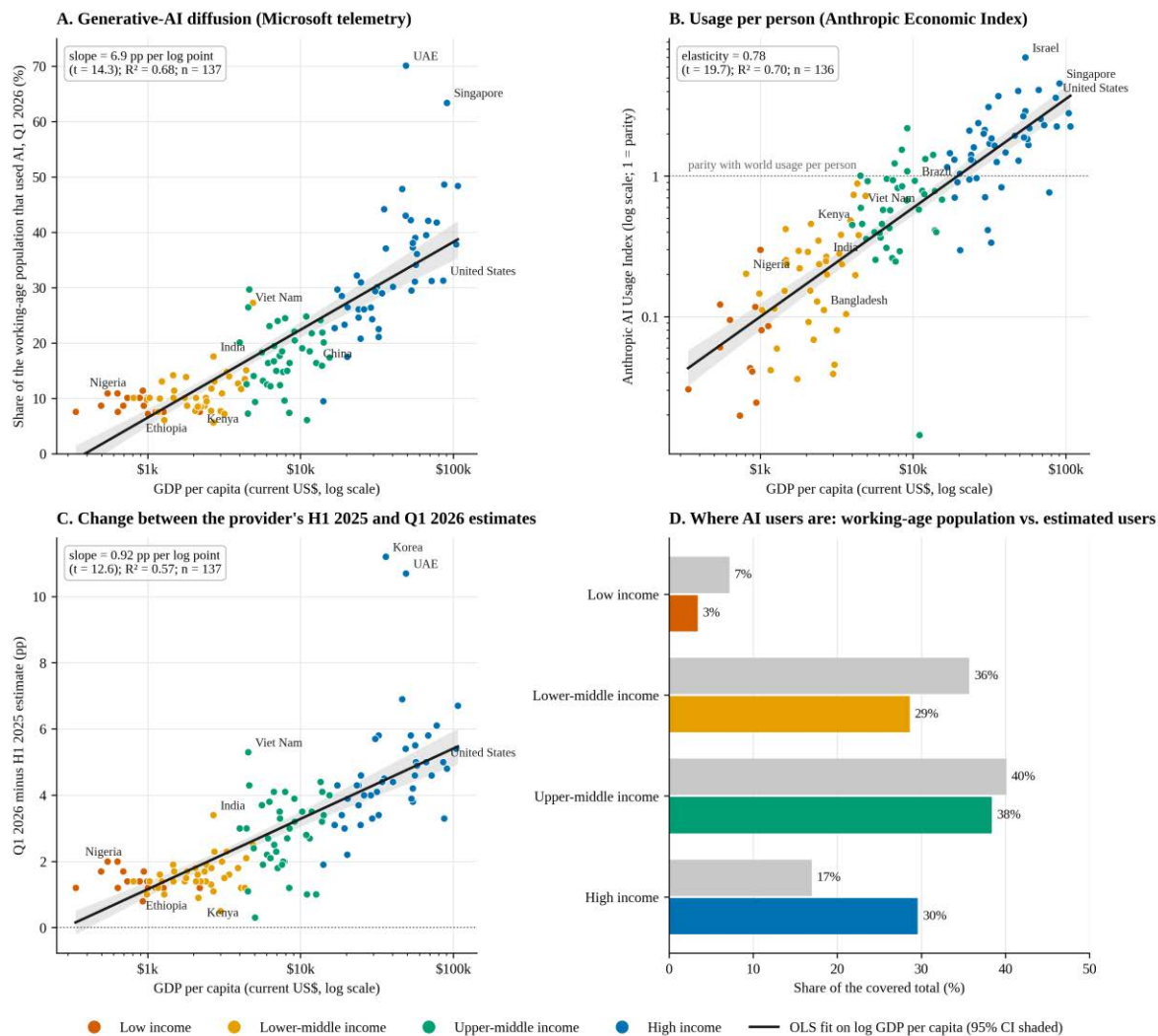
Table 5. What the evidence shows and what remains uncertain

Finding	Evidence	Confidence	Further evidence needed
1. AI use rises with income; the absolute gap has widened	Two provider measures agree (correlation 0.74)	High for the observed pattern; not causal	Representative household surveys
2. AI assistance raises output on tested tasks	Nine evaluations: +29% (10% to 51%); eight gains, one loss	Positive average; moderate. Size: low	Informal work and local-language evaluations
3. Supported AI tutoring can improve learning; unrestricted use can harm it	Eight comparisons, four in poorer countries: 0.21 SD (0.05 to 0.36); sensitive to exclusions	Positive average: low to moderate. Size: very low to low. Harm: low	Longer follow-up; replication of the Türkiye trial
4. Job and earnings effects differ by setting	Exposed freelance demand fell a fifth or more; Denmark earnings effects within $\pm 2\%$; US earnings higher; no average profit gain in Kenya	Low	Worker panels in poorer countries
5. Performance gaps narrow in some settings and widen in others	Ten within-study comparisons; subgroup definitions and statistical support differ	Low; mechanisms remain uncertain	Trials testing tool design and user support
6. AI predicts wealth levels well, but changes less reliably (in small-area estimation settings)	17 validation studies; selected R^2 : levels 0.55 to 0.83; changes 0.04 to 0.50	High for level prediction; limited for changes and consumption	Validation for changes and consumption
7. Targeting gains depend on the alternative and its cost	Phone data beat most feasible Togo rules; a survey-based method performed better in Bangladesh	Moderate for variation across settings; low for a general ranking	Comparisons including data and administrative costs
8. Our search found no causal study linking generative AI to consumption poverty in poorer countries	Search and six reviews' study lists	Limited by search scope	Trials measuring consumption poverty
9. AI can improve service delivery; patient benefits remain unclear	12 studies, including earlier AI: mostly process gains; one inconclusive generative-AI patient-outcome trial	Moderate for processes; low for patient benefits	Public-system trials measuring patient outcomes and costs

Source: Sections 2.2, 3, and 4; Figures 1, 3, and 4; Tables A6 to A9, A12, and A14.

Note: Confidence ratings are the authors' judgments, not formal ratings; they consider bias, consistency, precision, and applicability. In rows 2 and 3, confidence concerns the selected-study average, not a new application; parentheses give 95% confidence intervals. Learning is measured without the tool. SD = standard deviation; R^2 = share of variation explained. In row 6, the contrast between levels and changes is clearest in the four paired results from two studies in Figure 4. Row 8 records a search result, not an intervention effect.

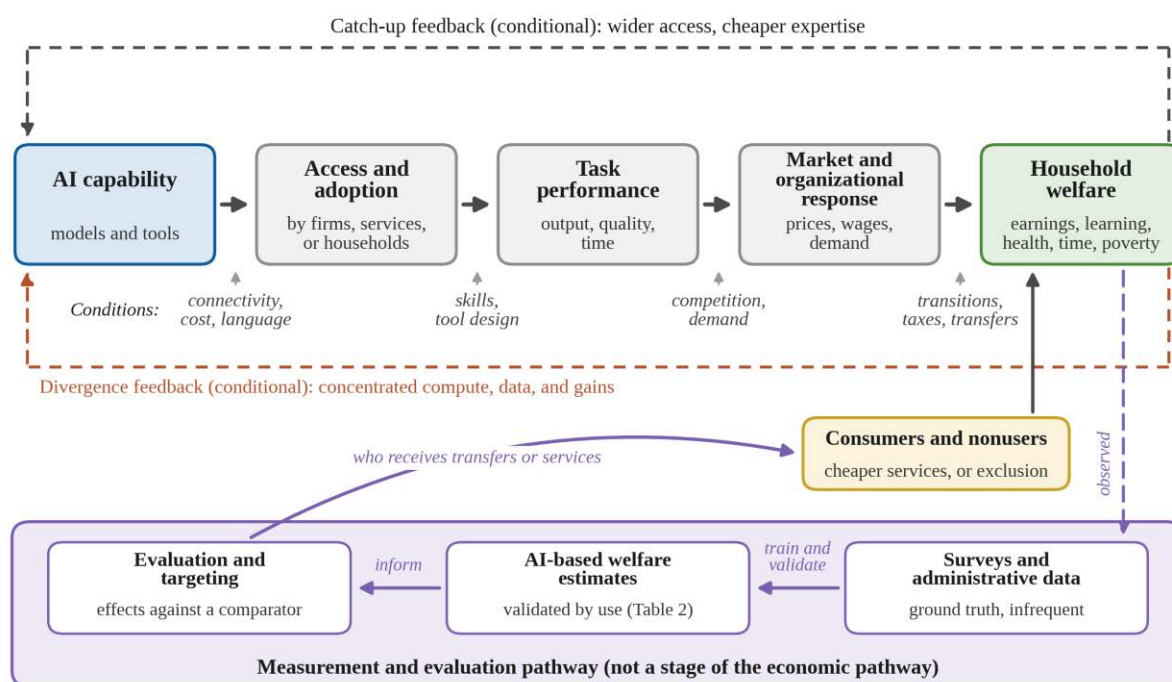
Figure 1. Measured AI use rises steeply with national income



Source: Anthropic (2025); Microsoft AI Economy Institute (2026); World Development Indicators; authors' calculations.

Note: Economies with population above 1 million; World Bank FY2027 income groups. Panel A: share of people aged 15 to 64 who used a generative-AI product in Q1 2026, modeled from Microsoft telemetry adjusted for device-market share, internet penetration, and population. Panel B: Anthropic AI Usage Index, August 2025 (usage per person on one provider; world average = 1; log scale). Panel C: change between Microsoft's H1 2025 and Q1 2026 estimates, in percentage points (pp), not a growth rate of use. Panel D: each group's share of the working-age population (gray bars) and of estimated users, the diffusion rate times the population aged 15 to 64 (colored bars), in the 143 covered economies. Lines: OLS fits on log GDP per capita; every slope is significant at the 1% level, though the HC1 standard errors ignore error in the modeled estimates. Regressions, group means, and vintages: Table A1.

Figure 2. How AI capability reaches household welfare, and the measurement and evaluation pathway that observes it



Source: Authors' elaboration.

Note: Solid arrows show the economic pathway, dashed black and red lines conditional feedbacks, and purple arrows the measurement and evaluation pathway. The pathway is schematic: learning, health, and time can change directly through an application, while earnings and poverty may also depend on market and organizational responses. Because adoption can occur anywhere in the service or production system, consumers and nonusers can gain or lose without adopting AI, through prices, services, or public allocation. Household welfare also covers jobs, prices, service access, and inequality. Further conditions include electricity, structured support, whether a task lies within what the tool does well, implementation costs, and pass-through to prices and wages; catch-up can also come from local data and capability, and divergence from lost demand for tradable services. Table A3 sets out what statistics observe at each stage. Evaluations can use observed survey or administrative outcomes directly; AI-based estimates are an optional measurement approach, validated against survey and administrative ground truth for each use (Table 2) before evaluation and targeting rely on them.

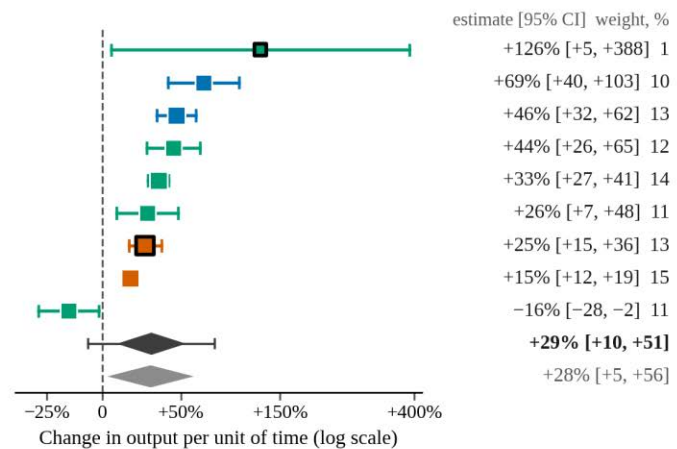
Figure 3. Effects on work with AI assistance and on learning in the selected studies: exploratory summary

A. Task performance with AI

Peng et al. (2023) Upwork, mostly India, Pakistan · completers
 Noy and Zhang (2023) United States · online
 Haslberger et al. (2025) United Kingdom · online · SE bound
 Dell'Acqua et al. (2025) Procter & Gamble, Europe, Americas
 Dell'Acqua et al. (2026) Global consultancy
 Cui et al. (2026) Three firms · 3 field RCTs · IV (adopters)
 Gambacorta et al. (2026) China · matched rollout
 Brynjolfsson et al. (2025) Mainly Philippines · rollout
 Becker et al. (2025) Open-source developers

Pooled ($k = 9$); $I^2 = 90\%$

Without high or serious risk of bias ($k = 7$); $I^2 = 92\%$



B. Learning measured without the tool

AI tutoring with support or guardrails vs no AI (pooled)

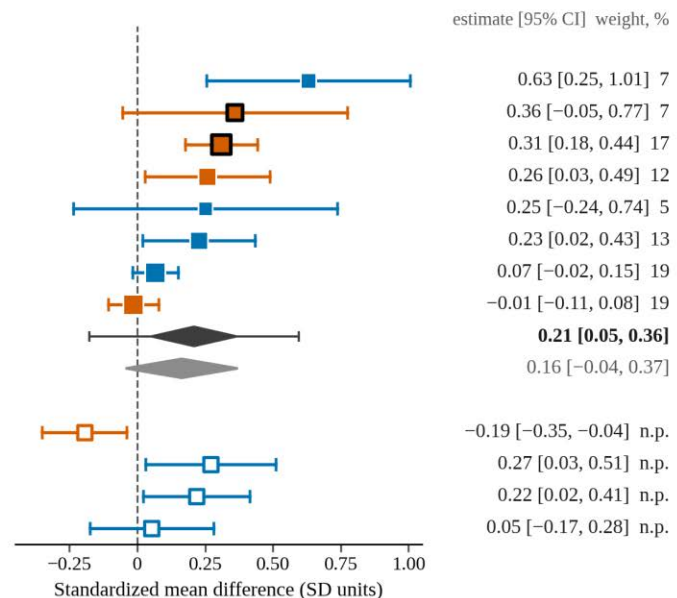
Kestin et al. (2025) US university physics, 1 lesson · SE bound
 Henkel et al. (2024) Ghana, grades 3–8, 8 months · SE adjusted
 De Simone et al. (2025) Nigeria, secondary, 6 weeks
 LearnLM Team, Fab AI (2026) Sierra Leone, 8 weeks
 Vanzo et al. (2025) Italy, high school, 8 weeks
 Fischer et al. (2025) Germany, laboratory, 25 minutes
 Oreopoulos and Low (2026) US middle schools, per school year
 Bastani et al. (2025), hint-only tutor Türkiye, 4 sessions

Pooled ($k = 8$); $I^2 = 76\%$

Without high risk of bias ($k = 6$); $I^2 = 71\%$

Other comparisons (shown, not pooled)

Bastani et al. (2025), unrestricted Türkiye; shared control
 Contractor and Reyes (2026) US college; search allowed
 Gallegos (2026), guidance Chile; vs AI access alone
 Gallegos (2026), encouragement Chile; vs AI access alone



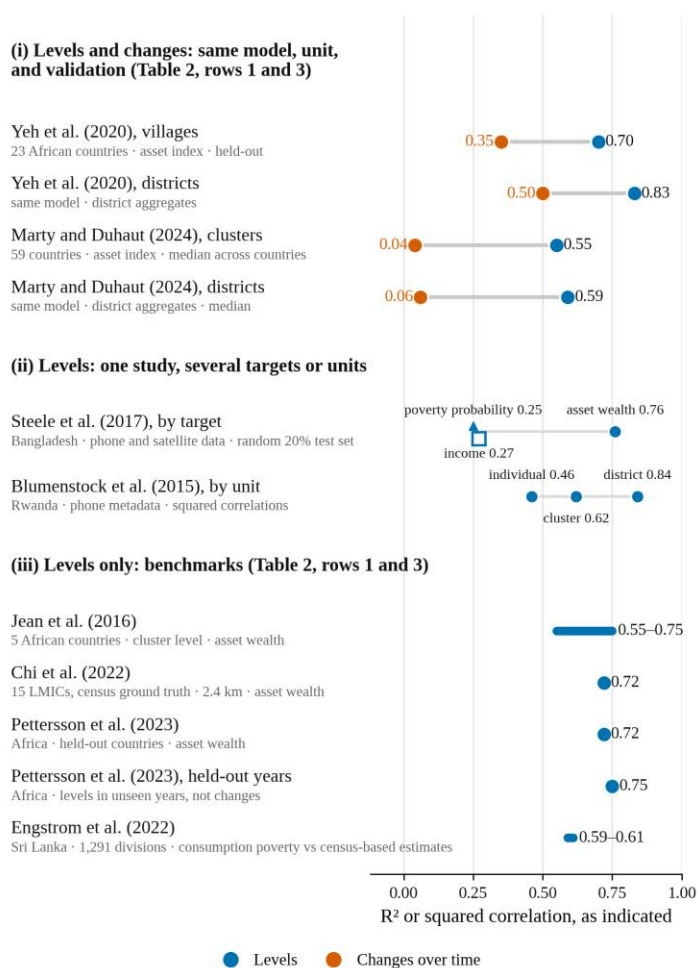
■ Poorer-country sample Shown, not pooled (n.p.) — Line through diamond: prediction interval
■ High-income setting Black edge: high or serious risk of bias (Table A14) Gray diamond: pooled without those studies
■ Mixed or unstated geography Pooled: Hartung–Knapp 95% interval

Source: Studies in Table A8; authors' calculations.

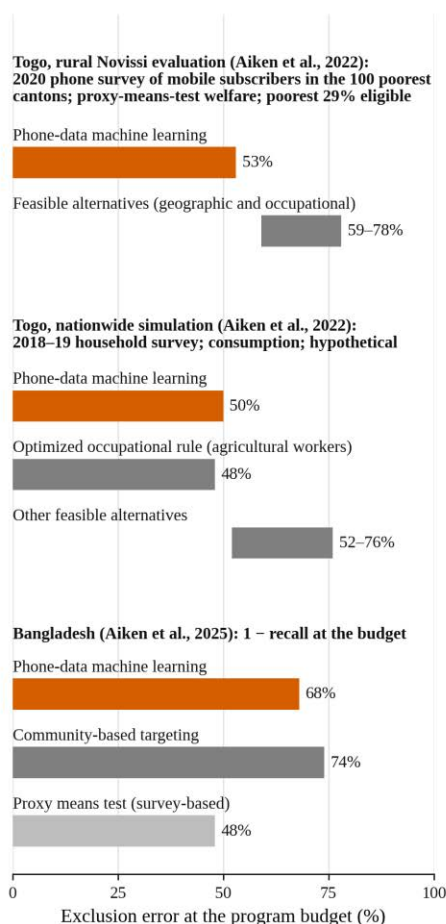
Note: Squares are study estimates with 95% intervals, sized by weight; hollow squares (n.p., not pooled) are comparisons shown but not pooled. Diamonds are small-sample (modified Hartung–Knapp) 95% intervals of the pooled estimates; the lines through them are prediction intervals, which describe the spread of effects expected in settings comparable to those studied, under the model. Squares with a black edge are studies rated at high or serious risk of bias (Table A14); the gray diamond pools the remaining studies. In panel B, Kestin *et al.* (2025) and Vanzo *et al.* (2025) do not report whether AI was unavailable during testing; without them, the pooled estimate is 0.17 SD (Table A12, row 8m). Business and labor-market outcomes are compared, not pooled, in Table A8. Contrasts, variances, and sensitivity analyses: Tables A8 and A12; by sample geography: Figure A1.

Figure 4. Selected validation results: AI-based measures predict wealth levels well, changes less well, and targeting results depend on the setting

A. Selected illustrations: accuracy for levels, changes, targets, and units



B. Targeting: exclusion error by setting



Source: Blumenstock *et al.* (2015); Jean *et al.* (2016); Steele *et al.* (2017); Yeh *et al.* (2020); Aiken *et al.* (2022, 2025b); Chi *et al.* (2022); Engstrom *et al.* (2022); Pettersson *et al.* (2023); Marty and Duhaut (2024); authors' compilation.

Note: Panel A: illustrations selected by one rule and compared, not averaged: (i) paired level and change results with the model, unit, and validation fixed; (ii) within-study contrasts across targets or units, with Blumenstock *et al.* (2015) as squared correlations; (iii) level benchmarks for multi-country asset wealth and for Sri Lankan small-area poverty against census-based estimates. Ten of the 17 validation studies in Table A6 are shown (the table says why the others are not); their range is not that of the literature. Panel B: exclusion error at the program budget (share of eligible poor not reached), one setting at a time: Togo's rural evaluation (2020 phone survey, poorest 29% eligible), Togo's nationwide simulation (2018 to 2019 survey, consumption), and Bangladesh (one minus recall).

Box 1. Two lenses in one country: Togo

Togo shows how the two lenses of this review meet in one program. In 2020, the government had no comprehensive registry of poor households. To select recipients of the Novissi emergency cash transfers, it used machine learning on mobile-phone records. In the rural evaluation, this method reached more of the eligible poor than the feasible geographic and occupational rules. In a nationwide simulation, it beat most alternatives. One occupation-based rule did slightly better: sending all transfers to agricultural workers (Aiken *et al.* 2022; Figure 4, panel B).

The transfers' effects were then estimated twice, once from household surveys and once from phone data (Aiken *et al.* 2025a). Comparing the two estimates tests the digital measure. Neither estimate shows the welfare gain from the targeting method itself. Evidence that a transfer improves welfare is not evidence that AI-based targeting improved it more than another method would have.

For an agency facing the same choice, the policy message has four parts. First, judge AI-based targeting against the best feasible alternative in the same setting, not against perfect data. The comparison must cover the same population and the same total budget. An advantage in reach at a fixed transfer budget can disappear once data costs come out of the same total. Second, ask what a registry would cost to build and keep current, against the accuracy it buys. Third, ask who is missing from the phone data, and give them a route to appeal. Fourth, test any digital measure against a survey sample before trusting it, whether it selects recipients or evaluates a program.

References

- Abramoff, M. D., Whitestone, N., Patnaik, J. L., Rich, E., Ahmed, M., Husain, L., *et al.* (2023). Autonomous artificial intelligence increases real-world specialist clinic productivity in a cluster-randomized trial. *npj Digital Medicine*, 6, 184.
- Acemoglu, D. (2024). Harms of AI. In J. B. Bullock *et al.* (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press.
- Acemoglu, D. (2025). The simple macroeconomics of AI. *Economic Policy*, 40(121), 13–58.
- Acemoglu, D., Autor, D., Hazell, J., and Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40(S1), S293–S340.
- Acemoglu, D., and Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6), 1488–1542.
- Adedinsewo, D. A., Morales-Lara, A. C., Afolabi, B. B., Kushimo, O. A., Mbakwem, A. C., Ibiyemi, K. F., *et al.* (2024). Artificial intelligence guided screening for cardiomyopathies in an obstetric population: A pragmatic randomized clinical trial. *Nature Medicine*, 30(10), 2897–2906.
- Aghion, P., Jones, B. F., and Jones, C. I. (2019). Artificial intelligence and economic growth. In *The Economics of Artificial Intelligence: An Agenda* (pp. 237–282). University of Chicago Press.
- Agweyu, A., Mwaniki, P., Menon, V., Korom, R., Isaaka, L., Wanyama, C., *et al.* (2026). Generative AI-enabled clinical decision support system in primary care: A pragmatic, cluster-randomized trial. *Nature Medicine*, 32, 3032–3039.
- Aiken, E., Ashraf, A., Blumenstock, J. E., Guiteras, R. P., and Mobarak, A. M. (2025b). Scalable targeting of social protection: When do algorithms out-perform surveys and community knowledge? Yale Economic Growth Center Discussion Paper 1117.
- Aiken, E., Bedoya, G., Blumenstock, J., and Coville, A. (2023). Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan. *Journal of Development Economics*, 161, 103016.
- Aiken, E., Bellue, S., Blumenstock, J. E., Karlan, D., and Udry, C. (2025a). Estimating impact with surveys versus digital traces: Evidence from randomized cash transfers in Togo. *Journal of Development Economics*, 175, 103477.
- Aiken, E., Bellue, S., Karlan, D., Udry, C., and Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603, 864–870.
- Aker, J. C., and Mbiti, I. M. (2010). Mobile phones and economic development in Africa. *Journal of Economic Perspectives*, 24(3), 207–232.
- Algama, C., Chandana, M., Dias, V., and Amarasinghe, K. (2026). Decision-centered evaluation of machine learning poverty maps using mobile phone and satellite data. arXiv:2609.23805.
- Alonso, C., Berg, A., Kothari, S., Papageorgiou, C., and Rehman, S. (2020). Will the AI revolution cause a great divergence? IMF Working Paper WP/20/184.

- Amodei, D. (2026). *The Adolescence of Technology*. Essay, January. darioamodei.com/essay/the-adolescence-of-technology
- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671), 669–674.
- Angrist, N., Bergman, P., and Matsheng, M. (2022). Experimental evidence on learning using low-tech when school is out. *Nature Human Behaviour*, 6, 941–950.
- Anthropic. (2025). *Anthropic Economic Index report: Uneven geographic and enterprise AI adoption*. September. arXiv:2511.15080. anthropic.com/research/economic-index-geography
- Anthropic. (2026). *Anthropic Economic Index: January 2026 report*. anthropic.com/research/anthropic-economic-index-january-2026-report
- Argyle, L. P. (2025). Political persuasion by artificial intelligence. *Science*, 390(6777), 983–984.
- Aron, J. (2018). Mobile money and the economy: A review of the evidence. *World Bank Research Observer*, 33(2), 135–188.
- Autor, D. (2024). Applying AI to rebuild middle class jobs. NBER Working Paper 32140.
- Bahia, K., Castells, P., Cruz, G., Masaki, T., Rodríguez-Castelán, C., and Sanfelice, V. (2024). The welfare effects of mobile broadband internet: Evidence from Nigeria. *Journal of Development Economics*, 170, 103314.
- Barriga-Cabanillas, O., Blumenstock, J. E., Lybbert, T. J., and Putman, D. S. (2025). Probing the limits of mobile phone metadata for poverty prediction and impact evaluation. *Journal of Development Economics*, 174, 103462.
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122 (see also the 2025 correction).
- Becker, J., Rush, N., Barnes, E., and Rein, D. (2025). Measuring the impact of early-2025 AI on experienced open-source developer productivity. arXiv:2507.09089.
- Bick, A., Blandin, A., and Deming, D. J. (2025). The rapid adoption of generative AI. *Management Science*, forthcoming (NBER Working Paper 32966).
- Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4), 401–416.
- Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073–1076.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Wiley.
- Brynjolfsson, E., Chandar, B., and Chen, R. (2026). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Stanford Digital Economy Lab working paper, revised August 2026.

- Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2), 889–942.
- Brynjolfsson, E., Rock, D., and Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1), 333–372.
- Burke, M., Driscoll, A., Lobell, D. B., and Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535), eabe8628.
- Burneo, A., Dinarte-Diaz, L., Lopez, C., and Molina, E. (2026). Can EdTech close learning gaps? Global evidence from digital interventions. Background paper for the *World Development Report 2026*. World Bank. thedocs.worldbank.org
- Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., *et al.* (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, 368, l6890.
- Capoot, A. (2026). OpenAI rules out IPO this year as Altman, Musk & Amodei warn AI is moving too fast. CNBC, September 12. [cnbc.com/2026/09/12/anthropics-amodei-proposes-plan-to-slow-the-pace-of-advancing-ai-capabilities.html](https://www.cnbc.com/2026/09/12/anthropics-amodei-proposes-plan-to-slow-the-pace-of-advancing-ai-capabilities.html)
- Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., Rockall, E., and Tavares, M. M. (2024). Gen-AI: Artificial intelligence and the future of work. IMF Staff Discussion Note SDN/2024/001. [imf.org/en/publications/staff-discussion-notes/issues/2024/01/14/gen-ai-artificial-intelligence-and-the-future-of-work-542379](https://www.imf.org/en/publications/staff-discussion-notes/issues/2024/01/14/gen-ai-artificial-intelligence-and-the-future-of-work-542379)
- Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., and Wadman, K. (2025). How people use ChatGPT. NBER Working Paper 34255.
- Chen, D., Kane, C., Kozlowski, A., Kunievsky, N., and Evans, J. A. (2025). The (short-term) effects of large language models on unemployment and earnings. arXiv:2509.15510.
- Cheng, X., Yin, L., Lin, C., Shi, Z., Zheng, H., Zhu, L., Liu, X., Chen, K., and Dong, R. (2024). Chatbot dialogic reading boosts comprehension for Chinese kindergarteners with higher language skills. *Journal of Experimental Child Psychology*, 240, 105842.
- Chi, G., Fang, H., Chatterjee, S., and Blumenstock, J. E. (2022). Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3), e2113658119.
- Comin, D., and Mestieri, M. (2018). If technology has arrived everywhere, why has income diverged? *American Economic Journal: Macroeconomics*, 10(3), 137–178.
- Contractor, Z., and Reyes, G. (2026). Experimental evidence on the learning impact of generative AI. IZA Discussion Paper 18792.
- Corral, P., Henderson, H., and Segovia, S. (2025). Poverty mapping in the age of machine learning. *Journal of Development Economics*, 172, 103377.
- Cruces, G., Fernández Meijide, D., Galiani, S., Gálvez, R. H., and Lombardi, M. (2026). Does generative AI narrow education-based productivity gaps? Evidence from a randomized experiment. NBER Working Paper 34851.

- Cui, K. Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., and Salz, T. (2026). The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers. *Management Science*, advance online publication.
- Czernich, N., Falck, O., Kretschmer, T., and Woessmann, L. (2011). Broadband infrastructure and economic growth. *Economic Journal*, 121(552), 505–532.
- DANE. (2025). *Encuesta de Tecnologías de la Información y las Comunicaciones en Hogares (ENTIC Hogares) 2024: Boletín técnico*. Departamento Administrativo Nacional de Estadística, Bogotá, August 1. dane.gov.co
- Dang, H.-A., and Lanjouw, P. (2026). Poverty mobility in developing countries: Stylized facts, data constraints, and subnational illustrations. IZA Discussion Paper.
- Dang, H.-A., and Rogers, F. H. (2008). The growing phenomenon of private tutoring: Does it deepen human capital, widen inequalities, or waste resources? *World Bank Research Observer*, 23(2), 161–200.
- Dang, H.-A., Hallegatte, S., Nguyen, M. C., and Trinh, T.-A. (2026a). Impacts of global warming on subnational poverty and inequality. *Nature Climate Change*, 16(2), 207–213.
- Dang, H.-A., Kilic, T., Abanokova, K., and Carletto, C. (2026b). Imputing poverty indicators without consumption data: An exploratory analysis. *Oxford Bulletin of Economics and Statistics*, forthcoming.
- Dell’Acqua, F., Ayoubi, C., Lifshitz, H., Sadun, R., Mollick, E., Mollick, L., Han, Y., Goldman, J., Nair, H., Taub, S., and Lakhani, K. R. (2025). The cybernetic teammate: A field experiment on generative AI reshaping teamwork and expertise. NBER Working Paper 33641.
- Dell’Acqua, F., McFowland III, E., Mollick, E., Lifshitz, H., Kellogg, K. C., Rajendran, S., Kraymer, L., Candelon, F., and Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2), 403–423.
- Demirci, O., Hannane, J., and Zhu, X. (2025). Who is AI replacing? The impact of generative AI on online freelancing platforms. *Management Science*, 71(10), 8097–8108.
- Demombynes, G., Langbein, J., and Weber, M. (2025). The exposure of workers to artificial intelligence in low- and middle-income countries. World Bank Policy Research Working Paper 11057.
- Deng, R., Jiang, M., Yu, X., Lu, Y., and Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, 227, 105224.
- DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177–188.
- De Simone, M. E., Tiberti, F., Barron Rodriguez, M., Manolio, F., Mosuro, W., and Dikoru, E. J. (2025). From chalkboards to chatbots: Evaluating the impact of generative AI on learning outcomes in Nigeria. World Bank Policy Research Working Paper 11125, revised December 2025.

- de Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, 7(10), 2191–2194.
- Dillon, E. W., Jaffe, S., Immorlica, N., and Stanton, C. T. (2025). Shifting work patterns with generative AI. NBER Working Paper 33795.
- Egger, M., Davey Smith, G., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634.
- Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1), 355–364.
- Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., *et al.* (2017). Living systematic review: 1. Introduction, the why, what, when, and how. *Journal of Clinical Epidemiology*, 91, 23–30.
- Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702), 1306–1308.
- Engstrom, R., Hersh, J., and Newhouse, D. (2022). Poverty from space: Using high-resolution satellite imagery for estimating economic well-being. *World Bank Economic Review*, 36(2), 382–412.
- Eurostat. (2025). *Use of Artificial Intelligence in Enterprises*. Statistics Explained, Eurostat, Luxembourg, December. ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises
- Fabregas, R., Harigaya, T., Kremer, M., and Ramrattan, R. (2025). Digital information provision and behavior change: Lessons from six experiments in East Africa. *American Economic Journal: Applied Economics*, 17(1), 527–566.
- Fang, L., Yuan, Z., Zhang, K., Donati, D., and Sarvary, M. (2026). Generative AI and sales productivity: Field experiments in online retail. arXiv:2510.12049.
- Ferman, B., Lima, L., and Riva, F. (2021). Artificial intelligence, teacher tasks and individualized pedagogy. SocArXiv working paper (also "Experimental evidence on artificial intelligence in the classroom," MPRA Paper 103934, 2020).
- Fischer, M., Rau, H. A., and Rilke, R. M. (2025). AI tutoring enhances student learning without crowding out reading effort. IZA Discussion Paper 18338.
- Fütterer, T., Bardach, L., Kuhn, J., Keller, S. D., and Gerjets, P. (2026). Enhancing school students' self-regulated learning through generative AI support: A randomized controlled trial. *Educational Psychology Review*, advance article.
- Gallegos, S. (2026). Guidance over adoption: Experimental evidence on AI-assisted learning. IZA Discussion Paper 18513.
- Gambacorta, L., Qiu, H., Shan, S., and Rees, D. M. (2026). Generative AI and labour productivity: A quasi experiment on coding. *Journal of Financial Stability*, 84, 101543.
- Gelbach, J. B. (2016). When do covariates matter? And which ones, and how much? *Journal of Labor Economics*, 34(2), 509–543.

- Gmyrek, P., Berg, J., and Bescond, D. (2023). Generative AI and jobs: A global analysis of potential effects on job quantity and quality. ILO Working Paper 96. [ilo.org \(Working Paper 96\)](https://ilo.org/publications/working-paper-96)
- Gmyrek, P., Berg, J., *et al.* (2025). Generative AI and jobs: A refined global index of occupational exposure. ILO Working Paper 140. ilo.org/publications/generative-ai-and-jobs-refined-global-index-occupational-exposure
- Gmyrek, P., Viollaz, M., and Winkler, H. (2026). Disruption without dividend? How the digital divide and task differences split GenAI's global impact. ILO Working Paper 166 (also World Bank Policy Research Working Paper 11328). ilo.org/publications/disruption-without-dividend-how-digital-divide-and-task-differences-split
- Gmyrek, P., Winkler, H., and Garganta, S. (2024). Buffer or bottleneck? Employment exposure to generative AI and the digital divide in Latin America. World Bank Policy Research Working Paper 10863.
- Google, Temasek, and Bain. (2025). *e-Conomy SEA 2025*. economysea.withgoogle.com
- Hall, O., Dompae, F., Wahab, I., and Dzanku, F. M. (2023). A review of machine learning and satellite imagery for poverty prediction: Implications for development research and applications. *Journal of International Development*, 35(7), 1753–1768.
- Harari, Y. N. (2024). *Nexus: A Brief History of Information Networks from the Stone Age to AI*. Random House.
- Haslberger, M., Gingrich, J., and Bhatia, J. (2025). No great equalizer: Experimental evidence on productivity effects of generative AI use in the UK labor market. Working paper, University of St. Gallen and University of Oxford (SSRN 4594466).
- Hedges, L. V., Gurevitch, J., and Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, 80(4), 1150–1156.
- Hedges, L. V., Tipton, E., and Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1), 39–65.
- Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., and Lee, A. (2024). Effective and scalable math support: Experimental evidence on the impact of an AI math tutor in Ghana. arXiv:2402.09809.
- Higgins, J. P. T., Li, T., and Deeks, J. J. (2024). Choosing effect measures and computing estimates of effect. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions*, version 6.5, chapter 6. Cochrane.
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414), 557–560.
- Hjort, J., and Poulsen, J. (2019). The arrival of fast internet and employment in Africa. *American Economic Review*, 109(3), 1032–1079.
- Huang, L. Y., Hsiang, S., and Gonzalez-Navarro, M. (2021). Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs. NBER Working Paper 29105.

- Hui, X., Reshef, O., and Zhou, L. (2024). The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. *Organization Science*, 35(6), 1977–1989.
- Humlum, A., and Vestergaard, E. (2025a). Still waters, rapid currents: Early labor market transformation under generative AI. NBER Working Paper 33777 (revised; earlier version titled “Large language models, small labor market effects”).
- Humlum, A., and Vestergaard, E. (2025b). The unequal adoption of ChatGPT exacerbates existing inequalities among workers. *Proceedings of the National Academy of Sciences*, 122(1), e2414972121.
- IBGE. (2025). *Pesquisa de Inovação Semestral 2024: Tecnologias digitais avançadas, teletrabalho e cibersegurança*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, September 24. ibge.gov.br
- Idan, L., and Anand, B. (2026). Generative AI and the productivity divide: Human–AI complementarities in education and knowledge work. arXiv:2605.18143.
- IEA. (2026). *Key Questions on Energy and AI*. World Energy Outlook Special Report. International Energy Agency. iea.org/reports/key-questions-on-energy-and-ai
- International AI Safety Report. (2026). *International AI Safety Report 2026*. February. internationalaisafetyreport.org
- IntHout, J., Ioannidis, J. P. A., Rovers, M. M., and Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, 6(7), e010247.
- Jagnani, M., and Pande, R. (2024). Forecasting fate: Experimental evaluation of a flood early warning system. Working paper, Yale University, April.
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790–794.
- Jensen, R. (2007). The digital provide: Information (technology), market performance, and welfare in the South Indian fisheries sector. *Quarterly Journal of Economics*, 122(3), 879–924.
- Jerven, M. (2013). *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It*. Cornell University Press.
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293.
- Ju, H., and Aral, S. (2026). Collaborating with AI agents: Field experiments on teamwork, productivity, and performance. arXiv:2503.18238 (revised February 2026).
- Kauhanen, A., and Rouvinen, P. (2025). Assessing early labour market effects of generative AI: Evidence from population data. *Applied Economics Letters*, advance online publication.
- Kestin, G., Miller, K., Klaes, A., Milbourne, T., and Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15, 17458.

- Knapp, G., and Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22(17), 2693–2710.
- Koo, W. W. (2025). Cross-lingual effects of AI-generated content on human work. *Scientific Reports*, 15, 30949.
- Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4), 1281–1317.
- Korinek, A., and Stiglitz, J. E. (2021). Artificial intelligence, globalization, and strategies for economic development. NBER Working Paper 28453.
- Korom, R., Kiptinness, S., Adan, N., Said, K., Ithuli, C., Rotich, O., *et al.* (2025). AI-based clinical decision support for primary care: A real-world study. arXiv:2507.16947.
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253.
- LearnLM Team and Eedi. (2025). AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms. arXiv:2512.23633.
- LearnLM Team and Fab AI. (2026). *Teaching with Gemini: Measuring the Impact of Guided Learning on Student Mathematics Progress in Sierra Leone*. Technical report, Google DeepMind, May 15 (AEA RCT Registry AEARCTR-0016651). storage.googleapis.com/deepmind-media/LearnLM/learnLM_sierraleone_may26.pdf
- Lee, K., and Braithwaite, J. (2022). High-resolution poverty maps in Sub-Saharan Africa. *World Development*, 159, 106028.
- Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., and Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- Lehdonvirta, V., Wu, B., and Hawkins, Z. (2024). Compute North vs. Compute South: The uneven possibilities of compute-based AI governance around the globe. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)*.
- Lin, H., Li, R., Liu, Z., Chen, J., Yang, Y., Chen, H., *et al.* (2019). Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: A multicentre randomized controlled trial. *EClinicalMedicine*, 9, 52–59.
- Liu, Y., Gao, J., Liu, J., Walline, J. H., Liu, X., Zhang, T., *et al.* (2021). Development and validation of a practical machine-learning triage algorithm for the detection of patients in need of critical care in the emergency department. *Scientific Reports*, 11, 24044.
- Lucchetti, L., Corral, P., Ham, A., and Garriga, S. (2025). An application of LASSO and multiple imputation techniques to income dynamics with cross-sectional data. *Review of Income and Wealth*, 71(1), e12693.
- Ma, Y., Feng, T., Fairlie, R., Liu, C., Loyalka, P. K., Rozelle, S., and Zhang, X. (2026). Putting AI to the test: Evidence from a large-scale RCT in rural China. CESifo Working Paper 12837 (SSRN 7195260).

- MacPherson, P., Webb, E. L., Kamchedzera, W., Joeke, E., Mjoli, G., Lalloo, D. G., *et al.* (2021). Computer-aided X-ray screening for tuberculosis and HIV testing among adults with cough in Malawi (the PROSPECT study): A randomised trial and cost-effectiveness analysis. *PLOS Medicine*, 18(9), e1003752.
- Maier, S., Gunzenhäuser, M., Schweisthal, J., Schneider, M., and Feuerriegel, S. (2026). A meta-analysis of the effect of generative AI on productivity and learning in programming. arXiv:2605.04779.
- Marty, R., and Duhaut, A. (2024). Global poverty estimation using private and public sector big data sources. *Scientific Reports*, 14, 3160.
- Mathenge, W., Whitestone, N., Nkurikiye, J., Patnaik, J. L., Piyasena, P., Uwaliraye, P., *et al.* (2022). Impact of artificial intelligence assessment of diabetic retinopathy on referral service uptake in a low-resource setting: The RAIDERS randomized trial. *Ophthalmology Science*, 2(4), 100168.
- Mehmood, S., Goessmann, C., and Ash, E. (2026). Courts of tomorrow: Evidence from a nationwide rollout of generative AI. CEPR Discussion Paper DP21783, July.
- Merfeld, J. D., Dang, H.-A., and Newhouse, D. (2025). Improving estimates of mean welfare and uncertainty in developing countries. World Bank Policy Research Working Paper 10348, revised June 2025.
- METR. (2026). We are changing our developer productivity experiment design. METR blog, February 24. metr.org/blog/2026-02-24-uplift-update
- Microsoft AI Economy Institute. (2026). *AI Diffusion Report: Q1 2026 update* (data release). Microsoft. microsoft.com/research/group/aiei/ai-diffusion
- Misra, A., Wang, J., McCullers, S., White, K., and Lavista Ferres, J. (2025). Measuring AI diffusion: A population-normalized metric for tracking global AI usage. arXiv:2511.02781.
- Muralidharan, K., Singh, A., and Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review*, 109(4), 1426–1460.
- Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., *et al.* (2024). Global prediction of extreme floods in ungauged watersheds. *Nature*, 627, 559–563.
- Newhouse, D. (2024). Small area estimation of poverty and wealth using geospatial data: What have we learned so far? *Calcutta Statistical Association Bulletin*, 76(1), 7–32.
- Noy, S., and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192.
- Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.
- OECD. (2025). *Competition in Artificial Intelligence Infrastructure*. OECD Publishing, Paris. oecd.org/en/publications/2025/11/competition-in-artificial-intelligence-infrastructure_69319aee.html

- Oreopoulos, P., Liut, M., Sungu, A., and Low, N. (2026). Making AI tutoring productive: Evidence from a mastery-based math practice experiment. NBER Working Paper 35621.
- Oreopoulos, P., and Low, N. (2026). One click away: AI tutoring with Khanmigo in a two-year school experiment. NBER Working Paper 35620.
- Otis, N., Clarke, R., Delecourt, S., Holtz, D., and Koning, R. (2026). The uneven impact of generative artificial intelligence on entrepreneurial performance: Evidence from a field experiment in Kenya. *Management Science*, advance online publication.
- Oxford Insights. (2025). *Government AI Readiness Index 2025*. Oxford Insights. oxfordinsights.com/ai-readiness/government-ai-readiness-index-2025
- Park, H., Buchmann, C., Choi, J., and Merry, J. J. (2016). Learning beyond the school walls: Trends and implications. *Annual Review of Sociology*, 42(1), 231–252.
- Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv:2302.06590.
- Pettersson, M. B., Kakooei, M., Ortheden, J., Johansson, F. D., and Daoud, A. (2023). Time series of satellite imagery improve deep learning estimates of neighborhood-level poverty in Africa. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*.
- Pokhriyal, N., and Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, 114(46), E9783–E9792.
- Qazi, I. A., Ali, A., Khawaja, A. U., Akhtar, M. J., Sheikh, A. Z., and Alizai, M. H. (2026). Large language model diagnostic assistance for physicians in a lower-middle-income country: A randomized controlled trial. *Nature Health*, 1, 198–205.
- Qiao, D., Rui, H., and Xiong, Q. (2025). AI and jobs: Has the inflection point arrived? Evidence from an online labor platform. Working paper, 2025 version (first released as arXiv:2312.04180).
- Ratledge, N., Cadamuro, G., de la Cuesta, B., Stigler, M., and Burke, M. (2022). Using machine learning to assess the livelihood impact of electricity access. *Nature*, 611, 491–495.
- Rodriguez-Segura, D. (2022). EdTech in developing countries: A review of the evidence. *World Bank Research Observer*, 37(2), 171–203.
- Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., *et al.* (2022). Tackling climate change with machine learning. *ACM Computing Surveys*, 55(2), 1–96.
- Rounding, N., Arif, L. S., Berg, J., *et al.* (2026). Impact of LLM assistance on physician decision-making: A multi-country randomized controlled trial. *npj Digital Medicine*, advance online publication, September 9.
- Röver, C., Knapp, G., and Friede, T. (2015). Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology*, 15, 99.

- Sautmann, A., and Weinert, J. (2026). Artificial intelligence interventions to support frontline healthcare workers in low-resource settings: A review. Background paper for the *World Development Report 2026*. World Bank. thedocs.worldbank.org
- Serajuddin, U., Uematsu, H., Wieser, C., Yoshida, N., and Dabalen, A. (2015). Data deprivation: Another deprivation to end. World Bank Policy Research Working Paper 7252.
- Smythe, I. S., and Blumenstock, J. E. (2022). Geographic microtargeting of social assistance with high-resolution poverty maps. *Proceedings of the National Academy of Sciences*, 119(32), e2120025119.
- Stanley, T. D., and Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1), 60–78.
- Stanley, T. D., Doucouliagos, H., and Steel, P. (2018). Does ICT generate economic growth? A meta-regression analysis. *Journal of Economic Surveys*, 32(3), 705–726.
- Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., *et al.* (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society Interface*, 14(127), 20160690.
- Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., *et al.* (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, 355, i4919.
- Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., *et al.* (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, 366, l4898.
- Suleyman, M., with Bhaskar, M. (2023). *The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma*. Crown.
- Suri, T., and Jack, W. (2016). The long-run poverty and gender impacts of mobile money. *Science*, 354(6317), 1288–1292.
- Tekic, Z., Omelchenko, Y., and Dunaeva, A. (2026). Generative AI and the uneven disruption of freelance labor: Exposure, diffusion, and adaptation under constraint. *Technology in Society*, 88, 103464.
- Teutloff, O., Einsiedler, J., Kässi, O., Braesemann, F., Mishkin, P., and del Rio-Chanona, R. M. (2025). Winners and losers of generative AI: Early evidence of shifts in freelancer demand. *Journal of Economic Behavior and Organization*, 235, 106845.
- Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, 20(3), 375–393.
- Triyana, M., Jiang, A. W., Hu, Y., and Naoaj, S. (2026). Climate shocks and the poor: A review of the literature. *Wiley Interdisciplinary Reviews: Climate Change*, 17(1), e70039.
- UN News. (2026). UN panel calls for stronger safeguards as AI agents advance. United Nations. news.un.org/en/story/2026/09/1168380
- UNCTAD. (2025). *Technology and Innovation Report 2025: Inclusive Artificial Intelligence for Development*. United Nations. unctad.org/publication/technology-and-innovation-report-2025

- UNDP. (2025). *Human Development Report 2025. A Matter of Choice: People and Possibilities in the Age of AI*. United Nations Development Programme. hdr.undp.org/content/human-development-report-2025
- Vaccaro, M., Almaatouq, A., and Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8, 2293–2303.
- van der Weide, R., Blankespoor, B., Elbers, C., and Lanjouw, P. (2024). How accurate is a poverty map based on remote sensing data? An application to Malawi. *Journal of Development Economics*, 171, 103352.
- Vanzo, A., Pal Chowdhury, S., and Sachan, M. (2025). GPT-4 as a homework tutor can improve student engagement and learning outcomes. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 31119–31136.
- Verme, P. (2025). Predicting poverty. *World Bank Economic Review*, 39(4), 1039–1064.
- Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., and Demszky, D. (2025). Tutor CoPilot: A human-AI approach for scaling real-time expertise. arXiv:2410.03017.
- Weidenkaff, F., and Huynh, P. (2026). Generative AI and jobs in Viet Nam: Labour market exposure and policy considerations. ILO Brief, April. Geneva: International Labour Organization. ilo.org/publications/generative-ai-and-jobs-viet-nam-labour-market-exposure-and-policy
- Weidlich, J., Gašević, D., Drachsler, H., and Kirschner, P. (2025). ChatGPT in education: An effect in search of a cause. *Journal of Computer Assisted Learning*, 41, e70105.
- Winkler, H. (2026). Technology, labor markets, and inequality across the development spectrum: Reviewing the evidence. *Oxford Open Economics*, advance article.
- World Bank. (2021). *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank. worldbank.org/en/publication/wdr2021
- World Bank. (2024). *Poverty, Prosperity, and Planet Report 2024: Pathways Out of the Polycrisis*. Washington, DC: World Bank. worldbank.org/en/publication/poverty-prosperity-and-planet
- World Bank. (2026). *World Development Report 2026: The Promise of Artificial Intelligence*. World Bank. worldbank.org/en/publication/wdr2026
- World Economic Forum, and Kearney. (2026). *Global Value Chains Outlook 2026: Orchestrating Corporate and National Agility*. World Economic Forum. weforum.org/publications/global-value-chains-outlook-2026-orchestrating-corporate-and-national-agility
- Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., and Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11, 2583.
- Yu, S., Yan, C., Qin, G., Pazo, E. E., He, X., Qi, P., et al. (2025). Assessing the impact of AI-assisted portable slit lamps on rural primary ophthalmic medical service. *Current Eye Research*, 50(5), 551–558.

Appendix A. Search procedure and additional tables and figures

A.1 Search and screening procedure for the evidence synthesis

Eligibility criteria

Three kinds of evidence enter the review under separate rules, because no single set of conditions fits all three. Tool-assignment studies (task performance and learning) are eligible if they meet conditions (i) to (iv) below; condition (v) governs entry into a pooled estimate, not into the review, so that a null or directional result, or one that cannot be converted to a common effect size, stays on the evidence map and in the structured synthesis. The criteria are applied in the fixed order given, and the first failing criterion is recorded as the exclusion reason. The business and labor-market domain covers profits, sales, earnings, employment, and demand. It includes deployments of a tool to firms or workers that meet conditions (i) to (iv), such as an advisory service offered to small businesses, and market-response studies, which concern populations that need not use any tool. Market-response studies are eligible if a dated release or rollout of a generative-AI system supplies the variation, the design is a difference-in-differences, event-study, or staggered-exposure comparison whose identifying assumption can be stated, and the outcome is earnings, employment, demand, or profits.

Findings in the business and labor-market domain are compared by outcome and design, as a synthesis without meta-analysis (Campbell *et al.* 2020), and never pooled with task-performance effects: none of its subgroups meets the rule of at least five independent estimates of the same outcome and treatment comparison (the five platform studies measure postings, jobs, earnings, or participation on different platforms), and several platform and register studies use ChatGPT's release as a common source of variation, so their estimates are not independent draws from many technology shocks. Welfare-prediction validation studies are eligible if a predictive model, by any machine-learning or statistical method from 2015 onward, predicts a welfare quantity in a low- or middle-income country from satellite, phone, or other proxy data and is validated against survey or census ground truth with an out-of-sample design; the accuracy metric is recorded as published and compared by use (Table 2), never averaged, and evaluations that use AI-predicted welfare as their outcome are recorded separately (Section 3.5). Comparisons between AI designs, such as pedagogical prompts against a standard tool, are recorded and shown but not pooled with comparisons of AI against none. The task-performance domain pools comparisons of work with the tool against work without it; comparisons of AI with a human collaborator or tutor are recorded but not pooled. Older digital technologies and forecasting applications are labeled background evidence wherever they appear.

- (i) Intervention. Human participants use a generative-AI tool, meaning a large language model or a comparable generative system released in 2020 or later. Examples are a general chat assistant, an embedded coding assistant, an AI tutor, an AI advisory service, and a clinical decision-support tool built on a language model.
- (ii) Comparator. There is a group without the tool, or with a non-generative alternative.
- (iii) Design. Assignment is random, or the design has a credible source of quasi-random variation, meaning a staggered rollout, a regression discontinuity, a difference-in-differences design, or a synthetic control. The identifying assumption of each quasi-experiment is stated and assessed at coding; a design label alone does not establish credibility.

- (iv) Outcome. At least one quantitative outcome falls in one of four categories: task output, quality, speed, or errors; earnings, profits, sales, employment, hiring, household consumption, income, or poverty status; learning or test scores; clinical quality, patient outcomes, farm decisions, or yields.
- (v) Effect size, for pooling only. The paper reports enough information to compute a standardized mean difference, in the source's standard-deviation units or converted with the control-arm standard deviation, for learning and quality outcomes, or a log response ratio for output per unit time, together with a standard error or a confidence interval. A study that fails this condition remains in the review and is reported descriptively (Table A8).

Excluded by construction are opinion studies, descriptive adoption surveys, benchmark evaluations of a model without human participants, and studies whose only outcomes are satisfaction or perceptions. The window runs from January 1, 2020, which precedes the release of GPT-3, to the end of September 2026, when the search ended. The search is limited to English, and that limit is a stated weakness rather than a design choice.

How the search was conducted

The search was conducted from the project's conception to the end of September 2026. It started from the studies known to the authors, listed as the initial list in Tables A5 and A6, and extended that list with two rounds of recorded web searches. The first ran 27 queries in a general web search engine (19 for the first three outcome domains and 8 for validation studies), each combining a generative-AI tool block, an experimental or quasi-experimental design block, and an outcome block with country and region terms, and returned 230 links; three compilations were screened for citation chasing. The second ran 28 queries (251 links), 22 of them site-restricted searches of working-paper series and organizational repositories (NBER, SSRN, IZA, CEPR, the World Bank Policy Research Working Paper series and Open Knowledge Repository, arXiv, RePEc, J-PAL and Innovations for Poverty Action, 3ie, and Stanford SCALE), chased the reference list of Burneo *et al.* (2026) and the J-PAL Project AI Evidence list backwards, and re-checked ten records for newer versions.

Supplementary checks added two studies: a multi-country clinical trial and a mastery-practice experiment. Full texts obtained in September 2026 resolved seven first-round records that had been left unresolved (Table A5, rows 46 to 51, and a later version of row 36). A final check on September 29, 2026 added the Pakistan courts rollout (Table A5, row 52), the seven earlier-AI health trials of the WDR 2026 health review (rows 53 to 59), and a seven-country validation study (Table A6, row 22). Three distinct tasks were involved: an AI assistant ran the queries and the initial data extraction against a fixed query log set by the authors; the two authors screened the returned records; and the authors verified every pooled input against the full text of its source. Every query, source, hit count, and screening decision of the two rounds is recorded in the search logs, and the supplementary checks in Table A5 (rows 43 and 45). No record entered the verified list on the strength of a web hit alone: each verified study was resolved to a digital object identifier or a stable working-paper URL and read before coding.

The search was limited to English and its results were not screened by an independent second screener, so the list is not exhaustive; that is why the article is a structured critical review rather than a systematic review and why every quantitative summary built on the list is labeled exploratory (Section 4.1). All candidates were read in full and coded, with the decision and its

reason recorded in Table A5. Nine of them evaluate tools that are not generative: an adaptive practice system built on pre-generative machine learning (Ma *et al.* 2026) and the seven health trials (rows 53 to 59), all kept as benchmarks from earlier AI (Table 3), and a scripted chatbot compared with a human reading partner, which is out of scope (Cheng *et al.* 2024).

Screening and coding

Screening was done by the two authors; no record was screened independently by both, so no agreement statistic is available. Every returned link was screened on title and abstract with a decision code and a reason recorded per record. Of the 155 unique records of the first round, 68 were not studies, 20 were reviews or starting lists of studies, 14 were capability or perception studies, and 8 were descriptive studies set aside at title screening; 5 were excluded on the criteria, 12 could not be resolved from the retrieved material, none of them on the intervention lens, 27 new candidates entered the candidate lists, and 1 was a later version of a study already listed. Seven of these records became resolvable only when their full texts were obtained in September 2026: six candidates (Table A5, rows 46 to 51) and the published version of row 36. The second round added the candidates of Table A5, rows 33 to 42. The screening decisions are auditable record by record in the search logs, but they were not replicated, and that is the main procedural limitation of the list.

Coding follows a fixed sheet. Two of its fields place every estimate in time: the data-collection window of the study and the version and release date of the model tested, so that effects can be read against the release calendar of the models. Learning and quality effects are entered in standard-deviation units as the source reports them, as Cohen's d or as a regression coefficient on an outcome standardized by the source, or converted by dividing a reported coefficient and its standard error by the control-arm standard deviation of the outcome (Borenstein *et al.* 2009); no entry is re-standardized as Hedges' g , and the small-sample correction would change no pooled entry by more than 0.003 at these sample sizes. Binary outcomes are converted from log odds. Effects on output per unit of time are converted to log response ratios (Hedges *et al.* 1999), so that a fall in task time is a positive effect (Tables A8 and A12). Assigned, observed, and analyzed samples are recorded separately, and the unit of assignment is recorded so that student-level inference from a school-randomized trial is flagged. Each pooled record carries its source table, page, or column, arm counts, assignment unit, estimator, effect conversion, and variance calculation.

Where a trial randomized clusters but reports only student-level inference, the variance is inflated by a design effect under a stated intraclass correlation, and where a study reports only a p-value threshold, the largest standard error consistent with it at the unit of assignment, under a normal reference distribution, is used and labeled a bound; where a study reports effects for several tasks but not for their total, the variance of the total is bounded by assuming perfectly correlated task errors and varied in sensitivity rows, as recommended for imputed quantities (Higgins *et al.* 2024). All three cases are flagged in the variance field and dropped in a sensitivity run (Tables A8 and A12). Where a study reports only a one-sided p-value for its estimate, the standard error it implies under a normal reference distribution is used. Each pooled record also states what its contrast represents: assignment of a tool or encouragement to use it, an IV estimate for adopters, effects among users after a rollout, or a contrast conditional on completion. Multiple outcomes from one study are retained; robust variance estimation under a correlated-effects weighting scheme (Hedges *et al.* 2010; Tipton 2015), the specification of Burneo *et al.* (2026), would need

more estimates per domain than the list contains and is not used. At the current counts, one estimate enters per study and domain, DerSimonian–Laird random effects are used, and the modified Hartung–Knapp interval is the primary interval (Appendix B).

How the review treats its limitations

This article is a structured critical review rather than a systematic review: every study on the list is appraised on the same recorded fields, the study-level discussion of Section 3 is its main synthesis, and the pooled estimates of Section 4 are an exploratory summary of the studies the list contains. Risk of bias in the pooled studies is rated with RoB 2 for the trials and ROBINS-I for the two workplace rollouts (Sterne *et al.* 2016, 2019; Table A14); one author made the ratings from the full texts with AI assistance and the other author reviewed them. Design, assignment unit, attrition where above 30%, provider participation, and the identifying assumption of each quasi-experiment are also recorded (Table A8; Section 3). Small-study tests and precision-effect corrections are not run at nine or fewer studies per domain (Egger *et al.* 1997; Stanley and Doucouliagos 2014).

Table A1. AI use and income per capita: regressions behind Figure 1

No.	Dependent variable	Sample	Coefficient (robust SE)	R ²	n
1	Microsoft AI diffusion, Q1 2026 (% of population aged 15 to 64)	Population ≥ 1 million	6.89 (0.48)	0.68	137
2	Microsoft AI diffusion, H1 2025 (% of population aged 15 to 64)	Population ≥ 1 million	5.97 (0.44)	0.66	137
3	Microsoft AI diffusion, Q1 2026 (% of population aged 15 to 64)	All economies in the release	6.79 (0.49)	0.66	139
4	Change in diffusion, H1 2025 to Q1 2026 (pp)	Population ≥ 1 million	0.92 (0.07)	0.57	137
5	ln Anthropic AI Usage Index, August 2025	Population ≥ 1 million; index > 0	0.78 (0.04)	0.70	136
6	Anthropic AI Usage Index (level), August 2025	Population ≥ 1 million; index > 0	0.57 (0.06)	0.54	136
<i>Growth regressions referred to in footnote 5 (coefficient on the Q1 2026 diffusion rate, per percentage point; growth in constant 2015 US\$)</i>					
7	Average annual growth of GDP per capita, 2023 to 2024 (pp)	Population ≥ 1 million	-0.056 (0.020)	0.04	140
8	Average annual growth of GDP per capita, 2015 to 2024 (pp)	Population ≥ 1 million	0.008 (0.015)	0.00	140
9	Average annual growth of GDP per capita, 2010 to 2019 (pp): placebo window that closes before generative AI	Population ≥ 1 million	-0.015 (0.014)	0.01	140

Source: Authors' calculations from Anthropic (2025), Microsoft AI Economy Institute (2026), and World Development Indicators.

Note: OLS with HC1 heteroskedasticity-robust standard errors in parentheses. Rows 1 to 6 regress the AI measure on ln GDP per capita; every coefficient is significant at the 1% level (t above 9). Rows 7 to 9 regress average annual growth of GDP per capita (log difference in constant 2015 US\$; World Development Indicators, September 2026 vintage) on the diffusion rate: $t = -2.8$ for 2023 to 2024, 0.6 for 2015 to 2024, and -1.1 for 2010 to 2019, a window that closes before the release of ChatGPT; none of the three estimates AI's effect (footnote 5). GDP per capita is in current US\$, 2024 (2023 where missing). Income-group means (FY2027) of Q1 2026 diffusion: low income 8.9%, lower-middle 11.0%, upper-middle 17.1%, high income 33.6% (weighted by population aged 15 to 64: 8.5%, 14.3%, 17.0%, 31.0%); of the August 2025 index: 0.08, 0.25, 0.69, and 1.97. Shares of working-age population and of estimated users (diffusion rate times population aged 15 to 64) in the 143 covered economies (Figure 1, panel D): 7% and 3% (low income), 36% and 29%, 40% and 38%, and 17% and 30% (high income). The provider's Q1 2026 estimates exceed its H1 2025 estimates by 16% to 20% in every group (unweighted), so the high-minus-low gap widened from 21 to 25 percentage points while the ratio stayed near 3.8. The Microsoft and Anthropic measures correlate at 0.74 across the 123 economies with both, which corroborates the gradient without validating either. The Microsoft release models sparse-traffic economies; lacking its country flags, we cannot exclude modeled observations, and the standard errors ignore shared estimation inputs. The Democratic Republic of Congo is matched to its own covariates; the provider reports identical values for the two Congos, and Anthropic indices for it are missing. Data file and code available from the authors.

Table A2. Risks of AI for poorer countries, organized by strength of evidence

No.	Risk and mechanism	Evidence type and date	Particular relevance for poorer countries	Feasible domestic response	International cooperation
1	Cyberattacks at lower cost and larger scale	Capability demonstrations and provider reporting (e.g., Anthropic Frontier Red Team 2026); documented incidents; realized aggregate losses not yet measured	Weak cybersecurity around banking, government databases, and mobile-money platforms	Basic cyber hygiene; incident reporting; AI-assisted defense; regional response capacity	Intelligence sharing; provider safeguards and disclosure norms
2	Biological misuse by hostile actors	Capability evaluations and expert assessment (Suleyman 2023; CSIS 2025 ; Amodei 2026; International AI Safety Report 2026); no realized AI-attributable event	Weak biosurveillance and health systems reduce detection and containment capacity	Health-system and pandemic preparedness; detection investment	Screening standards; model safeguards
3	Chatbot harms to mental health	Case reports, litigation, and early clinical reviews (e.g., Psychiatric Times 2025); systematic evidence thin as of the 2026 safety synthesis; assessment dated, not asserted as absence	Scarce mental-health services make both harms and assistive uses higher-stakes	Deployment standards for minors and crisis routing; local evaluation before scale	Liability frameworks; safety reporting
4	Cognitive offloading and deskilling	One randomized learning harm under unrestricted use (Bastani <i>et al.</i> 2025); survey evidence on reduced critical evaluation (Lee <i>et al.</i> 2025)	Larger classes and thinner teacher support weaken classroom management of AI use	Deployment with built-in guidance; assess competence without AI; teacher training	Shared evaluation standards
5	Amplification of misogyny and group bias	Audit and content studies; press synthesis of the literature (e.g., Financial Times 2026); prevalence estimates weak	High baseline gender inequality; thin local-language moderation and audit capacity	Local-language audits; redress channels	Audit access to models
6	Information-environment degradation	Persuasion experiments (Argyle 2025); platform studies; population-level effects uncertain (Harari 2024, as commentary)	Fragile information environments; election pressures; ethnic-tension risks	Provenance and disclosure rules; media capacity	Platform transparency
7	Surveillance and authoritarian use	Documented deployments; scenario analysis (Amodei 2026)	Weaker institutional checks in many settings	Procurement safeguards; legal limits; oversight	Export controls; rights frameworks
8	Labor displacement	Early market evidence: exposed freelancing demand fell about a fifth (Demirci <i>et al.</i> 2025); descriptive early-career declines in exposed occupations (Brynjolfsson <i>et al.</i> 2026); estimated earnings and hours effects within $\pm 2\%$ in the studied Danish occupations so far (Humlum and Vestergaard 2025a)	Economies that rely heavily on tradable services, and young labor-market entrants, may be especially vulnerable to job losses; exposure estimates depend on definitions and current technology	Sections 3.1 and 3.2 and the transitions row of Table 4	Evidence sharing on transitions
9	Severe, low-probability scenarios	Expert scenario assessments, not probability estimates (Amodei 2026; International AI Safety Report 2026 ; Pro-Human AI Declaration 2026)	Poorer countries are largely rule-takers in global AI governance	Monitoring capacity; participation in coalitions	A Global South voice in governance (UNCTAD 2025)

Source: Authors' compilation from the studies and reports cited in the table.

Note: Capability demonstrations, documented incidents, and measured losses are distinguished, and a statement that evidence is lacking describes the record at the time of writing, not the absence of risk. Human review as a safeguard itself requires skill, time, incentives, and authority. “Regulation and liability” alone is not an implementation plan. Press articles and web pages that are linked in the table are not repeated in the reference list.

Table A3. What each AI indicator can and cannot measure

Indicator	What it measures	Whose behavior	What it cannot tell us
1. Capability	What models can do on set tasks	Models, not people	Performance in the field; who benefits
2. Exposure	Share of work tasks AI could do or assist	Employed persons, by occupation	Actual use, displacement, or benefit; informal work
3. Access and readiness	Who could use AI; whether the complements and institutions are in place	Population; governments	Whether they do; quality of access; readiness is not adoption
4. Adoption	Who used AI in a period	Adults, internet users, or workers, by source	Intensity; comparisons across sources with different denominators
5. Intensity	How much, and for which tasks	Users of one provider	Population reach; other providers
6. Integration	Whether firms embed AI in production	Formal firms	Informal firms; worker and household outcomes
7. Causal effects	What AI changes, against a comparator	Study samples	Effects outside the setting; effects beyond the study's follow-up
<i>Measurement approaches for the outcomes above (not a stage of the chain)</i>			
8. AI-based welfare estimates	Poverty or wealth predicted from proxies	Populations covered by poverty maps	AI's impact; accuracy without validation (Table 2)

Source: By row: 1, capability evaluations (International AI Safety Report 2026); 2, Gmyrek *et al.* (2023, 2025), Cazzaniga *et al.* (2024), Eloundou *et al.* (2024), and Demombynes *et al.* (2025); 3, ITU and WDI connectivity series; 4 and 5, the sources in Table 1; 6, firm surveys and vacancy data (Acemoglu *et al.* 2022); 7, the studies in Section 3 and Table A5; 8, the studies in Section 3.5 and Table A6.

Note: Each measure answers a useful question, but the measures are not interchangeable. Measures of capability, exposure, access, or use do not establish causal effects, and only row 7 supports a claim about impact; AI-based welfare estimates require validation for their intended use. Row 8 is a measurement approach for the outcomes of the chain, listed for completeness; it is not a later stage through which adoption turns into welfare (Figure 2).

Table A4. Draft AI module for household and labor-force surveys, ready for piloting

No.	Domain	Core items (wording) and response categories	Respondents, skips, and design notes
1	Definition and awareness	“By AI tools we mean programs that answer questions or produce text, images, or voice on request, such as [local examples: ChatGPT, Gemini, Meta AI in WhatsApp, Copilot]. Have you heard of such tools?” Yes / No / Don’t know	All persons aged 15 and over. Examples fixed per country after cognitive testing; the definition is read to every respondent before any use item, and deliberate use of a chat assistant is distinguished from exposure to embedded recommendations or search features. Awareness does not set the route: every respondent continues to item 2

No.	Domain	Core items (wording) and response categories	Respondents, skips, and design notes
2	Use in the last 12 months and the last 7 days	“In the last 12 months, did you deliberately use such a tool, on your own device, on someone else’s, or with someone’s help?” Yes / No / Don’t know. “In the last 7 days, did you use one (a) for personal matters, (b) for work, (c) for study?” Yes / No for each, allowing overlap. “On how many of the last 7 days did you use one, for any purpose?” 0 to 7	All respondents; this item sets the route: twelve-month users continue to items 3 to 8, nonusers and respondents who answer don’t know to items 3 and 7; a respondent who reports seven-day use but denies twelve-month use is prompted once (“including the last seven days?”) and, if the answer stands, recorded as a user. Days of use are recorded for any purpose, with the purpose split carried by the seven-day items; seven-day recall for recent intensity, twelve-month item for occasional use; not directly comparable with quarterly telemetry unless the bridge is stated
3	Access and its conditions	“Could you use such a tool today if you wanted to?” On my own device / on a shared or borrowed device / with someone’s help / no / don’t know. “Do you or your household pay for any AI tool or subscription?” Yes / No / Don’t know	All respondents, including nonusers and past users without current access (a person who used a shared device last month may have no access today); paid access is asked directly rather than inferred from barriers
4	Tools, language, and mode of use, including embedded AI	“Which tools?” named list plus other. “Did you use AI features inside messaging, search, or office software?” Yes / No / Don’t know. “In which language do you mostly use them?” [local list]. “Do you usually use them on your own, or does someone help you?” On my own / with help / both	Users in the last twelve months, answering for that period (so that former users describe the tools they used); the tool list is fixed for cross-country comparison and versioned by year, with local examples documented; language and use with others’ help are asked directly of users
5	Tasks	“For which tasks?” writing, translation, advice, coding, images, studying, farming or business information, other (multiple response)	Users in the last twelve months, for that period; the task list mirrors the exposure indices for linkage
6	Verification and workplace integration	“Do you check the output before using it?” Always / sometimes / never, and “How?” (open, coded). “Does your employer provide an AI tool for your work?” Yes / No / Don’t know / Not applicable. “Does your employer require you to use one?” Yes / No / Don’t know / Not applicable. “Have you received training on AI tools at work?” Yes / No / Don’t know	Verification asked of all users in the last twelve months, for that period, including students and personal users; the three employer items asked only of employed users (provision and requirement kept as separate items; not applicable for the self-employed); links to firm size, formality, occupation, and hours
7	Barriers	“What keeps you from using such tools, or from using them more?” cost, data, language, trust, skills, no need, not allowed, other / don’t know	All respondents, including nonusers, in a short route
8	Perceived changes in time and in earnings	“Has using AI changed the time you spend on your tasks?” less / same / more / don’t know. “Has using AI changed what you earn?” less / same / more / don’t know / not applicable, with optional amounts	Users; two separate items because time and earnings can move in opposite directions; recorded as perceptions, never as effects, and kept apart from measured outcomes

Source: Authors’ elaboration.

Note: Responses link to existing survey variables on occupation, informality, gender, age, location, firm size, hours, earnings, and household welfare. Each selected person aged 15 and over self-reports; flagged proxy answers are accepted only for members absent throughout fieldwork. Items 2 and 3 use the periods as worded; items 4 to 6 and 8 refer to the last twelve months. Items 1 to 8 are a fixed core; agencies may append optional detail. Wording, translations, and routes will be published with the pilot draft. The module identifies informal workers but does not replace firm panels. Comparisons with official sources keep their own denominators (Colombia’s rate is among internet users aged five and older), and a comparison with the providers’ quarterly estimates of the share of users needs a three-month item or a stated assumption. Self-assessed effects never substitute for measured outcomes. Precedents: Colombia’s ENTIC Hogares 2024 (DANE 2025) and Brazil’s innovation survey (IBGE 2025).

Table A5. Candidate studies and inclusion decisions: intervention lens (Table A7, rows 1a to 3)

No.	Study	Setting (worker geography)	Design	Outcome domain	Headline estimate (as published)	Inclusion decision
1	Noy and Zhang (2023)	United States (per the trial registration), online professionals	RCT	Task performance (writing)	Quality about +0.4 SD (0.41 in source Fig. 1B; 0.45 in the text); second-task time 11 minutes lower (0.75 SD) against a control mean of 27 minutes	Include; the published article's estimate and interval are used (Table A8)
2	Peng <i>et al.</i> (2023)	Upwork freelancers, mostly in India and Pakistan (mixed geography)	RCT	Task performance (coding)	Completion time −55.8% [21, 89] among the 35 per arm who completed the task (95 assigned: 45 treated, 50 control)	Include; contrast conditional on completion, not intention to treat
3	Becker <i>et al.</i> (2025)	Experienced open-source developers (countries not reported)	RCT	Task performance (coding)	19% slower with early-2025 tools [2, 39]; participants believed 20% faster	Include: key negative estimate; read with the METR (2026) design update (late-2025 follow-up: −18% and −4% time, both insignificant)
4	Cui <i>et al.</i> (2026)	Microsoft (developers mainly in the United States), Accenture (offices in Southeast Asia), and an anonymous firm (location not stated); assignment by developer or team	Field experiments (3)	Task performance (coding)	Completed pull requests per developer-week +26.1% (SE 10.3)	Include; an IV estimate for adopters, not an intention-to-treat effect; the intention-to-treat estimate, for two of the three firms in an early period, enters a sensitivity (Tables A8 and A12); verified in the Management Science article, whose estimates match the working paper
5	Brynjolfsson <i>et al.</i> (2025)	≈89% outside the US, mainly the Philippines	Staggered rollout	Task performance (support)	+15% resolutions per hour; novices gain most	Include; worker geography coded as mainly the Philippines (poorer-country evidence)
6	Dell'Acqua <i>et al.</i> (2026)	Global consultancy	Field experiment	Task performance (consulting)	+12.2% tasks; quality +34% and +30% in the two AI arms inside the frontier (final article); −19 pp accuracy outside	Include (final article: arm-level means and standard deviations from source Table 2, quality from source Table 4)
7	Bastani <i>et al.</i> (2025), both arms	Türkiye	RCT	Task performance (practice) and learning without the tool (exam)	Practice +48% (base), +127% (tutor); unaided exam −0.054 (SE 0.022) for the base arm and −0.004 (SE 0.013) for the tutor arm on a [0, 1] grade, i.e. −17% and ≈ 0 of the control mean	Include: the hint-only-tutor exam arm is pooled (learning) and the unrestricted exam arm shown but not pooled, because the two share a control group; practice arms descriptive (task quality); PNAS version
8	De Simone <i>et al.</i> (2025)	Nigeria	RCT	Learning without the tool	+0.31 SD composite; +0.24 SD English (facilitated package)	Include; effect attributed to the package, not the model
9	Henkel <i>et al.</i> (2024)	Ghana	RCT	Learning without the tool	+0.36 SD math over eight months (version 2 full text; 0.37 in the earlier version)	Include; 11-school cluster trial with student-level inference, variance widened by an assumed design effect
10	Kestin <i>et al.</i> (2025)	United States	RCT (crossover)	Learning without the tool	+0.63 SD versus active learning	Include; whether AI was unavailable during the post-lesson quiz is not reported

No.	Study	Setting (worker geography)	Design	Outcome domain	Headline estimate (as published)	Inclusion decision
11	Otis <i>et al.</i> (2026)	Kenya	RCT	Business (profits and revenues)	≈ 0 average; top performers $\approx +15\%$; bottom $\approx -8\%$	Include (business outcomes); published version
12	Humlum and Vestergaard (2025a)	Denmark	Surveys $\times$ registers	Labor market (earnings, hours)	≈ 0 ; confidence intervals exclude effects above 2% two years after launch (latest version; earlier version 1%)	Include (labor market); self-reported time savings excluded from the task-performance domain
13	Demirci <i>et al.</i> (2025)	Global platform	Difference-in-differences	Labor market (task demand)	Posts -21% (writing, coding); -17% (image creation)	Include (structured synthesis); platform selection discussed
14	Korom <i>et al.</i> (2025)	Kenya	Quality-improvement comparison	Task performance (clinical quality)	Diagnostic errors -16% ; treatment errors -13%	Sensitivity analyses only: non-random
15	Brynjolfsson <i>et al.</i> (2026)	United States	Descriptive payroll	Context (employment)	Relative early-career decline in most-exposed occupations	Context only; not pooled
16	Muralidharan <i>et al.</i> (2019)	India	RCT	Learning (pre-generative CAL)	$+0.37$ SD math; $+0.23$ SD Hindi	Benchmark from earlier AI, searched separately; not in the generative-AI pool
17	Angrist <i>et al.</i> (2022)	Botswana	RCT	Learning (non-AI)	$+0.12$ SD at very low cost	Exclude: digital precedent; retained as cost-effectiveness benchmark
18	Fabregas <i>et al.</i> (2025)	East Africa	Pooled RCTs (6)	Agriculture (non-AI)	Modest take-up gains at very low cost	Exclude: digital precedent
19	Vaccaro <i>et al.</i> (2024)	106 experiments to mid-2023	Meta-analysis	Benchmark	$g = +0.64$ versus human alone; -0.23 versus best of either	Starting list of studies and benchmark; contrasts kept separate; not a pooled row
<i>Added by the first round of web searches</i>						
20	LearnLM Team and Fab AI (2026)	Sierra Leone, 12 junior secondary schools	Cluster RCT (48 classrooms)	Learning without the tool	Math $+0.258$ SD [0.027, 0.488] after eight weeks; $+0.380$ SD for completing at least 12 hours (IV, treatment on the treated)	Include (pooled); trial run by the tool's developer, flagged; preregistered
21	Fischer <i>et al.</i> (2025)	Germany, university lab	RCT	Learning without the tool	$+0.23$ SD overall; $+0.34$ SD versus control with the tutor available from the start ("unrestricted access"); $+0.13$ SD, not significant, with access after 10 minutes of reading	Include (pooled); immediate test after a 25-minute session
22	Fütterer <i>et al.</i> (2026)	Germany, grades 7–9	RCT (three arms)	Learning without the tool	No additional knowledge effect of either pedagogical-prompt condition over standard ChatGPT; all three arms used AI	Shown, not pooled: a comparison between AI designs, not of AI against none; no moments published
23	Agweyu <i>et al.</i> (2026)	Kenya, 16 primary care facilities	Cluster RCT (clinical officers)	Clinical quality (patient outcomes)	Treatment failure aOR 0.77 [0.55, 1.08]; documentation quality aOR ≈ 1.7	Descriptive; kept separate from learning; supersedes the quality-improvement comparison as the causal evidence on AI Consult

No.	Study	Setting (worker geography)	Design	Outcome domain	Headline estimate (as published)	Inclusion decision
24	Qazi <i>et al.</i> (2026)	Pakistan, physicians	RCT (single-blind)	Task performance (clinical reasoning)	Diagnostic reasoning +27.5 pp [22.8, 32.2] with GPT-4o after AI-literacy training	Descriptive (quality with AI); poorer-country evidence
25	Cruces <i>et al.</i> (2026)	Argentina (UMIC), online sample aged 25 to 45; preregistered (AEARCTR-0016607)	RCT (online)	Task performance (business problem solving)	Education gap in task performance 0.55 SD without AI, 0.14 SD with AI; 1,174 completers of 1,795 assigned; partial retention without AI	Descriptive; average-effect moments not published
26	Idan and Anand (2026)	Early-career knowledge workers in a self-study task; participant country not stated in the source (coded as online, country not stated)	RCT	Task performance (technical self-study task)	Gains concentrated among high AI-interaction competence; built-in guidance reduces variance	Descriptive; moments not published
27	Dillon <i>et al.</i> (2025)	66 firms, 7,137 workers	Field RCT	Work patterns (time use)	Email time –2 hours per week among adopters; no change in task composition	Descriptive; not an output-per-hour outcome
28	Wang <i>et al.</i> (2025)	United States, Title I schools	RCT (human tutors)	Task performance (tutoring)	Student mastery 66% versus 62% with AI-assisted tutors	Descriptive; comparator is human tutoring without AI
29	LearnLM Team and Eedi (2025)	United Kingdom, five schools	RCT	Task performance (tutoring)	AI tutoring at least as effective as human tutoring; novel problems 66.2% versus 60.7%	Descriptive; comparator is human tutoring, not no-AI
30	Hui <i>et al.</i> (2024)	Upwork, global	Difference-in-differences	Labor market (jobs, earnings)	Jobs –2%; monthly earnings –5.2% for exposed freelancers	Include (structured synthesis)
31	Qiao <i>et al.</i> (2025)	Online labor platform, global	Difference-in-differences	Labor market (work volume, earnings)	Translation: transaction volume –9.0%, earnings –29.7%; web development: +6.4% and +66.5%; larger US gains in volume and, at the 10% level only, in earnings	Include (structured synthesis); 2025 revision read in full
32	Ferman <i>et al.</i> (2021)	Brazil, high schools	RCT	Learning (pre-generative automated essay feedback)	Higher college-exam essay scores; human graders add no effect	Benchmark from earlier AI, searched separately
<i>Added by the second round of web searches, backward citation chasing, and supplementary checks</i>						
33	Haslberger <i>et al.</i> (2025)	United Kingdom (HIC), working-age online panel with ChatGPT accounts; 1,041 participants (504 treated, 537 control); preregistered (OSF)	Preregistered RCT (encouragement)	Task performance	Time saved and quality gains on all three tasks (for example, email task time –134 seconds, SE 15; quality +1.26 points on a 5-point scale, SE 0.12); interactions by gender, age, education, and occupation insignificant in the main IV models; the few significant intention-to-treat interactions favor younger, university-educated, or professional respondents	Verified (full text, March 23, 2025 version, with supplementary Tables S19 to S22); pooled in task performance at the intention-to-treat effect on total time over the three tasks, with a variance bound (Table A8); the IV estimates for compliers are descriptive; counterexample for the skill gradient (Table A9)

No.	Study	Setting (worker geography)	Design	Outcome domain	Headline estimate (as published)	Inclusion decision
34	Dell'Acqua <i>et al.</i> (2025)	Procter & Gamble, 776 professionals in Europe and the Americas (mixed geography)	Preregistered RCT	Task performance (product innovation)	Log time -0.366 (SE 0.070) for individuals with AI versus without; quality $+0.37$ SD; individuals with AI matched teams without AI	Verified (NBER Working Paper 33641, April 2025); pooled in task performance (tool versus no tool, individuals) as a difference in mean log time; the 16.4% reduction in average time in an assumption-based sensitivity (Table A12)
35	Ju and Aral (2026)	United States (HIC), online workers on Prolific, 2,234 participants	RCT	Task performance (ads produced)	Human–AI teams versus human–human teams: $+1.95$ ads per participant (SE 0.20), about 50% more	Verified; not pooled: the comparator is a human collaborator rather than work without a tool (task-performance rule, Appendix A.1); recorded in Table A8
36	Gambacorta <i>et al.</i> (2026)	Ant Group, China (UMIC), programmers	Matched comparison after tool rollout	Task performance (tasks completed; code output)	Tasks completed $+0.220$ log points (SE 0.043); code output up by more than 50%, significant for junior but not senior staff (no test of the difference)	Verified (Journal of Financial Stability article, which supersedes BIS Working Paper 1208); pooled in task performance at tasks completed, a non-randomized rollout like row 5 (Table A8); code output not used because generated code inflates it
37	Fang <i>et al.</i> (2026)	Cross-border online retail platform (mixed geography)	Randomized field experiments, seven workflows	Business (sales, conversion)	Sales effects by workflow from -4.5% to $+16.3\%$, several not statistically significant	Verified (arXiv:2510.12049); business and labor-market outcomes, structured synthesis: firm sales by workflow, not pooled by rule (Table A8)
38	Gallegos (2026)	Chile (HIC), university econometrics course, ≈ 300 students	Two RCTs	Learning (final exam)	Guidance on learning-oriented use $+0.22$ SD (SE 0.10) on final grades, $+12.8$ pp passing; encouragement alone $+0.05$ SD (SE 0.12); both against AI access alone	Verified (IZA DP 18513); shown, not pooled: both arms had AI access, so the contrasts compare ways of using AI rather than AI against none (Figure 3, panel B)
39	Contractor and Reyes (2026)	United States (HIC), college undergraduates, proctored sessions	RCT, proctored sessions	Learning without the tool	$+0.27$ SD [0.03, 0.51] on an unaided test one week later; essay written with or without off-the-shelf AI (search allowed)	Verified (IZA DP 18792); shown, not pooled: unrestricted general-purpose assistance against conventional search, a different comparison from the pooled one (Figure 3, panel B)
40	Oreopoulos and Low (2026)	United States, 18 middle schools	Two-year cluster RCT	Learning (mathematics)	Per-school-year effect $+0.066$ control-group SD (SE 0.043), stacking annual contrasts; year 2 $+0.116$ (0.053); median activated student in year 2 messaged the tutor on a third of practice days; summer-persistence estimate uninformative (0.005 SD, SE 0.030)	Verified (NBER Working Paper 35620); pooled in learning (Khan Academy practice with the tutor versus usual remedial instruction; the tutor's own contribution is not identified)

No.	Study	Setting (worker geography)	Design	Outcome domain	Headline estimate (as published)	Inclusion decision
41	Ma <i>et al.</i> (2026)	Rural Sichuan, China (UMIC): 105 primary schools in eight low-income counties; 8,647 students in grades 4 to 6 (315 classes)	Cluster RCT, three arms, classes randomized within strata; registered (AEARCTR-0007888)	Learning (mathematics; pre-generative adaptive practice)	Adaptive practice versus non-adaptive practice -0.040 SD (SE 0.025) and versus no computer-assisted learning -0.040 SD (SE 0.027) after 11 weeks in spring 2021 (source Table 1, column 2); non-adaptive versus none 0.000 (SE 0.025)	Verified (CESifo Working Paper 12837; read in full): excluded from the pool at criterion (i), because the tool is an adaptive system built on item response theory and deep knowledge tracing, not a generative model; kept as a benchmark from earlier, non-generative AI with rows 16 and 32 (Section 3.3; Table 3)
42	Jagnani and Pande (2024)	Bihar, India (LMIC), 319 communities (160 treated, 159 control), 5,582 households surveyed	Cluster RCT	Household preparedness, health, economic well-being (AI flood forecasts delivered as alerts)	Alerts received $+2.4$ (2022) and $+4.4$ (2023) per household; in severely flooded communities, adaptation and health indices $+0.18$ SD and medical spending -30% ; no detectable effect on economic well-being	Verified (full text, April 3, 2024 version); background evidence: predictive forecasting delivered through volunteers, not a generative tool; not pooled
43	Rounding <i>et al.</i> (2026)	Kenya, Indonesia, and the Netherlands (multi-country), 249 physicians	RCT	Task performance (clinical reasoning, simulated cases)	$+18.0$ pp Kenya [12.7, 23.2]; $+10.7$ pp Indonesia [5.7, 15.7]; $+7.2$ pp Netherlands [3.7, 10.7], all $p < 0.001$; GPT-4o; controls without internet or guideline access	Verified (full text, published version); descriptive (task quality); within-study cross-setting contrast; not a patient outcome
44	Cheng <i>et al.</i> (2024)	Beijing, China (UMIC): one kindergarten in a less-developed district; 148 children, mean age 5.8	RCT, 2×2 factorial (dialogic or plain reading; chatbot or human partner); one session	Learning (story comprehension and word learning, tested immediately)	Dialogic reading raised comprehension and word learning with either partner, with no significant difference in its benefit between the chatbot and the human partner; for story comprehension, chatbot-led dialogic reading helped children with higher vocabulary more	Verified (<i>Journal of Experimental Child Psychology</i> 240; read in full): out of scope at criterion (i), because the chatbot classifies spoken answers by semantic similarity and replies from a fixed script rather than generating text; it is compared with a human partner reading the same script; not counted in Table 3; listed by Burneo <i>et al.</i> (2026)
45	Oreopoulos <i>et al.</i> (2026)	United States (HIC), Hamilton County middle schools, more than 6,000 students	RCT (individual; 2×2 factorial)	Learning (mathematics practice; unaided delayed test)	AI tutor versus practice without AI, without a mastery requirement: practiced-minus-unpracticed number correct (each out of two questions) -0.014 (SE 0.034); with the requirement $+0.026$ (0.034)	Verified (NBER Working Paper 35621); sensitivity only: the outcome's standard deviation is not reported, so it enters one row of Table A12 at a calibrated value (Table A8)
Resolved from first-round records once their full texts were obtained (September 2026)						
46	Koo (2025)	Participants proficient in English, Arabic, or Chinese, recruited worldwide (multiple countries); preregistered	RCT (online), 480 participants	Task performance (email quality)	AI-generated content improved completeness and relevance similarly across languages; actionability improved 0.59 SD less in Arabic and 0.47 SD less in Chinese than in English	Verified (Scientific Reports); descriptive: quality ratings of work done with pre-generated GPT-4 content rather than live use (Table A8); language contrast in Table A9

No.	Study	Setting (worker geography)	Design	Outcome domain	Headline estimate (as published)	Inclusion decision
47	Vanzo <i>et al.</i> (2025)	Italy (HIC), technical high school, four classes	RCT (individual, stratified by grades)	Learning without the tool (English as a second language)	Learning gain $d = 0.251$ (one-sided $P = 0.156$) after eight weeks of GPT-4 homework sessions (six curriculum units); third-year classes 0.603, fifth-year -0.004 ; 76 students	Verified (ACL 2025); pooled in learning, with the SE derived from the reported P (Table A8); whether AI was unavailable during the tests is not reported
48	Teutloff <i>et al.</i> (2025)	Online freelancing platform, global	Difference-in-differences by skill cluster	Labor market (job postings)	Postings for substitutable skills -24% relative to unaffected clusters (raw fall 7%); falls in individual clusters of up to 59% , some not significant; complementary clusters mixed	Verified (Journal of Economic Behavior and Organization); business and labor-market outcomes, structured synthesis
49	Tekic <i>et al.</i> (2026)	Russia (HIC), freelance platform (Profi.ru), 11,894 freelancers in the full panel	Difference-in-differences with panel fixed effects	Labor market (participation, orders, income)	Among 2,190 continuing providers, monthly participation relative to offline services -6.7 pp (word processing), -8.7 pp (translation), -4.4 pp (design); no comparable decline in programming, marketing, or web development	Verified (Technology in Society); business and labor-market outcomes, structured synthesis; access to frontier tools constrained
50	Kauhanen and Rouvinen (2025)	Finland (HIC), all wage earners	Synthetic difference-in-differences by occupational exposure	Labor market (wages, employment)	No statistically significant change in wages or employment two years after ChatGPT's release	Verified (Applied Economics Letters); business and labor-market outcomes, structured synthesis
51	Chen <i>et al.</i> (2025)	United States (HIC), Current Population Survey	Synthetic difference-in-differences by occupational exposure	Labor market (earnings, unemployment)	Weekly earnings about \$89 higher in exposed occupations; unemployment about 0.2 pp higher, which the authors call negligible	Verified (arXiv:2509.15510); business and labor-market outcomes, structured synthesis
Added on September 29, 2026: the Pakistan courts rollout and seven trials of earlier, non-generative AI in poorer countries from the WDR 2026 health review (Sautmann and Weinert 2026), retained as benchmarks						
52	Mehmood <i>et al.</i> (2026)	Pakistan (LMIC), 118 district courts; 1,559 judges, about half of the country's trial judges	RCT, three arms (a GPT-4 assistant grounded in Pakistani case law, with targeted or with generic training; generic training only); court outcomes by district exposure to the targeted-training arm	Public services (case resolution)	Resolved cases $+6.3\%$ at the median district share of judges given the assistant with targeted training (about 1,848 cases a year); appeals per resolved case did not rise (if anything, they fell slightly)	Verified (CEPR Discussion Paper DP21783, July 2026; read in full); descriptive record (Table A8); counted in Table 3; not pooled: a district-level dose response rather than a per-judge assignment effect

No.	Study	Setting (worker geography)	Design	Outcome domain	Headline estimate (as published)	Inclusion decision
53	Lin <i>et al.</i> (2019)	China (UMIC), five eye clinics; 350 children (700 eyes)	RCT (children randomized 1:1 to diagnosis by the AI platform or by senior consultants)	Health (diagnosis of childhood cataract; earlier AI)	Diagnostic accuracy 87.4% with the AI platform versus 99.1% for senior consultants, and treatment decisions 70.8% versus 96.7%; mean diagnosis time 2.8 versus 8.5 minutes	Verified (EClinicalMedicine; read in full); benchmark from earlier AI (Table 3); process outcome; not pooled
54	MacPherson <i>et al.</i> (2021)	Malawi (LIC), one urban primary clinic in Blantyre; 1,462 adults with cough (473 usual care, 492 HIV testing, 497 HIV testing with computer-aided X-ray)	RCT (three arms: usual care; universal HIV testing; HIV testing plus digital chest X-ray read by software)	Health (tuberculosis screening; earlier AI)	Time to tuberculosis treatment, X-ray arm against usual care: hazard ratio 2.86 [1.04, 7.87]; US\$4,620 per QALY gained, not cost-effective at the study's US\$1,200 threshold	Verified (PLOS Medicine; read in full); benchmark from earlier AI; the only cost-effectiveness analysis among the seven earlier-AI health trials; not pooled
55	Liu <i>et al.</i> (2021)	China (UMIC), one tertiary emergency department in Beijing; 17,072 visits over 60 days (16,775 analyzed)	30 days randomly selected for the standard protocol and the other 30 for the triage algorithm (March to April 2020)	Health (emergency triage; earlier AI)	Life-threatening mis-triage 0.9% versus 1.2% ($p = 0.037$); calls to supervising physicians up 24%	Verified (Scientific Reports; read in full); benchmark from earlier AI; days randomized, not patients; process outcome; not pooled
56	Mathenge <i>et al.</i> (2022)	Rwanda (LIC), four clinics in and around Kigali; 275 patients with referable diabetic retinopathy (136 intervention, 139 control)	RCT (immediate AI grading with same-day feedback versus human-grader results three to five days later)	Health (referral uptake; earlier AI)	Referral within 30 days 51.5% versus 39.6% ($p = 0.048$; adjusted OR 1.74 [1.05, 2.88])	Verified (Ophthalmology Science; read in full); benchmark from earlier AI; not pooled
57	Abramoff <i>et al.</i> (2023)	Bangladesh (LMIC), one eye hospital in Rangpur with three retina specialists; 105 clinic days (51 intervention, 54 control), 993 patients	Cluster RCT (clinic days randomized and concealed)	Health (clinic throughput; earlier AI)	Completed patient encounters per specialist hour 1.59 [1.37, 1.80] with autonomous AI eye screening versus 1.14 [1.02, 1.25] ($p < 0.001$); no patient outcomes measured	Verified (npj Digital Medicine; read in full); benchmark from earlier AI; process outcome; not pooled

No.	Study	Setting (worker geography)	Design	Outcome domain	Headline estimate (as published)	Inclusion decision
58	Adedinsewo <i>et al.</i> (2024)	Nigeria (LMIC), six hospitals; 1,232 pregnant and postpartum women randomized, 1,195 analyzed	Pragmatic open-label RCT (AI-guided screening with a digital stethoscope and 12-lead ECG versus usual care)	Health (detection of cardiomyopathy; earlier AI)	Detection of left-ventricular systolic dysfunction 4.1% versus 2.0% (OR 2.12 [1.05, 4.27]); 12 deaths in the intervention arm against 3 in the control arm, which the authors attribute tentatively to closer follow-up and uneven reporting of deaths	Verified (Nature Medicine; read in full); benchmark from earlier AI; not pooled
59	Yu <i>et al.</i> (2025)	Rural China (UMIC), four districts of one county; 63 village doctors (26 intervention, 37 control)	Cluster RCT (districts randomized in pairs: an AI-assisted portable slit lamp with training versus training alone, after a one-year baseline)	Health (cataract referral and surgery; earlier AI)	Cataract referrals: adjusted incidence rate ratio 2.67 [1.57, 4.55]; surgeries 5.85 [2.54, 13.47]; detection 1.19 [0.94, 1.50], not significant	Verified (Current Eye Research; read in full); benchmark from earlier AI; two districts per arm, so the estimates rest on a before-after comparison of two pairs; not pooled

Source: Authors' compilation from the studies listed.

Note: Decisions form the study list (procedure: Appendix A.1). Rows 33 to 42 come from the second round of web searches, row 44 from backward chasing of Burneo *et al.* (2026) (Table A11, panel B), rows 43 and 45 from the supplementary checks, rows 46 to 51 from first-round records resolved once their full texts were obtained, and rows 52 to 59 from a final check on September 29, 2026: row 52 from a discussion paper released in July 2026, and rows 53 to 59 from the study list of Sautmann and Weinert (2026), retained as benchmarks from earlier AI (Table 3, row 4). Twelve first-round records, none of them on the intervention lens, remain unresolved in the search log. Task-performance, learning, and business and labor-market outcomes are never pooled together; the effect sizes used are in Table A8. All rows were read in full; rows 41, 44, and 53 to 59 use tools that are not generative and enter no pooled estimate (rows 41 and 53 to 59 count in Table 3 as benchmarks; row 44 is out of scope). De Simone *et al.* (2025) was excluded from the main pooling of Burneo *et al.* (2026) for attrition above 30%.

Table A6. Candidate studies and inclusion decisions: measurement lens (Table A7, row 4)

No.	Study	Setting	Predictor and target (Table 2 row)	Validation design and metric, as published	Inclusion decision
1	Jean <i>et al.</i> (2016)	Nigeria, Tanzania, Uganda, Malawi, Rwanda	Daytime satellite imagery with transfer learning; asset wealth and consumption (rows 1 and 3)	Cross-validated cluster-level R^2 of 0.55 to 0.75 for assets; up to three-quarters of local variation explained	Include (rows 1 and 3)
2	Yeh <i>et al.</i> (2020)	23 African countries	Satellite imagery; asset wealth, levels and changes (rows 1 and 3)	Held-out countries: R^2 0.70 for village-level wealth, 0.83 at district level; changes over time R^2 0.35 (villages) and up to 0.50 (districts)	Include; change-detection metrics recorded separately
3	Chi <i>et al.</i> (2022)	135 poorer countries	Satellite, connectivity, and other geospatial data; relative wealth index at 2.4 km (row 1)	Trained on surveys from 56 countries; R^2 0.72 against independent census data from 15 countries	Include (row 1)

No.	Study	Setting	Predictor and target (Table 2 row)	Validation design and metric, as published	Inclusion decision
4	Blumenstock <i>et al.</i> (2015)	Rwanda	Mobile phone metadata; individual and area wealth (row 1)	Cross-validated individual prediction ($r = 0.68$); cluster ($r = 0.79$) and district ($r = 0.92$, 30 districts) aggregates against survey data	Include (row 1); displayed in Figure 4 as a within-study contrast across units, shown as squared correlations
5	Steele <i>et al.</i> (2017)	Bangladesh	Phone and satellite data combined; several poverty measures (rows 1 and 3)	Random 20% test set: out-of-sample R^2 0.76 for an asset-wealth index, 0.27 for income, and 0.25 for a poverty-probability index (phone and satellite data combined)	Include (rows 1 and 3); displayed in Figure 4 as a within-study contrast across targets
6	Pokhriyal and Jacques (2017)	Senegal	Combined phone and environmental data; multidimensional poverty (row 1)	Spatially cross-validated across 552 communes against census-based deprivations: Pearson r 0.91 and rank correlation 0.88 for the index with phone and environmental data combined (0.89 and 0.84 for each source alone); 0.49 for the health dimension	Include (listed only); not displayed: its target is a multidimensional poverty index at commune level, which no displayed study shares
7	Engstrom <i>et al.</i> (2022)	Sri Lanka	High-resolution imagery features; small-area consumption and poverty rates (rows 1 and 3)	Cross-validated R^2 0.59 to 0.61 for small-area poverty rates and log consumption across 1,291 divisions, against census-based small-area estimates	Include (rows 1 and 3); displayed in Figure 4 as a level benchmark for consumption poverty
8	Aiken <i>et al.</i> (2022)	Togo	Phone metadata; eligibility ranking for emergency cash (row 2)	Rural evaluation (2020 phone survey of mobile subscribers in the 100 poorest cantons; proxy-means-test welfare; poorest 29% eligible): exclusion error 53% versus 59 to 78% for the feasible geographic and occupational alternatives. Nationwide simulation (2018 to 2019 household survey; consumption): 50% versus 52 to 76% for most feasible alternatives, and 48% for an optimized occupation-based rule that assigns all transfers to agricultural workers; relative to methods requiring a comprehensive social registry, which does not exist in Togo, the phone-based approach increases exclusion errors by 9 to 35% (reported relative terms)	Include (row 2); plotted setting by setting with the study's own numbers, including the optimized occupational rule; no registry value is derived
9	Aiken <i>et al.</i> (2023)	Afghanistan	Phone metadata; ultra-poor targeting (row 2)	Errors versus survey-based targeting; who is missing from phone data	Include (row 2); not displayed: it reports classification accuracy for ultra-poor status among program applicants against a survey-based benchmark, not exclusion error at a program budget
10	Aiken <i>et al.</i> (2025a)	Togo	Digital traces as outcome data in a randomized transfer	Agreement between survey-measured and trace-measured impacts	Evaluation using AI-predicted outcomes (Section 3.5); not pooled
11	Ratledge <i>et al.</i> (2022)	Uganda	Satellite-derived asset wealth as the outcome of electrification	Machine-learned outcomes with corrected inference	Evaluation using AI-predicted outcomes (Section 3.5); not pooled
12	Huang <i>et al.</i> (2021)	Kenya	Satellite imagery and deep learning; housing outcomes of cash transfers	Learned outcomes evaluated against experimental ground truth	Evaluation using AI-predicted outcomes (Section 3.5); not pooled

No.	Study	Setting	Predictor and target (Table 2 row)	Validation design and metric, as published	Inclusion decision
13	Burke <i>et al.</i> (2021)	Global	Review of satellite-based measurement	Compiles reported accuracy across studies	Starting list of studies for the validation search; not a study row
<i>Added by the first round of web searches</i>					
14	Marty and Duhaut (2024)	59 countries, 63,854 survey clusters	Satellite, connectivity, and other public and private big data; asset wealth, levels and changes (rows 1 and 3)	Levels: median r^2 0.55 at cluster level and 0.59 at district level (pooled 0.64 and 0.74); changes between rounds: median 0.04 and 0.06	Include (rows 1 and 3); verified
15	Pettersson <i>et al.</i> (2023)	Africa, 1990–2019	Time series of satellite imagery with recurrent networks; asset wealth (rows 1 and 3)	R^2 0.72 across held-out countries; 0.75 across held-out time spans	Include (row 1; held-out time spans recorded separately); verified
16	Aiken <i>et al.</i> (2025b)	Bangladesh (Togo validation)	Phone metadata versus proxy means test and community-based targeting; targeting at a fixed budget (row 2)	Precision and recall at the budget: proxy means test 52% (AUC 0.82), phone-based 32% (0.61), community-based 26% (0.58); phone-based cheaper below about \$10 to \$50 per screened household	Include (row 2); verified; the survey-based alternative wins on accuracy
17	Smythe and Blumenstock (2022)	Nigeria	High-resolution poverty maps for geographic microtargeting (row 2)	Targeting performance of map-based allocation against survey-based benchmarks	Include (row 2); not displayed: it reports the performance of map-based geographic allocations, a simulated targeting outcome rather than a validation statistic on either axis
18	van der Weide <i>et al.</i> (2024)	Malawi	Remote-sensing predictors in a spatial error regression, not machine learning; poverty map validated against a census-based map (rows 1 and 3)	Poverty rates from a survey and remote-sensing data against census-based small-area estimates across traditional authorities: R^2 0.83; correlation 0.91 with exact survey coordinates and 0.88 with public displaced coordinates; large gaps in a few areas	Include (rows 1 and 3); not displayed: Figure 4 shows one small-area benchmark against census-based estimates, filled by Engstrom <i>et al.</i> (2022, row 7)
19	Corral <i>et al.</i> (2025)	Mexico	Machine learning on remotely sensed data versus traditional small-area estimation (rows 1 and 3)	Against census ground truth, ML underperforms the conventional benchmark and is more biased for the poorest areas	Include (rows 1 and 3); not displayed: it reports bias and mean squared error against census ground truth; error metrics that do not map to R^2
20	Lee and Braithwaite (2022)	25 Sub-Saharan African countries	Geospatial data and machine learning; village-level poverty maps (row 1)	Average R^2 0.88 within country (five-fold cross-validation) and 0.86 across countries (leave-one-country-out), against survey-cluster wealth (published version)	Include (row 1); not displayed: the authors cross-validate with separate training and test data, but their refining process uses the model's earlier estimates and each cluster's survey wealth to narrow the cluster's likely location, so the R^2 values are not directly comparable with those displayed
21	Barriga-Cabanillas <i>et al.</i> (2025)	Haiti	Phone metadata predictions of expenditure as outcome data for a cash transfer	CDR-based expenditure predictions far less accurate than wealth predictions; CDR-based impact estimates insignificant where survey estimates are positive	Evaluation using AI-predicted outcomes (Section 3.5); the cautionary case; not pooled

No.	Study	Setting	Predictor and target (Table 2 row)	Validation design and metric, as published	Inclusion decision
22	Merfeld <i>et al.</i> (2025)	Burkina Faso, Madagascar, Malawi, Mozambique, Sri Lanka, Tanzania, and Viet Nam	Satellite-derived geospatial indicators; small-area asset index in all seven countries and poverty rates in Malawi and Tanzania (rows 1 and 3)	100 two-stage samples drawn from unit-level census data in each country; the standard small-area model (empirical best predictor) and three tree-based learners against census benchmarks: average correlation 0.83 for the standard model and 0.87 to 0.89 for Cubist and XGBoost; coverage of 95% intervals 78% (parametric bootstrap) against 84% to 89% (two-stage residual bootstrap); accuracy lower in unsampled areas	Include (rows 1 and 3); not displayed: accuracy is an average correlation over simulated samples rather than an out-of-sample R^2 on one held-out set; the one study that validates uncertainty estimates

Source: Authors' compilation from the studies listed.

Note: Decisions form the study list (Appendix A.1); the first round of web searches also found two further reviews used as starting points for citation searching (Hall *et al.* 2023; Newhouse 2024). The list includes 17 validation studies (rows 1 to 9, 14 to 20, and 22), 4 evaluations that use AI-predicted outcomes (rows 10 to 12 and 21; Section 3.5), and one such review (row 13). Figure 4 displays 10 validation studies selected by one rule (rows 1 to 5, 7, 8, and 14 to 16); the other seven validation rows say why they are not shown. Metrics are compared only within a Table 2 row, never averaged. Language-model instruments for survey coding and translation show promise but are not listed here, as no validation study with representative ground truth was found.

Table A7. Outcome domains: the comparison each shares, what the exploratory summary finds, and what remains unknown

Outcome domain (scale)	Comparison shared by the pooled studies; k	Exploratory result (modified Hartung–Knapp 95% interval) and what it means	What remains unknown	Decision it informs (Table 4)
1a. Task performance with AI (% change in output per unit of time)	Tool assignment or encouragement, an IV estimate for adopters, effects among users after a rollout, or a completion-conditional contrast, on task time, resolutions per hour, or completed tasks; k = 9	+29% [+10, +51]; +33% [+6, +67] with randomized comparisons only. People working with the tool produced over a quarter more per unit of time on the tested tasks; the studies disagree widely, but no single study drives the average. The two workplace estimates from mainly poorer-country workforces, +15% and +25%, come from non-randomized rollouts.	Whether output with AI becomes earnings or demand; effects in informal work; long-run effects	Worker transitions (row 3): pilots measured on earnings
1b. Learning measured without the tool (SD)	AI tutoring with human support or guardrails versus usual instruction or no AI; k = 8 (four school trials in poorer countries)	0.21 SD [0.05, 0.36]; 0.30 [0.18, 0.43] without the two most precise estimates; 0.16 [−0.04, 0.37] without the two trials rated at high risk of bias. Students who practiced with a supported AI tutor or one with guardrails did modestly better on tests taken without it, although the two most precise trials found the smallest effects. Unrestricted answers lowered exam scores in Türkiye (−0.19 SD) but raised a one-week test among US undergraduates (+0.27 SD); both are shown, not pooled.	Persistence after the tool is withdrawn; transfer to routine schooling; cost per unit of learning	Schools (row 1)
1c. Business and labor-market outcomes: firm interventions and market responses (native units)	Not pooled: profits and sales after a tool is assigned; earnings, employment, and platform demand after a release; ten records	Structured summary (Table A8). When studies follow AI to profits, earnings, or demand, no consistent gain appears: no detectable average profit gain for small firms in Kenya, earnings and hours effects within 2% in Denmark, no statistically significant change in Finland, higher earnings in exposed US occupations, and demand for exposed freelance work down a fifth or more.	Wages in a poorer country; prices of expertise for nonusers	Worker transitions (row 3)

Outcome domain (scale)	Comparison shared by the pooled studies; k	Exploratory result (modified Hartung–Knapp 95% interval) and what it means	What remains unknown	Decision it informs (Table 4)
1d. Health and other public services: process and outcome measures (native units)	Not pooled: twelve records on different outcomes, one generative-AI cluster trial with patient outcomes, two clinical-reasoning trials, one court rollout, seven earlier-AI screening or triage trials, and one quality-improvement comparison (Table 3, row 4)	Structured summary (Section 3.3). Screening, referral, detection, throughput, diagnostic reasoning, and case resolution improve in most studies; the one generative-AI trial with patient outcomes is inconclusive (treatment failure adjusted odds ratio 0.77 [0.55, 1.08]); the one cost-effectiveness analysis among the seven earlier-AI trials finds its tool not cost-effective. The court rollout puts the tool’s annual infrastructure cost for eight judges, excluding development and maintenance, at about one-twentieth of one judge’s salary.	Patient and service outcomes; costs at scale; generative tools in public systems	Other public services (row 4)
2. Generative-AI tool → household consumption poverty in a low- or middle-income country, ex post	No qualifying study; k = 0	n.a. Our search found no study that measures whether a generative-AI tool changes household consumption poverty in a low- or middle-income country; this is the paper’s main evidence gap (Section 4.2).	The whole link from application to consumption poverty	Measurement plan for every large evaluation (Section 5)
3. Digital precedents (ICT, mobile money, broadband)	Not pooled: already synthesized	n.p. (Section 3.1). Earlier digital technologies raised welfare slowly and through complements such as mobile-money agents, regulation, and electricity; they are reminders, not evidence on AI.	Whether the lessons apply to AI	Timing of welfare effects left open (Section 5)
4. Validation of AI-based welfare measures (R ² ; exclusion error)	17 studies listed; 10 displayed as selected illustrations: 8 as R ² or squared correlations (2 with paired level and change statistics) and 2 with exclusion errors at a budget	Asset-wealth levels R ² 0.55 to 0.83 in multi-country studies; income and a poverty-probability index about 0.25 in one study; changes 0.04 to 0.50. The models tell where people are wealthier far better than whether they are getting richer or poorer. Phone-data targeting beat most feasible alternatives in Togo but lost to a survey-based proxy means test in Bangladesh; the statistics are compared, not averaged.	Reliability in new populations and periods, for changes and consumption; value of accuracy in decisions	Poverty statistics and targeting (row 2)

Source: Rows 1a and 1b, Table A12 and the effect-size table (Table A8); row 1c, Table A8; row 3, Stanley *et al.* (2018); row 4, Table A6 and Figure 4.

Note: Exploratory results on the study list of Section 4.1. Row 1a pools log response ratios, reported as % changes; row 1b pools standardized mean differences. Brackets give modified Hartung–Knapp 95% intervals; heterogeneity statistics, conventional and prediction intervals, and every sensitivity analysis are in Table A12. n.a.: no qualifying study; n.p.: not pooled (rows 1c and 1d by the rule in Section 4.1; row 3 already synthesized; row 4 compared, not averaged). Income-group averages are limited to one exploratory sensitivity row of Table A12 (Figure A1).

Table A8. Effect-size table for the exploratory summary, by outcome domain: inputs verified against the published sources

No.	Study	Domain	Setting (FY2027 group)	Design	n	Headline estimate, as published	Estimate (SE)	Weight, %	Variance source and status
Task performance with AI: pooled (output per unit of time)									
1	Peng <i>et al.</i> (2023)	Task performance, output per hour	Upwork freelancers, mostly in India and Pakistan (mixed geography)	RCT	95 assigned (45 treated, 50 control); 35 per arm completed	Task time 160.9 → 71.2 minutes (–55.8%) among completers	0.82 (0.39)	1	Reported CI, delta method; contrast conditional on completion (tool assignment among completers)

No.	Study	Domain	Setting (FY2027 group)	Design	n	Headline estimate, as published	Estimate (SE)	Weight, %	Variance source and status
2	Noy and Zhang (2023)	Task performance, output per hour	United States (HIC)	RCT	453 recruited; attrition 6% (control) and 11% (treatment); time effect estimated on the linear and convex incentive arms (80% of participants)	Second-task time 11 minutes lower, an adjusted contrast (0.75 SD, 95% CI 0.55 to 0.95), against a control mean of 27 minutes (−41%); raw means 17 and 27 minutes	0.52 (0.09)	10	Adjusted within-participant effect and 95% CI in pretreatment SDs (published article, source Fig. 1A; self-reported time, with a similar result for the objective measure of time active), converted to minutes (1 SD = 11/0.75 = 14.7 minutes; SE of the difference 1.5 minutes, from rounded inputs) and to a control-mean-normalized log response ratio, $\ln[27/(27 - 11)]$, by the delta method with the control mean treated as fixed; tool assignment
3	Haslberger <i>et al.</i> (2025)	Task performance, output per hour	United Kingdom (HIC), adults with ChatGPT accounts in an online panel	Preregistered RCT (encouragement)	1,041 (504 encouraged, 537 discouraged)	Total time on three work tasks 310 seconds lower (−32%) against a control total of 978 seconds, intention to treat; quality higher on all three tasks	0.38 (0.05)	13	Reported intention-to-treat effects on time and control means for the three tasks (supplementary Table S19, panel A: −89 s, SE 9.8; −157 s, 13.1; −64 s, 11.4), summed and converted to a log response ratio by the delta method; the SE of the sum assumes perfectly correlated task errors, an upper bound (assumption-based, flagged; correlations of 0.5 and 0 in Table A12); encouraged versus discouraged use
4	Dell’Acqua <i>et al.</i> (2026)	Task performance, output per hour	Global consultancy, consultants recruited across offices (mixed geography)	RCT	385 in the inside-frontier tasks (255 in the two AI arms, 130 control; 758 in the experiment)	Time to reach the last of 18 questions, 3,763 versus 5,023 seconds (−25.1%), converted to output per unit time +33.5%	0.29 (0.03)	14	Reported arm-level means and standard deviations (final article, source Table 2): AI arms combined mean 3,763 s (SD 1,396, n 255), control 5,023 s (SD 827, n 130); unadjusted contrast, delta method; tool assignment
5	Dell’Acqua <i>et al.</i> (2025)	Task performance, output per hour	Procter & Gamble, product-development professionals in Europe and the Americas (mixed geography)	RCT (four arms)	550 in the time regression (776 professionals)	Log time −0.366 (SE 0.070), individuals with AI versus without; average time −16.4%	0.37 (0.07)	12	Reported coefficient and robust SE of log time (source Table 4, column 1), sign reversed to output per unit time: a difference in mean log time, that is, a geometric-mean ratio, whereas the other time-based records use arithmetic means, so its pooled contribution is an approximation; the 16.4% reduction in average time enters an assumption-based sensitivity with the log-time SE, because the source reports no SE for that contrast (Table A12); tool assignment (individuals)

No.	Study	Domain	Setting (FY2027 group)	Design	n	Headline estimate, as published	Estimate (SE)	Weight, %	Variance source and status
6	Cui <i>et al.</i> (2026)	Task performance, completed pull requests per developer-week	Microsoft (developers mainly in the United States), Accenture (Southeast Asia), and an anonymous firm (location not stated); mixed geography	Field RCTs (3)	4,867 developers analyzed (1,521; 316; 3,030) of 5,120 at the start (1,746; 320; 3,054)	Completed tasks +26.1% (SE 10.3); IV estimate for adopters, three experiments; intention to treat +4.7% (SE 3.6); Microsoft and Accenture only, early period before control-group adoption	0.23 (0.08)	11	Reported SE, delta method; weighted IV (local average treatment) estimate for adopters, with assignment instrumenting use (adoption about 60% to 76%); assignment by developer and by team at Microsoft, by developer at Accenture, and by team in a staggered rollout at the third firm; the intention-to-treat estimate, for Microsoft and Accenture in the early period only, enters a sensitivity that changes the sample and window, not only the kind of estimate (Table A12)
7	Brynjolfsson <i>et al.</i> (2025)	Task performance, output per hour	Support agents, 89% outside the US, mainly the Philippines (UMIC)	Staggered rollout	5,172 agents; main model 12,295 agent-months	Resolutions per hour +0.301 (SE 0.033) on a pre-treatment mean of 1.97, i.e. +15.3%	0.14 (0.01)	15	Reported coefficient and SE (source Table II of the published version), delta method with the pre-treatment mean treated as known; effects among users after a staggered rollout
8	Gambacorta <i>et al.</i> (2026)	Task performance, completed tasks per two-week period	China (UMIC), programmers in one firm	Matched rollout (propensity-score matching with difference-in-differences)	1,219 programmers (335 treated, 884 comparison) in the code-output sample; task outcome re-matched: 8,911 programmer-periods	Tasks completed +0.220 log points (SE 0.043); code output +0.552 (SE 0.103); by seniority, significant for junior but not senior staff (no test of the difference)	0.22 (0.04)	13	Reported coefficient and SE (Journal of Financial Stability article, source Table 7, log-linear difference-in-differences, tasks completed): a difference in mean log output per two-week period, that is, a geometric-mean ratio; clustering of the standard errors not stated; the Poisson estimate (0.125, SE 0.050) enters a sensitivity (Table A12); code output not used because generated code inflates it; effects among users after a non-random rollout
9	Becker <i>et al.</i> (2025)	Task performance, output per hour	Open-source developers (countries not reported)	RCT within developer	16 developers, 246 tasks	Task time +19% [2, 39]	-0.17 (0.08)	11	Reported 95% CI; tool assignment (task within developer)
Task performance with AI: <i>descriptive records (quality, other outcomes, other comparators, non-randomized designs)</i>									
10	Noy and Zhang (2023)	Task performance, quality	United States (HIC)	RCT	453 recruited; quality effect estimated on the linear and convex incentive arms (80% of participants); analysis N not reported	Writing quality +0.41 SD [0.23, 0.59] (source Fig. 1B; the text reports 0.45)	0.41 (0.09)	n.a.	Reported CI, published article; descriptive

No.	Study	Domain	Setting (FY2027 group)	Design	n	Headline estimate, as published	Estimate (SE)	Weight, %	Variance source and status
11	Dell'Acqua <i>et al.</i> (2026)	Task performance, quality	Global consultancy	RCT	385 inside the frontier; 373 outside	Quality +34% and +30% in the two AI arms relative to the control mean inside the frontier; –19 pp accuracy outside the frontier	n.a.	n.a.	Arm means relative to control (final article, source Table 4); descriptive
12	Dell'Acqua <i>et al.</i> (2025)	Task performance, quality	Procter & Gamble (mixed geography)	RCT (four arms)	776	Solution quality +0.37 SD (SE 0.11), individuals with AI versus without; individuals with AI matched teams without AI	0.37 (0.11)	n.a.	Reported coefficient and SE (source Table 2); descriptive (quality)
13	Bastani <i>et al.</i> (2025), practice arms	Task performance, quality	Türkiye (UMIC)	RCT	≈1,000	Practice performance +48% (unrestricted chatbot), +127% (tutor)	n.a.	n.a.	% of control mean; descriptive
14	Korom <i>et al.</i> (2025)	Task performance, quality	Kenya (LMIC)	Quality-improvement comparison	15 clinics	Diagnostic errors –16%; treatment errors –13%	n.a.	n.a.	Non-randomized; sensitivity only
15	Qazi <i>et al.</i> (2026)	Task performance, quality	Pakistan (LMIC)	RCT (single-blind)	58 physicians, 342 cases	Diagnostic reasoning score 71.4% versus 42.6%: +27.5 pp [22.8, 32.2]	n.a.	n.a.	Reported CI; percentage points, descriptive
16	Rounding <i>et al.</i> (2026)	Task performance, quality	Kenya (LMIC), Indonesia (UMIC), and the Netherlands (HIC)	RCT	249 physicians	Case-performance score +18.0 pp Kenya [12.7, 23.2]; +10.7 pp Indonesia [5.7, 15.7]; +7.2 pp Netherlands [3.7, 10.7]; GPT-4o against controls without internet or guideline access	n.a.	n.a.	Reported CIs; percentage points on simulated cases, descriptive (quality); a within-study contrast across settings, not a patient outcome
17	Cruces <i>et al.</i> (2026)	Task performance, quality	Argentina (UMIC), online sample aged 25 to 45	RCT (online)	1,174 completers of 1,795 assigned	Education gap in task performance 0.55 SD without AI, 0.14 SD with AI; gains for all	n.a.	n.a.	Average effect moments not published; descriptive; preregistered
18	Idan and Anand (2026)	Task performance, quality	Early-career knowledge workers, self-study task (participant country not stated)	RCT	n.r.	Task performance up on average; gains concentrated among high AI-interaction competence	n.a.	n.a.	Moments not published; descriptive
19	Wang <i>et al.</i> (2025), Tutor CoPilot	Task performance, quality	United States (HIC), Title I schools	RCT (tutors)	1,800 students	Student mastery 66% versus 62% with AI-assisted human tutors	n.a.	n.a.	AI-assisted human tutors; descriptive
20	LearnLM Team and Eedi (2025)	Task performance, quality	United Kingdom (HIC)	RCT	165	AI tutoring at least as effective as human tutoring; novel problems 66.2% versus 60.7%	n.a.	n.a.	Comparator is human tutoring, not no-AI; descriptive
21	Ju and Aral (2026)	Task performance, output per hour	United States (HIC), online workers (Prolific)	RCT	2,234	Human–AI teams versus human–human teams: +1.954 ads per participant (SE 0.197), about 50% more	n.a.	n.a.	Not pooled: the comparator is a human collaborator, not work without a tool (task-performance rule, Appendix A.1)

No.	Study	Domain	Setting (FY2027 group)	Design	n	Headline estimate, as published	Estimate (SE)	Weight, %	Variance source and status
22	Dillon <i>et al.</i> (2025)	Task performance, work patterns	66 firms (mixed geography)	Field RCT	7,137	Email time –2 hours per week among adopters; no change in task composition	n.a.	n.a.	Time-use outcome, not output per hour; descriptive
23	Koo (2025)	Task performance, quality	Participants proficient in English, Arabic, or Chinese, recruited worldwide (multiple countries)	Preregistered RCT (online)	480 (160 per language; 240 given AI-generated content)	Email quality: completeness and relevance improved similarly across languages; actionability improved 0.59 SD less in Arabic and 0.47 SD less in Chinese than in English, and creativity 0.49 SD less in Chinese	n.a.	n.a.	Reported interaction terms ($p = 0.015, 0.022, \text{ and } 0.016$); pre-generated GPT-4 content rather than live use; descriptive (quality); heterogeneity by language in Table A9
Learning measured <i>without the tool</i> : pooled comparison (AI tutoring with human support or guardrails versus usual instruction or no AI)									
24	Henkel <i>et al.</i> (2024)	Learning	Ghana (LMIC)	Cluster RCT (11 schools), analyzed at student level	477 analyzed (637 at baseline; 241 control, 236 treatment)	Math growth +0.36 SD (Cohen's d), eight months (version 2 full text)	0.36 (0.21)	7	Student-level SE from the arm sizes (0.09) inflated by the design effect $1 + (477/11 - 1) \times 0.10 = 5.24$ for school-level assignment (intraclass correlation 0.10 assumed; 0.05, 0.20, and the student level in Table A12); no cluster-adjusted inference reported in the source; assumption-based, flagged; supported tutoring versus usual instruction
25	De Simone <i>et al.</i> (2025)	Learning	Nigeria (LMIC)	RCT (individual assignment)	654 in the regression; 759 of 1,328 assigned took the final test (422 treatment, 337 control)	Composite +0.31 SD (SE 0.07); English +0.24 SD (SE 0.07), six weeks	0.31 (0.07)	17	Reported coefficient and heteroskedasticity-robust SE (source Table 2, column 1, a score with test items weighted by difficulty); 43% of those assigned did not take the final test (excluded from the main pool of Burneo <i>et al.</i> 2026); the attrition-weighted estimate, 0.272 (SE 0.067; source Table 11), enters Table A12, row 8I; supported tutoring versus usual instruction
26	Kestin <i>et al.</i> (2025)	Learning	United States (HIC)	Crossover RCT (randomized by peer-instruction group of two or three; regression clustered by student, with a group-clustered robustness result)	194 students (316 student-lesson observations)	Post-test +0.63 SD versus active learning (student-clustered $p < 10^{-8}$; group-clustered robustness result $p < 0.001$)	0.63 (0.19)	7	Bound: $SE \leq 0.63/3.29 = 0.19$, the largest value consistent with the group-clustered result ($p < 0.001$) at the unit of assignment under a normal reference distribution; the reported coefficient controls for time on task, which the source reports as uncorrelated with post-test scores; assumption-based, flagged; AI tutor versus active-learning class; AI access during the post-lesson quiz not reported

No.	Study	Domain	Setting (FY2027 group)	Design	n	Headline estimate, as published	Estimate (SE)	Weight, %	Variance source and status
27	LearnLM Team and Fab AI (2026)	Learning	Sierra Leone (LIC)	Cluster RCT (48 classrooms, 12 schools)	1,423 analyzed (1,547 at baseline)	Math +0.258 SD [0.027, 0.488], eight instructional weeks over nine calendar weeks; +0.380 SD for completing at least 12 hours (IV, treatment on the treated); baseline-score interaction +0.195 SD [0.074, 0.315]	0.26 (0.12)	12	Reported CI, intent to treat (the source's clustered SE, 0.115, also rounds to 0.12); run by the tool's developer; registered before the trial (AEARCTR-0016651); supported tutoring versus usual instruction
28	Fischer <i>et al.</i> (2025)	Learning	Germany (HIC)	Lab RCT	334 (112 control, 222 treated)	Test score +0.227 SD (SE 0.106), AI tutor (either arm) versus control; +0.34 SD versus control with the tutor available from the start ("unrestricted access")	0.23 (0.11)	13	Reported coefficient and SE (source Table 1, column 1); a GPT-4 tutor grounded in the course text, whose two arms differ in when it becomes available (after 10 minutes or throughout), combined; AI tutor versus no AI
29	Bastani <i>et al.</i> (2025), hint-only-tutor exam arm	Learning	Türkiye (UMIC)	Cluster RCT (about 50 classes; classroom-clustered SE)	≈1,000 students	Unaided exam −0.004 (SE 0.013) on a [0, 1] grade; control mean 0.321, SD 0.277	−0.01 (0.05)	19	Reported coefficient and classroom-clustered SE (source Table 1, exam) divided by the control-arm SD of the normalized grade (0.277); hint-only AI tutor versus no AI (textbook practice); shares its control group with row 32
30	Vanzo <i>et al.</i> (2025)	Learning	Italy (HIC), technical high school, English as a second language	RCT (individual, stratified by grade point average within four classes)	76 students (37 treatment, 39 control); analysis N not reported	Learning gain on a teacher-set multiple-choice test $d = 0.251$ (one-sided $P = 0.156$), eight weeks (six curriculum units); third-year classes 0.603, fifth-year −0.004	0.25 (0.25)	5	SE derived from the reported one-sided P under a normal reference distribution, $SE = d/z(1 - P)$ (ACL 2025, Section 4.4); GPT-4 homework sessions with a tutoring prompt versus usual homework; AI access during the tests not reported
31	Oreopoulos and Low (2026)	Learning	United States (HIC), Tennessee, 18 middle schools	Two-year cluster RCT (53 grade-within-school clusters)	18 schools, 53 clusters	Stacked school-year effect +0.066 control-group SD (SE 0.043), per school year (annual contrasts pooled over two years); year 1 +0.016 (0.043), year 2 +0.116 (0.053); +0.050 per term (0.023); low median use	0.07 (0.04)	19	Reported estimate and cluster-robust SE in control-group SD units; each student-year contrast spans that year's own treatment window, so the effect is per school year, not cumulative; Khan Academy practice with the Khanmigo tutor versus usual remedial instruction, so it does not isolate the tutor's contribution
Learning measured without the tool: <i>other comparisons, sensitivity-only and descriptive records</i>									
32	Bastani <i>et al.</i> (2025), unrestricted exam arm	Learning (shown, not pooled)	Türkiye (UMIC)	Cluster RCT (shared control)	≈1,000 students	Unaided exam −0.054 (SE 0.022) on a [0, 1] grade, i.e. 17% below the control mean	−0.19 (0.08)	n.p.	Reported coefficient and classroom-clustered SE (source Table 1, exam) divided by the control-arm SD (0.277); not pooled because its control group is shared with row 29; unrestricted assistance versus no AI

No.	Study	Domain	Setting (FY2027 group)	Design	n	Headline estimate, as published	Estimate (SE)	Weight, %	Variance source and status
33	Contractor and Reyes (2026)	Learning (shown, not pooled)	United States (HIC), college undergraduates, proctored sessions	RCT (individual)	204 completed both sessions (256 randomized)	Unaided test one week later +0.27 SD [0.03, 0.51]; essay written with or without off-the-shelf AI (search allowed)	0.27 (0.12)	n.p.	Reported 95% CI; unrestricted general-purpose assistance versus conventional search, a different comparison from the pooled one; shown, not pooled
34	Gallegos (2026)	Learning (shown, not pooled)	Chile (HIC), university econometrics course	Two RCTs (individual)	289 (guidance); 303 (encouragement)	Final grade: guidance on learning-oriented use +0.218 SD (SE 0.100); encouragement alone +0.053 SD (SE 0.116); both against AI access alone	0.22; 0.05 (0.10; 0.12)	n.p.	Reported coefficients and SEs; both arms had AI access, so these compare ways of using AI, not AI against none; shown, not pooled
35	Fütterer <i>et al.</i> (2026)	Learning	Germany (HIC)	RCT (three arms)	371	No additional knowledge effect of either pedagogical-prompt condition over standard ChatGPT; all three arms used AI	n.a.	n.p.	Moments not published; shown, not pooled: added design features versus an existing AI tool
36	Oreopoulos <i>et al.</i> (2026), mastery practice	Learning	United States (HIC), Hamilton County middle schools	RCT (individual; 2 × 2 × 2 factorial)	6,327 in the delayed-test sample (6,997 in week 1)	AI tutor versus computer-assisted practice without AI, non-mastery condition: practiced-minus-unpracticed number correct (each out of two questions) on an unaided delayed test −0.014 (SE 0.034); within the mastery condition +0.026 (0.034)	n.a.	n.a.	Not in the pooled estimate: the outcome is a difference between two counts of correct answers, each out of two questions, whose SD is not reported; entered in one sensitivity row of Table A12 at a calibrated SD of 0.96 correct answers, derived from the reported robust standard error assuming four equal cells with equal variances; 0.6 to 1.2 gives a pooled value of 0.16 to 0.17 SD
37	Agweyu <i>et al.</i> (2026)	Clinical quality	Kenya (LMIC)	Cluster RCT (103 clinical officers)	9,691 patients	14-day treatment failure 2.2% versus 2.0%: aOR 0.77 [0.55, 1.08]; documentation quality aOR ≈ 1.7	n.a.	n.a.	Patient outcome kept separate from learning; descriptive
38	Mehmood <i>et al.</i> (2026)	Public services (case resolution)	Pakistan (LMIC), 118 district courts	RCT (judges randomized; court outcomes by district share of judges assigned to the assistant with targeted training)	1,559 judges	Resolved cases +6.3% at the median share of treated judges; no rise in appeals per resolved case	n.a.	n.a.	Descriptive; the estimate is a district-level dose response to the share of treated judges, not a per-judge assignment effect, so it is not pooled
Business and labor-market outcomes (structured synthesis, not pooled)									
39	Otis <i>et al.</i> (2026)	Business and labor market	Kenya (LMIC)	RCT	≈640	Average ≈ 0 (n.s.); top performers +15%, bottom −8%; gap 0.27 SD	n.a.	n.a.	Structured synthesis; not pooled; qualitative null for the average (no point estimate extracted)

No.	Study	Domain	Setting (FY2027 group)	Design	n	Headline estimate, as published	Estimate (SE)	Weight, %	Variance source and status
40	Humlum and Vestergaard (2025a)	Business and labor market	Denmark (HIC)	Surveys × registers, DiD	≈25,000 workers	Earnings and hours ≈ 0; effects above 2% ruled out	n.a.	n.a.	Structured synthesis; reported bounds
41	Demirci <i>et al.</i> (2025)	Business and labor market	Global platform	Difference-in-differences	Platform posts	Posts −21% (writing, coding); −17% (image creation)	n.a.	n.a.	Structured synthesis; not pooled; no standard error extracted
42	Hui <i>et al.</i> (2024)	Business and labor market	Upwork freelancers (mixed geography)	Difference-in-differences	Platform workers	Number of jobs −2%; monthly earnings −5.2% for exposed freelancers	n.a.	n.a.	Structured synthesis; not pooled; no standard error extracted
43	Qiao <i>et al.</i> (2025)	Business and labor market	Online labor platform (mixed geography)	Difference-in-differences	Platform workers	Translation: transaction volume −9.0%, earnings −29.7%; web development: +6.4% and +66.5%, in the first ten months; larger US gains in volume and, at the 10% level only, in earnings	n.a.	n.a.	Structured synthesis; not pooled; exponentiated log coefficients from the 2025 revision; over the extended window to January 2025 (source Table 8), web-development gains are smaller (+4.6% and +38.0%) and translation losses slightly larger (−9.8% and −33.6%)
44	Teutloff <i>et al.</i> (2025)	Business and labor market	Online freelancing platform, 116 skill clusters (mixed geography)	Difference-in-differences against unaffected skill clusters	Job postings	Postings for substitutable skills (writing, translation) −24% relative to unaffected clusters (raw fall 7%); falls in individual clusters of up to 59%, some not significant; complementary clusters mixed (machine learning up about a quarter: 24% in the abstract, 29% in the results)	n.a.	n.a.	Structured synthesis; not pooled
45	Tekic <i>et al.</i> (2026)	Business and labor market	Russia (HIC), freelance platform (Profi.ru), six service categories	Difference-in-differences with panel fixed effects	2,190 continuing providers, offline controls included (2,412 units, 57,888 unit-months); full panel 11,894	Monthly participation (at least one completed order) relative to offline services −6.7 pp (word processing), −8.7 pp (translation), −4.4 pp (design), smaller in the full panel; no comparable decline in programming, marketing, or web development	n.a.	n.a.	Structured synthesis; not pooled; access to frontier tools constrained; February 2023 diffusion anchor; sanctions-era shocks not fully separable
46	Kauhanen and Rouvinen (2025)	Business and labor market	Finland (HIC), all wage earners	Synthetic difference-in-differences by occupational exposure	Population registers	No statistically significant change in wages or employment in the two years after ChatGPT's release	n.a.	n.a.	Structured synthesis; not pooled
47	Chen <i>et al.</i> (2025)	Business and labor market	United States (HIC), Current Population Survey	Synthetic difference-in-differences by occupational exposure	Occupation panel	Weekly earnings about \$89 higher (bootstrap SE 10.5; January 2010 dollars) in exposed occupations; unemployment +0.0024 (bootstrap SE 0.0003), about 0.2 pp, which the authors call negligible	n.a.	n.a.	Structured synthesis; not pooled; working paper; occupational exposure after release, not observed adoption

No.	Study	Domain	Setting (FY2027 group)	Design	n	Headline estimate, as published	Estimate (SE)	Weight, %	Variance source and status
48	Fang <i>et al.</i> (2026)	Business and labor market	Cross-border online retail platform (mixed geography)	Randomized field experiments (seven workflows)	Millions of users and products	Sales effects by AI workflow from -4.5% to +16.3%, several not statistically significant	n.a.	n.a.	Structured synthesis; not pooled; firm sales rather than worker or household outcomes

Source: Authors' extraction from the studies listed.

Note: Task-performance estimates of output per unit of time are log response ratios (time reductions converted to output gains; two records count completed tasks per developer-week or per two-week period, Cui *et al.* 2026 and Gambacorta *et al.* 2026); learning estimates are standardized mean differences, as reported or converted with the control-arm standard deviation; weights are random-effects weights within each domain. All seventeen pooled inputs were verified against their sources; this table is generated from the full extraction record. Three pooled variances rest on stated assumptions (the Ghana design effect, the Kestin bound, and the Haslberger bound on the correlation of task errors); Table A12 drops or varies each. Two task-performance records are differences in mean log output, that is, geometric-mean ratios (Dell'Acqua *et al.* 2025; Gambacorta *et al.* 2026), and the mastery-practice record enters one sensitivity row only. Validation metrics: Figure 4 and Table A6.

Table A9. Within-study evidence on who gains most from AI assistance: heterogeneity by experience, education, baseline performance, or language

No.	Study (setting)	Population and subgroup definition	Outcome	Contrast reported	Uncertainty of the contrast	Prespecified?
1	Brynjolfsson <i>et al.</i> (2025), support agents, mainly the Philippines	Agents by pre-AI productivity quintile and by tenure	Resolutions per hour	Lowest-skill quintile +36% (0.5 resolutions per hour); highest-skill quintile close to zero; agents with under one month of tenure +0.7 resolutions per hour, no effect after one year	Quintile estimates with standard errors in the source; no pooled test of the interaction in Table A8	Not stated
2	Noy and Zhang (2023), US professionals	Baseline writing ability from the first task	Writing quality; time	Weaker writers gained more in quality; the slope of second-task on first-task grades was 0.41 in the control group and 0.14 with ChatGPT; time savings similar across ability	Interaction reported in the source	Preregistered experiment
3	Dell'Acqua <i>et al.</i> (2026), consultants (mixed geography)	AI-treated consultants only: bottom and top halves of baseline performance	Task quality inside the frontier	Scores rose from the baseline task by about 31% in the bottom half and about 10% in the top half	Descriptive before-and-after means without a control group (source Figure 4); halves defined on the baseline score, so regression to the mean may favor the bottom half; no test of the difference	Not stated
4	Cruces <i>et al.</i> (2026), online sample	Education, more versus less educated	Task performance	Gap 0.55 SD without AI, 0.14 SD with AI; gains for all	Gap estimates reported; average-effect moments not published	Not stated
5	Otis <i>et al.</i> (2026), Kenyan entrepreneurs	Baseline business performance, above or below median	Profits and revenues	Top performers +15%, bottom performers -8%; gap 0.27 SD	Subgroup effects with standard errors in the source	Heterogeneity analysis reported; preregistration of the split not stated

No.	Study (setting)	Population and subgroup definition	Outcome	Contrast reported	Uncertainty of the contrast	Prespecified?
6	LearnLM Team and Fab AI (2026), Sierra Leone students	Baseline mathematics score (continuous)	Mathematics test	Positive baseline-score-by-treatment interaction, +0.195 SD [0.074, 0.315]	Interaction with 95% confidence interval	Preregistered trial; interaction prespecification not stated
7	Haslberger <i>et al.</i> (2025), UK working-age online sample	Gender, age, education, occupation	Task productivity	Gains on all tasks; no narrowing of gaps across any of the four groupings	Interaction tests: none significant in the main IV models; the few significant intention-to-treat interactions favor younger, university-educated, or professional respondents	Preregistered
8	Gambacorta <i>et al.</i> (2026), programmers in one Chinese firm	Seniority, junior versus senior	Code output	Code-output estimate significant for junior staff (0.67 log points, which the source reads as 67%, for staff with one year or less) but not for senior staff	Separate estimates by group; no test of the difference reported; non-randomized rollout	Not stated
9	Idan and Anand (2026), early-career knowledge workers	Competence in interacting with AI	Task performance	Gains concentrated among high-competence users; built-in guidance narrows the spread	By group; moments not published	Not stated
10	Koo (2025), participants writing in English, Arabic, or Chinese	Language of the task (English versus Arabic or Chinese)	Email quality (completeness, relevance, actionability, creativity)	Similar gains in completeness and relevance; actionability gains 0.59 SD smaller in Arabic and 0.47 SD smaller in Chinese, and creativity gains 0.49 SD smaller in Chinese, than in English	Interaction terms with p-values (0.015 to 0.022)	Preregistered experiment

Source: Authors' compilation from the studies listed (Tables A5 and A8).

Note: One row per study with a reported within-study contrast. Subgroup definitions (productivity quintile, tenure, baseline ability, education, baseline performance, baseline score, competence, language) and outcomes (output per unit of time, quality, profits, test scores) differ, so no contrast is pooled; contrasts would be pooled only where definitions and outcomes are comparable and the covariance or an interaction standard error is reported. A significant estimate in one subgroup beside an insignificant one in another is not evidence of a difference between them. Rows 7 and 10 were verified from their full texts (Table A5), and row 8 comes from a non-randomized rollout (Table A8).

Table A10. Correlates of the gap in AI diffusion between poorer and high-income economies: regressions and decomposition

Dependent variable: share of people who used a generative-AI tool, Q1 2026 (%)	(1)	(2)	(3)	(4)	(5)
<i>A. Regressions: coefficient (robust standard error)</i>					
Low income	−24.25 (1.77)	8.40 (5.79)	6.94 (5.60)	5.95 (5.25)	
Lower-middle income	−22.60 (1.89)	1.46 (4.33)	1.36 (4.34)	−0.61 (4.03)	
Upper-middle income	−15.91 (1.95)	−2.36 (2.54)	−2.06 (2.72)	−2.78 (2.61)	
ln GDP per capita		8.30 (1.50)	8.31 (1.80)	7.58 (1.65)	8.26 (0.83)
Internet users, 2023 (% of population)			0.03 (0.06)	0.04 (0.06)	0.01 (0.05)

Dependent variable: share of people who used a generative-AI tool, Q1 2026 (%)	(1)	(2)	(3)	(4)	(5)
Access to electricity, 2023 (% of population)			−0.06 (0.04)	−0.02 (0.04)	−0.10 (0.04)
Official language among the seven best-resourced languages (0/1)				3.15 (1.39)	2.73 (1.39)
English is an official language (0/1)				2.27 (1.41)	2.20 (1.48)
Constant	33.61	−54.20	−51.39	−50.79	−47.72
R ²	0.64	0.73	0.74	0.76	0.74
n	130	130	130	130	130
<i>B. Decomposition of the gap relative to high-income economies, columns (1) to (4) (percentage points)</i>					
	Low income	Lower-middle income	Upper-middle income		
Gap in column (1)	−24.25	−22.60	−15.91		
Explained by ln GDP per capita	−29.81	−21.96	−12.37		
Explained by internet users	−2.88	−1.60	−0.45		
Explained by access to electricity	1.02	0.38	0.02		
Explained by a best-resourced official language	0.94	0.78	−0.15		
Explained by English as an official language	0.51	0.41	−0.19		
Gap remaining in column (4)	5.95	−0.61	−2.78		

Source: Authors' calculations from Microsoft AI Economy Institute (2026), World Development Indicators, and the authors' coding of official languages.

Note: OLS with HC1 heteroskedasticity-robust standard errors in parentheses; economies with population above 1 million and complete data (n = 130). The dependent variable is the Figure 1, panel A measure in percentage points; income groups follow the FY2027 classification, with high income omitted. GDP per capita (current US\$, 2024, or 2023 where missing), internet use, and electricity access are from the World Development Indicators. The authors code two language indicators from official or de jure national languages: an official language in the best-resourced class of Joshi *et al.* (2020), namely English, Spanish, German, Japanese, French, Arabic, or Chinese; and English official, whose coefficient is the increment over the first. Panel B follows Gelbach (2016): each entry is a covariate's column (4) coefficient times the group's difference from the high-income mean, and a column's entries sum to the change in the group coefficient between columns (1) and (4). Income, internet use, and electricity move together, so their split in columns (3) to (5) is imprecise. Robustness (same economies): with PPP income, the column (4) group coefficients are 0.4 (5.2), −6.0 (3.7), and −6.5 (2.3) and the language coefficients 4.2 (1.5) and 2.1 (1.4); with a quadratic in log income, the group coefficients are about a point or less from zero and the English coefficient 0.9 (1.3). In column (4), the connectivity block is jointly insignificant (p = 0.76), while the language block and the three group indicators are jointly significant (p = 0.002 and 0.001); dropping six influential economies gives language coefficients of 1.5 (1.1) and 2.8 (1.0). The residual group gaps are therefore specification-sensitive, and an official language says little about the languages poor households speak.

Table A11. Existing syntheses of experimental evidence on generative AI, and what this review adds

No.	Synthesis	Scope, period, and coverage of poorer countries	Records screened; studies included	Headline estimate	What it leaves open for this review
1	Burneo <i>et al.</i> (2026), WDR 2026 background paper	Digital learning interventions, 2015 to May 2026. Ten economies, none low-income.	4,117 screened; 644 full texts; 14 randomized trials, 191 effect sizes	Learning 0.125 SD (0.030). Generative minus earlier tools 0.022 SD (0.075)	Education only. No cross-sector comparison. Teacher-facing evidence rests on three studies and is not pooled.
2	Sautmann and Weinert (2026), WDR 2026 background paper	AI support for frontline health workers in low-resource settings, conception to March 2026. Ten of seventeen studies in low- and middle-income countries.	68,049 screened; 230 full texts; 17 studies	Narrative synthesis; no pooled estimate	No common effect-size scale. Only one included study uses a generative model. Eight of seventeen sit in eye care and diabetes services.
3	Vaccaro <i>et al.</i> (2024), <i>Nature Human Behaviour</i>	Human-AI collaboration across task types, January 2020 to June 2023. No country-income coding.	106 experiments; 370 effect sizes	−0.23 SD against the better of human alone or AI alone	Closes before most generative-AI field evidence. No comparison by country income and no welfare outcome.
4	Maier <i>et al.</i> (2026), preprint	Generative AI in programming, 2019 to 2025. Countries not coded.	23 studies; 27 effect sizes	Productivity 0.33 SD; learning 0.14 SD, not distinguishable from zero	One task domain. No developing-country evidence and no welfare outcome.
5	ChatGPT-in-education meta-analyses, for example Deng <i>et al.</i> (2025)	ChatGPT and student learning, 2022 onward. Mostly rich-country or unspecified settings.	Twenty to thirty-five studies each	Around 0.6 to 0.7 SD	Weaker design screens. Weidlich <i>et al.</i> (2025) find four of nineteen comparisons clean on treatment, control, and outcome.
6	This review, outcome domains 1a to 1c	Generative AI and task, labor, learning, clinical, and farm outcomes, 2020 onward. Income group coded for every estimate.	Search procedure in Appendix A.1; 17 pooled estimates from 17 studies	Exploratory values, modified Hartung–Knapp intervals: output per unit of time +29% [+10, +51] (k = 9); AI tutoring with human support or guardrails 0.21 SD [0.05, 0.36] (k = 8). A different question from a sector-specific review: pooling within an outcome domain across sectors, with the income group of every estimate coded	n.a.
<i>B. Study-by-study reconciliation of the learning pool with Burneo et al. (2026)</i>					
	Study	Status in this review	Status in Burneo <i>et al.</i> (2026)	Outcome, timing, and comparator	Reason for the difference; appraisal notes
B1	Henkel <i>et al.</i> (2024), Ghana	Pooled (0.36 SD; variance widened for school-level assignment)	Included	Mathematics test without AI after eight months; business as usual	In both pools. Student-level inference in a trial randomized across 11 schools; co-authors affiliated with the tool’s developer

No.	Synthesis	Scope, period, and coverage of poorer countries	Records screened; studies included	Headline estimate	What it leaves open for this review
B2	De Simone <i>et al.</i> (2025), Nigeria	Pooled (0.31 SD composite; reported robust SE)	Excluded from the main pool for attrition above 30%; restored in a robustness check	Composite test without AI after six weeks; business as usual	Attrition 43% of those assigned; dropped in Table A12, row 8c, and reweighted for attrition in row 8l
B3	LearnLM Team and Fab AI (2026), Sierra Leone	Pooled (0.26 SD; reported CI)	Not included	Mathematics test without AI after eight instructional weeks; business as usual	Released in mid-May 2026, the month their search ended; run by the tool's developer; preregistered
B4	Kestin <i>et al.</i> (2025), US university	Pooled (0.63 SD; variance bounded from the group-clustered p-value)	Not included	Post-lesson quiz after one AI lesson in a two-week crossover (AI access during the quiz not reported); active-learning comparator	Outside their school scope (university course); crossover design
B5	Fischer <i>et al.</i> (2025), German laboratory	Pooled (0.23 SD; reported SE)	Not included	Immediate test after a 25-minute session; textbook comparator	Outside their school scope (laboratory, university students)
B6	Bastani <i>et al.</i> (2025), Türkiye	Hint-only-tutor arm pooled (−0.01 SD; reported coefficient and classroom-clustered SE divided by the control-arm SD); unrestricted arm shown, not pooled (−0.19 SD; shared control group)	Included as two arms with small negative effects on tests without AI	Unassisted exam at the end of each of four in-class practice sessions; business as usual	In both reviews; this review pools only the arm that matches the supported-tutoring comparison and keeps the unrestricted arm apart because the two share a control group (Section 3.3)
B7	Oreopoulos and Low (2026), United States	Pooled (0.07 SD per school year; reported cluster-robust SE)	Not included	District-administered NWEA MAP mathematics assessment, annual contrasts over two school years; usual remedial instruction	Released after their search cutoff of May 2026
B8	Cheng <i>et al.</i> (2024), China	Out of scope after full-text reading (Table A5, row 44): a scripted chatbot, not a generative model, compared with a human reading partner	Included	Comprehension test after one chatbot dialogic-reading session, kindergarten	Flagged from their list and read in full; a different tool and comparator
B9	Other trials in their pool	Out of scope	Included	Digital learning tools that are not generative AI	First-generation or pre-generative tools; tracked with the earlier-AI benchmarks where applicable
B10	Contractor and Reyes (2026); Gallegos (2026); Oreopoulos <i>et al.</i> (2026)	Shown, not pooled (other comparisons); sensitivity only (mastery practice)	Not included	One-week test; course grades; delayed practice test	Released after their cutoff or outside their school scope; none enters the pooled estimate

No.	Synthesis	Scope, period, and coverage of poorer countries	Records screened; studies included	Headline estimate	What it leaves open for this review
B11	Ma <i>et al.</i> (2026), China	Pre-generative benchmark, not pooled (Table A5, row 41): an adaptive practice system without a generative model	Not included	Mathematics test after 11 weeks; non-adaptive practice and no computer-assisted learning	Released after their search cutoff of May 2026; outside the generative-AI comparison of this review
B12	Vanzo <i>et al.</i> (2025), Italy	Pooled (0.25 SD; standard error derived from the reported one-sided p-value)	Not included	Teacher-set English test after eight weeks of GPT-4 homework sessions (AI access during the test not reported); usual homework	Not in their included list, though released before their cutoff; a four-class trial of 76 students

Source: Authors' compilation from the syntheses listed.

Note: Standard errors in parentheses. SD denotes a standard deviation of the outcome. Screening counts are records remaining after de-duplication. Panel B matches the two study lists study by study: two trials are in both pools, the Nigerian trial was excluded there for attrition, and the remaining differences arise from publication timing, setting, and scope. The two review averages therefore answer different questions: a sector-specific average for digital learning tools of all kinds, and a cross-sector average for one comparison involving generative tools, neither of which is more informative than the other in general.

Table A12. Pooled estimates and sensitivity analyses by outcome domain

No.	Domain and specification	k	Pooled estimate (SE)	95% interval: modified Hartung–Knapp first; conventional second; % equivalent	I ² , %	τ ²	95% prediction interval	Reading
1	Task performance with AI (output per unit of time): random effects, all nine	9	0.25 (0.07)	HK [0.10, 0.41]; conv. [0.15, 0.35]; +29% [+10, +51]; conv. [+17, +42]	90	0.017	[−0.07, 0.58]; −7 to +79%	Average gain positive; the small-sample interval excludes zero with or without any one study; the prediction interval applies to a new setting only if that setting is comparable to those studied (Q = 81.3, df = 8)
2	Task performance: fixed effect, all nine	9	0.19 (0.01)	[0.17, 0.21]; +21% [+19, +24]	n.a.	n.a.	n.a.	Shown for comparison only; heterogeneity rejects a common effect (Q = 81.3, df = 8)
3	Task performance: randomized designs only (drops the two rollouts)	7	0.29 (0.09)	HK [0.06, 0.51]; conv. [0.14, 0.43]; +33% [+6, +67]; conv. [+15, +53]	87	0.028	[−0.18, 0.75]; −17 to +113%	Randomized contrasts include an IV estimate for adopters and a completion-conditional contrast, so randomization does not make them all intention-to-treat effects
4	Task performance: reported variances only (drops Haslberger <i>et al.</i> , whose variance is bounded)	8	0.23 (0.08)	HK [0.06, 0.41]; conv. [0.13, 0.34]; +26% [+6, +51]; conv. [+14, +40]	90	0.016	[−0.10, 0.56]; −9 to +76%	Drops the one assumption-based variance; variances derived from reported statistics, some by the delta method, are kept; the interval still excludes zero

No.	Domain and specification	k	Pooled estimate (SE)	95% interval: modified Hartung–Knapp first; conventional second; % equivalent	I ² , %	τ ²	95% prediction interval	Reading
4a	Task performance: Haslberger <i>et al.</i> with task-error correlations of 0.5 / 0 (the bound assumes 1)	9	0.25 / 0.26	HK: +29% [+10, +51] / +29% [+10, +51]	91 / 92	n.a.	n.a.	Lower correlations narrow the study's interval and raise its weight, but the pooled value barely moves
4b	Task performance: Cui <i>et al.</i> (2026) entered at its intention-to-treat estimate (Microsoft and Accenture, early period only; +4.7%, SE 3.6)	9	0.23 (0.07)	HK [0.06, 0.39]; conv. [0.13, 0.32]; +26% [+6, +48]; conv. [+14, +38]	92	0.017	[−0.10, 0.56]; −10 to +74%	Replaces the IV estimate with an assignment effect for two of the three firms over a shorter window, so sample and period change too; the interval still excludes zero
4c	Task performance: Dell'Acqua <i>et al.</i> (2025) at its 16.4% reduction in average time, with the log-time standard error (assumption)	9	0.23 (0.07)	HK [0.08, 0.38]; conv. [0.14, 0.33]; +26% [+8, +47]; conv. [+15, +38]	89	0.015	[−0.08, 0.54]; −8 to +72%	Assumption-based: the source reports no standard error for average time; without the study (leave-one-out), +27% [+6, +52]
4d	Task performance: Gambacorta <i>et al.</i> (2026) at its Poisson estimate (0.125, SE 0.050)	9	0.24 (0.07)	HK [0.08, 0.40]; conv. [0.14, 0.34]; +27% [+8, +50]; conv. [+15, +41]	90	0.018	[−0.10, 0.58]; −9 to +78%	An alternative functional form for the same outcome; the pooled value barely moves
4e	Task performance: without the studies rated at high or serious risk of bias (Peng <i>et al.</i> 2023; Gambacorta <i>et al.</i> 2026; Table A14)	7	0.25 (0.08)	HK [0.05, 0.45]; conv. [0.13, 0.36]; +28% [+5, +56]; conv. [+14, +44]	92	0.020	[−0.14, 0.64]; −13 to +90%	Table A14 ratings; the average is little changed and the interval widens
5	Task performance: leave-one-out range	8 each	n.a.	Estimates 0.22 to 0.30 (+25 to +35%); the small-sample interval excludes zero in all nine cases	86 to 91	n.a.	n.a.	Lowest without Noy and Zhang, highest without Becker <i>et al.</i>
6	Learning measured without the tool: AI tutoring with human support or guardrails versus usual instruction or no AI, random effects, all eight	8	0.21 (0.07)	HK [0.05, 0.36]; conv. [0.08, 0.34]	76	0.021	[−0.18, 0.59]	Excludes zero; the two most precise estimates, the Türkiye hint-only arm and the Khanmigo trial, account for much of the heterogeneity (Q = 28.6, df = 7)
6a	Learning: unrestricted assistance versus no AI (Türkiye, shared control; US college, search allowed)	2	Not pooled	−0.19 [−0.35, −0.04]; 0.27 [0.03, 0.51] (study intervals)	n.a.	n.a.	n.a.	Shown, not pooled; estimates of opposite sign in different tasks and populations
6b	Learning: guidance or encouragement versus AI access alone (Chile)	2 arms	Not pooled	0.22 [0.02, 0.41]; 0.05 [−0.17, 0.28] (study intervals)	n.a.	n.a.	n.a.	Shown, not pooled; comparisons between ways of using AI
7	Learning: fixed effect, all eight	8	0.11 (0.03)	[0.06, 0.17]	n.a.	n.a.	n.a.	Comparison only; dominated by the two most precise estimates
8	Learning: without the four-session Türkiye hint-only arm	7	0.26 (0.07)	HK [0.09, 0.43]; conv. [0.12, 0.40]	66	0.019	[−0.14, 0.65]	Influence check; heterogeneity remains

No.	Domain and specification	k	Pooled estimate (SE)	95% interval: modified Hartung–Knapp first; conventional second; % equivalent	I ² , %	τ ²	95% prediction interval	Reading
8a	Learning: without the Türkiye hint-only arm and the Khanmigo trial	6	0.30 (0.05)	HK [0.18, 0.43]; conv. [0.21, 0.40]	0	0.000	[0.17, 0.44]	Influence check: no detected heterogeneity among the remaining six, which does not show that their effects are equal or which program features matter
8b	Learning: without the Khanmigo trial	7	0.26 (0.09)	HK [0.04, 0.47]; conv. [0.09, 0.43]	77	0.034	[−0.27, 0.78]	Sensitivity excluding the Khanmigo trial
8c	Learning: without the attrition-flagged Nigerian trial	7	0.18 (0.07)	HK [−0.002, 0.36]; conv. [0.05, 0.31]	68	0.016	[−0.19, 0.55]	De Simone <i>et al.</i> (2025) excluded, as in the main pool of Burneo <i>et al.</i> (2026); the interval narrowly includes zero
8d	Learning: Nigerian composite replaced by its English score (0.24 SD)	8	0.19 (0.06)	HK [0.04, 0.34]; conv. [0.07, 0.31]	70	0.016	[−0.15, 0.53]	Outcome aligned to a single subject, with its own reported standard error
8e	Learning: Ghana intraclass correlation 0.05 / 0.20 / student-level variance	8	0.21 / 0.20 / 0.22	HK [0.06, 0.37] / [0.05, 0.36] / [0.07, 0.38]	76 / 75 / 79	n.a.	n.a.	Design effects 3.1, 9.5, and 1; the pooled value is insensitive because the Ghana weight is small once clustering is allowed for
8f	Learning: school trials only (Ghana, Nigeria, Sierra Leone, Türkiye hint-only arm, Italy, US Khanmigo)	6	0.16 (0.07)	HK [−0.01, 0.33]; conv. [0.03, 0.29]	75	0.016	[−0.24, 0.56]	The small-sample interval includes zero
8g	Learning: poorer-country school trials	4	0.20 (0.11)	HK [−0.15, 0.55]; conv. [−0.01, 0.42]	84	0.037	Not informative at k = 4	Exploratory sensitivity: Ghana, Nigeria, Sierra Leone, and the Türkiye hint-only arm; the wide interval does not support a comparison of effects across country groups
8h	Learning: reported variances only (drops the two assumption-based variances: Ghana and the US crossover)	6	0.16 (0.06)	HK [−0.005, 0.32]; conv. [0.03, 0.28]	75	0.015	[−0.22, 0.54]	Checks influence; the interval narrowly includes zero, which does not show a zero effect
8i	Learning: without the US crossover (variance bound)	7	0.17 (0.06)	HK [0.02, 0.32]; conv. [0.05, 0.29]	72	0.015	[−0.18, 0.52]	Kestin <i>et al.</i> (2025) excluded; the lowest leave-one-out value
8j	Learning: adding the mastery-practice experiment (SD of the count difference 0.96 correct answers, calibrated from the reported standard error with equal cells and variances; 0.6 to 1.2)	9	0.17 (0.06)	HK [0.02, 0.31]; conv. [0.06, 0.28]	78	0.017	[−0.17, 0.50]	Oreopoulos <i>et al.</i> (2026) at −0.01 SD (non-mastery contrast, count scale); 0.16 to 0.17 across the SD range
8k	Learning: without the studies rated at high risk of bias (the Ghana and Nigeria trials; Table A14)	6	0.16 (0.08)	HK [−0.04, 0.37]; conv. [0.03, 0.30]	71	0.016	[−0.23, 0.56]	Table A14 ratings; the estimate remains positive but is imprecise, and its interval includes zero

No.	Domain and specification	k	Pooled estimate (SE)	95% interval: modified Hartung–Knapp first; conventional second; % equivalent	I ² , %	τ ²	95% prediction interval	Reading
8l	Learning: Nigerian composite at its attrition-weighted estimate (0.27 SD, SE 0.07; source Table 11)	8	0.20 (0.06)	HK [0.05, 0.35]; conv. [0.08, 0.32]	73	0.018	[−0.16, 0.56]	The same difficulty-weighted composite as in the main estimate (0.31 SD), reweighted by the source for attrition; the pooled value barely moves
8m	Learning: without the two studies that do not report whether AI was unavailable during testing (the US crossover and the Italian trial)	6	0.17 (0.06)	HK [0.002, 0.33]; conv. [0.04, 0.29]	76	0.016	[−0.23, 0.56]	Kestin <i>et al.</i> (2025) and Vanzo <i>et al.</i> (2025) excluded; the interval narrowly excludes zero
9	Learning: leave-one-out range	7 each	n.a.	Estimates 0.17 to 0.26; the small-sample interval includes zero in 1 of 8 cases (without the Nigerian trial, narrowly)	66 to 79	n.a.	n.a.	Lowest without Kestin <i>et al.</i> ; highest without the Türkiye hint-only arm
10	Business and labor-market outcomes (structured synthesis)	10	Not pooled	n.a.	n.a.	n.a.	n.a.	Contrasts differ (profits; earnings, hours, and employment; platform demand, earnings, and participation; sales); structured summary in Table A8
11	Validation of AI-based measures	10 of 17 displayed	Not pooled	Asset-wealth levels R ² 0.55 to 0.83 in the multi-country studies; income and poverty-probability about 0.25 in one; changes 0.04 to 0.50; exclusion error: Togo rural 53% versus 59 to 78%; Togo nationwide 50% versus 52 to 76% (most feasible alternatives) and 48% (optimized occupational rule); Bangladesh 68% versus 48% (proxy means test)	n.a.	n.a.	n.a.	Selected illustrations compared by use and setting, not averaged; Table A6 states why seven listed studies are not displayed (Figure 4)

Source: Authors’ calculations from the effect sizes in Table A8.

Note: Task-performance estimates are log response ratios and learning estimates standardized mean differences; standard errors in parentheses are modified Hartung–Knapp for random-effects rows and conventional for fixed-effect rows. Random effects follow DerSimonian and Laird (1986). The modified Hartung–Knapp interval (Knapp and Hartung 2003; Röver *et al.* 2015) is primary (“HK”) and the conventional interval second (“conv.”), both computed from the full-precision effect sizes rather than the rounded entries of Table A8. Prediction intervals (IntHout *et al.* 2016; k – 2 degrees of freedom) describe the spread of effects expected in settings comparable to those studied, under the model; they are not forecasts for settings unlike these. Row 8g is an exploratory sensitivity; its wide interval does not support a comparison of effects across country groups (Figure A1). Rows 4e and 8k use the risk-of-bias ratings of Table A14. Reported variances are derived from reported standard errors, intervals, P values, or arm-level means and standard deviations, some by the delta method; rows 4 and 8h drop only the

three assumption-based variances. The values are exploratory because the study list is not exhaustive (Section 4.1), and small-study tests are not run at these counts (Egger *et al.* 1997).

Table A13. Evaluation designs with AI-predicted outcomes: what each estimates, what it requires, and the sample-size implication (illustrative)

Design	What the evaluation estimates	What it requires	Sample size relative to a survey-measured outcome (Appendix B)	Illustration under stated assumptions
1. Uncorrected predicted outcome	The effect on the prediction; reading it as the effect on welfare requires that the prediction retain the treatment-induced change	To estimate the proxy effect, randomization only; to read it as an unbiased welfare effect, prediction error unrelated to treatment, which cannot be tested without labeled outcomes	The ratio of the within-arm variances of the proxy and the true outcome, divided by the squared ratio of their treatment effects	If the prediction retains a share s of the true effect with the same residual variance, the multiple is $1/s^2$: a proxy that retains half the change needs four times the sample, and one that retains none cannot detect the effect at any sample size
2. Prediction corrected with labeled outcomes in both treatment arms (additive correction)	The effect on welfare: the effect on the prediction plus the difference between arms in mean prediction residuals, an additive identity	Representative labeled samples in each arm, large enough to estimate the residual difference; the variance of the corrected effect adds the variance of the estimated residual difference and its covariance with the proxy effect, and the sampling and assignment design	As in row 1, with the estimator variance of the corrected effect in place of the simple ratio; it falls as the labeled samples grow	Under the two restrictions of Appendix B, a treatment signal retained in proportion to R^2 and a proxy variance equal to R^2 times the true variance, an R^2 of 0.35 for changes (Figure 4) implies a multiple of about 3 before correction uncertainty, and an R^2 of 0.04 about 25; these follow from the restrictions, not from the published R^2 values, which establish neither. A multiplicative calibration, dividing the proxy effect by the retained share of the signal, is a separate model and is unstable when that share is small

Design	What the evaluation estimates	What it requires	Sample size relative to a survey-measured outcome (Appendix B)	Illustration under stated assumptions
3. Labeled outcomes combined with a larger set of predictions (prediction-powered inference)	The effect on welfare, with valid confidence intervals under the method's sampling and regularity conditions	A labeled subsample that represents the corresponding target sample in each arm, and predictions for the rest; assignment and sampling that support the inference	Precision relative to the labeled sample alone improves when the prediction-only sample is large enough and the predictions are useful; it need not improve	For a mean, the variance is the variance of the predictions over the prediction sample plus the variance of the residual over the labeled sample when the two samples are independent (Angelopoulos <i>et al.</i> 2023); overlapping samples, fitted models, and clustered assignment need the corresponding covariance and design adjustments; the gain is largest when predictions are accurate and the labeled sample is small; the method does not repair unrepresentative labels

Source: Authors' elaboration; accuracy values from Figure 4, panel A; Angelopoulos *et al.* (2023).

Note: Illustrative. Power depends on the treatment signal the prediction retains, residual variation, the assignment unit and clustering, and the labeled sample's size and representativeness. Out-of-sample R^2 , a squared correlation, and a reliability ratio are different quantities, and the $1/R^2$ result of row 2 needs the restrictions of Appendix B.3. For poverty, the statistic must be targeted directly: classifying households by predicted mean consumption does not recover the headcount, and correcting a mean does not correct incidence or inequality. Under cluster assignment, required samples also scale with the ratio of the proxy's and the true outcome's design effects, and the number of clusters matters as well as their size.

Table A14. Risk-of-bias ratings of the pooled studies (RoB 2 and ROBINS-I)

No.	Study (domain)	Tool	Randomization or confounding	Deviations from intended interventions	Missing outcome data	Outcome measurement	Selection of the reported result	Overall	Main reasons for the rating
1	Peng <i>et al.</i> (2023), task performance	RoB 2	Low	Low	High	Low	Some concerns	High	Task time observed only for participants who completed the task (35 of 45 treated, 35 of 50 controls), so the contrast is conditional on completion; no registered analysis plan found in the source
2	Noy and Zhang (2023), task performance	RoB 2	Low	Low	Low	Low	Low	Low	Preregistered; attrition 6% (control) and 11% (treatment), with Lee bounds reported; self-reported time gives results similar to the objective measure of time active; 10% to 20% of controls used ChatGPT, which reduces the treatment-control difference in AI use

No.	Study (domain)	Tool	Randomization or confounding	Deviations from intended interventions	Missing outcome data	Outcome measurement	Selection of the reported result	Overall	Main reasons for the rating
3	Haslberger <i>et al.</i> (2025), task performance	RoB 2	Low	Low	Some concerns	Low	Low	Some concerns	Preregistered encouragement design, analyzed here by intention to treat; the analyzed sample consists of completed panel interviews, and incomplete interviews are not reported; time recorded by the survey software
4	Dell'Acqua <i>et al.</i> (2026), task performance	RoB 2	Low	Low	Low	Low	Low	Low	Preregistered; all 385 participants in the inside-frontier tasks analyzed; time recorded by the task platform and quality graded blind to condition
5	Dell'Acqua <i>et al.</i> (2025), task performance	RoB 2	Low	Low	Some concerns	Low	Some concerns	Some concerns	Preregistered, with the plan under the registry's embargo; the time regression uses 550 solutions (309 individuals and 241 two-person teams, 791 people), 15 more people than the 776 the source reports as randomized, a gap it does not explain; people from incomplete teams were moved to individual work after assignment
6	Cui <i>et al.</i> (2026), task performance	RoB 2 (effect of use)	Some concerns	Some concerns	Low	Low	Some concerns	Some concerns	IV estimate for adopters, which assumes that assignment affects output only through use; team-level staggered rollout at one firm; 4,867 of 5,120 developers analyzed; registered only after the experiments ended (AEARCTR-0014530, October 2024), with no analysis plan
7	Brynjolfsson <i>et al.</i> (2025), task performance	ROBINS-I	Moderate	Low	Moderate	Low	Low	Moderate	Timing of access set by training capacity and by managers' allocation of agents to sessions, with flat pre-trends; agents must remain employed to be treated, and AI access lowered attrition; outcome from administrative records
8	Gambacorta <i>et al.</i> (2026), task performance	ROBINS-I	Serious	Low	Low	Low	Moderate	Serious	Exposure depended on which teams were made aware of the tool; matching on programmer characteristics and fixed effects remove stable differences, but only three two-week periods test for diverging trends; several output measures reported
9	Becker <i>et al.</i> (2025), task performance	RoB 2	Low	Low	Some concerns	Some concerns	Some concerns	Some concerns	Tasks randomized within developer; completion times self-reported, checked against screen recordings for a subset; some tasks dropped; AI not used in some allowed tasks, which narrows the difference in use rather than biasing the effect of assignment; no registered analysis plan found
10	Henkel <i>et al.</i> (2024), learning	RoB 2 (cluster)	Some concerns	Low	High	Some concerns	Some concerns	High	Eleven schools randomly assigned (five treated; method not reported); 477 of 637 students tested at endline, with losses of 30% in treatment schools and 20% in control schools and lower baseline scores among those lost; assessors' blinding not stated; preliminary report without a registered plan

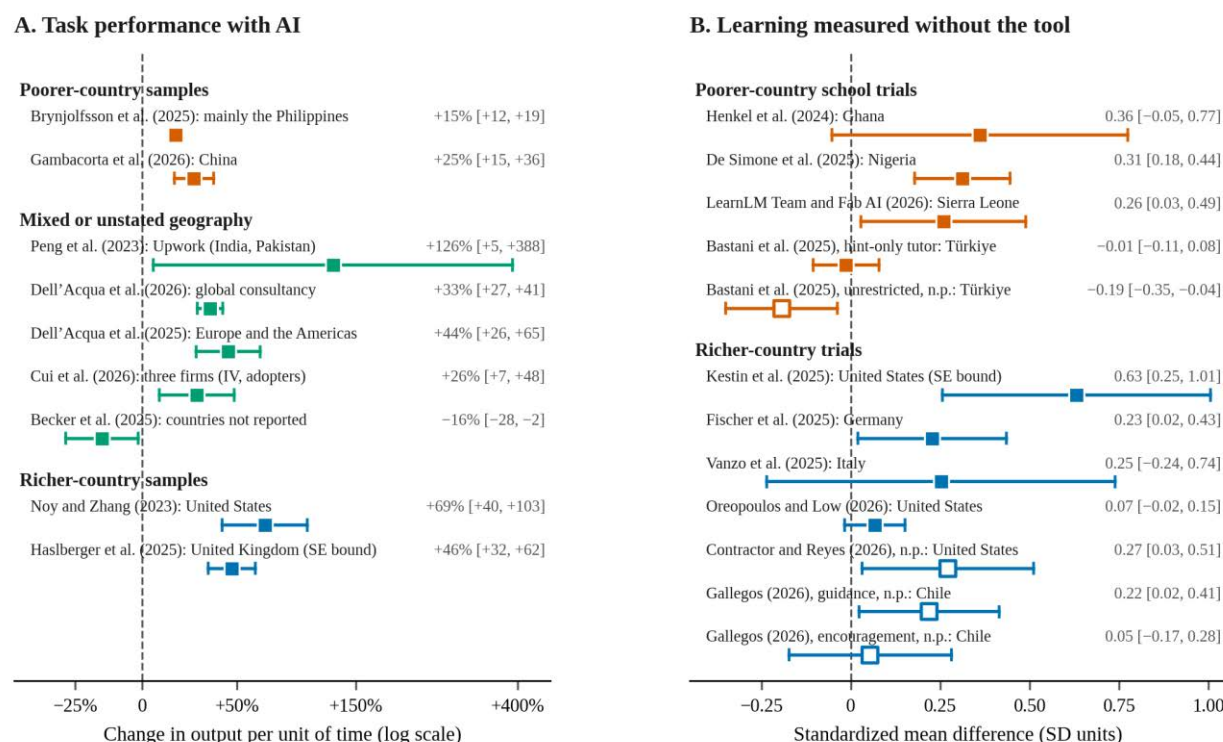
No.	Study (domain)	Tool	Randomization or confounding	Deviations from intended interventions	Missing outcome data	Outcome measurement	Selection of the reported result	Overall	Main reasons for the rating
11	De Simone <i>et al.</i> (2025), learning	RoB 2	Low	Low	High	Low	Some concerns	High	43% of those assigned did not take the final test (36% of the treatment group, 50% of controls) and about half are not in the regression; Lee bounds and inverse-probability weights are robustness checks only; monitored pencil-and-paper test; no registered analysis plan found in the source
12	Kestin <i>et al.</i> (2025), learning	RoB 2 (crossover)	Some concerns	Low	Some concerns	Some concerns	Some concerns	Some concerns	Crossover randomized by peer-instruction group; analysis restricted to the 194 of 233 enrolled students who completed all tests and both conditions; whether AI was unavailable during the post-lesson quiz is not reported; no registered analysis plan found
13	LearnLM Team and Fab AI (2026), learning	RoB 2 (cluster)	Low	Low	Low	Low	Some concerns	Some concerns	Registered before the trial began (AEARCTR-0016651), but the analysis plan was filed in January 2026, after endline testing; 124 of 1,547 students lost (9% of controls, 7% of treated students); tests developed and scored by an external assessor without knowledge of assignment; planned subdomain analyses dropped, with the reasons given; run by the tool's developer, a conflict of interest recorded separately (Section 3.3)
14	Fischer <i>et al.</i> (2025), learning	RoB 2	Low	Low	Low	Low	Low	Low	Preregistered laboratory experiment; multiple-choice test common to all arms; run in collaboration with the platform that supplied the tutor
15	Bastani <i>et al.</i> (2025), hint-only tutor arm, learning	RoB 2 (cluster)	Some concerns	Low	Low	Low	Some concerns	Some concerns	Classes allocated to arms by an integer program that balances observables, with students randomly assigned to classes; independent graders; no differential absenteeism; preregistered before data collection (AsPredicted 4DL_Q3J) with the unaided exam as the primary comparison, but the reported regression departs from the planned pairwise t tests, whose results the article reports as qualitatively similar
16	Vanzo <i>et al.</i> (2025), learning	RoB 2	Low	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	Randomized within classes, stratified by grades; controls were not forbidden to use ChatGPT, which reduces the treatment-control difference in use; some leakage of assignments to the teacher is likely; some students dropped for nonparticipation or incomplete surveys, in numbers not reported; whether AI was unavailable during the tests is not reported; one-sided tests without a registered plan
17	Oreopoulos and Low (2026), learning	RoB 2 (cluster)	Low	Low	Low	Low	Some concerns	Some concerns	Registered before the trial (AEARCTR-0013519), but the analysis plan was filed in May 2026, late in the second year, and results for the registered co-primary state test are pending; district-administered test, identical across arms; losses proportional across arms

No.	Study (domain)	Tool	Randomization or confounding	Deviations from intended interventions	Missing outcome data	Outcome measurement	Selection of the reported result	Overall	Main reasons for the rating
18	Oreopoulos <i>et al.</i> (2026), mastery practice, learning (sensitivity only)	RoB 2	Low	Low	Some concerns	Low	Some concerns	Some concerns	The registry entry cited (AEARCTR-0018678) dates from May 2026, after the March 2026 experiment, and describes a trial starting in August 2026; 6,327 of 6,997 week-1 students in the delayed test, with missing items coded as incorrect

Source: Authors' assessment of the full texts of the studies listed (Tables A5 and A8).

Note: RoB 2 (Sterne *et al.* 2019) is applied to the trials, for the effect of assignment (for Cui *et al.* 2026, the effect of use, which its IV estimate targets), and ROBINS-I (Sterne *et al.* 2016) to the two workplace rollouts, whose first domain column combines confounding, selection of participants, and classification of interventions. Judgments are low, some concerns, or high (RoB 2) and low, moderate, serious, or critical (ROBINS-I). One author made the ratings from the full texts with AI assistance and the other author reviewed them; provider involvement is recorded separately (Section 3.3 for Sierra Leone, Table A11 for Ghana, and row 14 above for the German laboratory study). Rows 4e and 8k of Table A12 omit the studies rated at high or serious risk. Registration dates are from the AEA RCT Registry and AsPredicted.

Figure A1. Where the evidence exists: the individual estimates of Figure 3 arranged by the geography of the study sample (no income-group averages shown)



Source: Studies in Table A8; authors' calculations.

Note: Each estimate is shown with its own conventional 95% interval, grouped by the geography of the study sample; mixed-geography samples draw participants from several income groups or do not report their countries. The display does not pool geography groups, because two or three studies per group would give unstable averages and geography varies with design, duration, and comparator; Table A12 (row 8g) reports one exploratory sensitivity for the four poorer-country school trials. Hollow markers (n.p.) are comparisons that enter no pooled estimate, for the reasons given in Table A8. The display shows where evidence exists, not national returns to AI; income groups follow the FY2027 classification of the study population.

Figure A2. Poorer-country evidence is concentrated in poverty measurement, health screening, and education: number of studies by domain and income group

	Income group of the study population (World Bank FY2027 classification)				
	Low income	Lower middle	Upper middle	High income	Multiple countries or geography unspecified
Work: tasks with AI assistance	0	0	3	3	8
Work: jobs, earnings, demand	0	1	0	5	5
Education	1	3	3	10	0
Health and public services	2	6	3	0	1
Agriculture and environment	0	1	0	0	1
Poverty measurement and targeting	4	8	2	0	7

Count: distinct studies, each read in full and coded. Studies listed in Tables A5 and A6, except the flood-forecasting study of Section 3.4.

Source: Authors' compilation from the lists of candidate studies (Tables A5 and A6) and the flood-forecasting study of Section 3.4; counts are those of Table 3.

Note: Distinct studies, each read in full, placed by the income group of the study population (FY2027) and by outcome domain (the earlier-AI benchmarks are identified in the Table 3 note and Section 3.3); multi-country studies, global platforms, and samples whose participants' countries are not stated are in the last column. The studies in every cell are listed in Tables A5 and A6, except the flood-forecasting study of Section 3.4.

Appendix B. Statistical methods

B.1 Pooling and small-sample inference

Within each outcome domain, study i contributes an estimate y_i with sampling variance v_i , for $i = 1, \dots, k$. Task performance uses log response ratios of output per unit time, so that a task-time reduction of r enters as $\ln(1/(1 - r))$, and learning uses standardized mean differences. Estimates are pooled by DerSimonian and Laird (1986) random effects. With fixed-effect weights $w_i = 1/v_i$ and fixed-effect mean $\bar{y}_w$, Cochran's statistic and the between-study variance are

$$Q = \sum_{i=1}^k w_i (y_i - \bar{y}_w)^2, \quad \tau^2 = \max\left\{0, \frac{Q - (k-1)}{c}\right\}, \quad C = \sum w_i - \frac{\sum w_i^2}{\sum w_i} \quad (\text{B.1})$$

and $I^2 = \max\{0, (Q - (k-1))/Q\}$ (Higgins *et al.* 2003). The pooled estimate, its conventional standard error, and the random-effects weights are

$$\theta = \frac{\sum w_i^* y_i}{\sum w_i^*}, \quad \text{se}(\theta) = (\sum w_i^*)^{-1/2}, \quad w_i^* = \frac{1}{v_i + \tau^2} \quad (\text{B.2})$$

With nine or fewer studies the conventional interval, $\theta \pm 1.96 \text{ se}(\theta)$, understates the uncertainty in τ^2 , so the primary interval of this review is the modified Hartung–Knapp interval (Knapp and Hartung 2003; Röver *et al.* 2015),

$$\theta \pm t_{k-1} \sqrt{\frac{q^*}{\sum w_i^*}}, \quad q^* = \max\left\{1, \frac{\sum w_i^* (y_i - \theta)^2}{k-1}\right\} \quad (\text{B.3})$$

where t_{k-1} is the 97.5th percentile of the t distribution with $k-1$ degrees of freedom and the floor at 1 keeps the interval from being narrower than the conventional one. The 95% prediction interval for the effect in a new setting is $\theta \pm t_{k-2} \sqrt{(\tau^2 + \text{se}(\theta)^2)}$ (IntHout *et al.* 2016). It applies only if the included studies describe a population of such settings, which is unlikely for task performance.

Where a trial randomized schools but reports student-level inference, the variance is multiplied by the design effect $1 + (m-1)\rho$ for average cluster size m and an assumed intraclass correlation ρ (0.10; 0.05, 0.20, and the student-level variance in Table A12). Where only a p-value threshold is reported, the largest standard error consistent with it at the unit of assignment, under a normal reference distribution, is used and labeled a bound. A t reference with the unreported number of clusters would give a slightly smaller bound. Where effects are reported for several tasks but not for their total, the variance of the total is bounded by assuming perfectly correlated task errors (correlations of 0.5 and 0 in Table A12).

What Section 4 reports, in detail. (1) A pooled estimate for each outcome domain from a random-effects model, which allows effects to differ across studies. Because each domain has nine or fewer studies, the primary interval is the modified Hartung–Knapp interval above, and the conventional interval and the prediction interval are secondary. (2) Descriptive comparisons by user skill, task type, and whether the tool came with support or guardrails, in place of the meta-regressions that the counts do not support, and the individual estimates arranged by the geography of the study sample (Figure A1), with income-group averages limited to one exploratory sensitivity row. (3) Within-study differences by experience, education, baseline performance, and language (Table A9); these contrasts are not pooled. (4) Sensitivity analyses that refit the pooled estimate under alternative variance assumptions, with reported variances only, with fixed effects, without each study in turn, without the studies rated at high or serious risk of bias (Table A14), and, for task

performance, without the two non-randomized rollouts and with alternative estimates for three studies (Table A12).

B.2 Decomposition of the diffusion gap

Table A10 compares a regression of diffusion on income-group indicators, column (1), with one that adds covariates, column (4). By Gelbach (2016), the change in a group coefficient between the two is

$$\gamma_g^{(1)} - \gamma_g^{(4)} = \sum_{j=1}^J \beta_j \delta_{jg} \quad (\text{B.4})$$

where β_j is the coefficient of covariate j in the full regression and δ_{jg} is the difference between the mean of covariate j in group g and its high-income mean.

B.3 Statistical power with AI-predicted outcomes

Consider a two-arm trial with n units per arm, a true welfare outcome y with within-arm variance σ_y^2 and treatment effect β_y , and a predicted outcome $\hat{y}$ with variance $\sigma_{\hat{y}}^2$ and treatment effect $\beta_{\hat{y}}$. Detecting the same underlying effect through $\hat{y}$ requires

$$\frac{n_{\hat{y}}}{n_y} = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2} \cdot \left(\frac{\beta_y}{\beta_{\hat{y}}} \right)^2 \quad (\text{B.5})$$

the ratio of variances divided by the squared ratio of effects (Table A13). Two special cases give $1/R^2$. Under classical error, $\hat{y} = y + u$ with u unrelated to y and to treatment, so that $\beta_{\hat{y}} = \beta_y$ and $\sigma_{\hat{y}}^2 = \sigma_y^2/R^2$, where R^2 is the share of the variance of $\hat{y}$ explained by y . For a fitted value whose response to treatment is proportional to its explanatory power, $\beta_{\hat{y}} = R^2 \beta_y$ and $\sigma_{\hat{y}}^2 = R^2 \sigma_y^2$. Neither case need hold: a prediction accurate in levels can have $\beta_{\hat{y}} = 0$, and then no sample size recovers the effect.

With the residual $e = y - \hat{y}$, $\beta_y = \beta_{\hat{y}} + (E[e | T = 1] - E[e | T = 0])$, and the variance of the corrected effect adds that of the estimated residual difference. Equation (B.5) compares power; it is not the variance of every corrected estimator.

Prediction-powered inference (Angelopoulos *et al.* 2023) combines survey-measured outcomes (labels) for part of the sample with predictions for the rest, and it improves on the labels alone only when the prediction sample is large and the predictions are useful. Poverty statistics and cluster designs are briefly discussed in the note to Table A13.