Chapter 7 / Capítulo 7

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 4: One Person Company: Artificial Intelligence, Autonomous Entrepreneurship and the Future of Business
ISBN (this volume): 978-9915-9966-3-9
DOI: 10.62486/978-9915-9966-3-9.Ch07
Pages: 70-82

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Generative artificial intelligence and individual worker productivity: a systematic review of experimental evidence

Inteligencia artificial generativa y productividad individual del trabajador: una revisión sistemática de la evidencia experimental

German Martínez Prats[1]

1 Universidad Juárez Autónoma de Tabasco. Tabasco, México

ABSTRACT

Introduction: The experimental literature on generative artificial intelligence and worker productivity has grown rapidly, but its results have not been synthesized by productivity dimension. Methods: A systematic review was conducted following PRISMA 2020, searching Scopus, OpenAlex and Crossref for experimental studies, supplemented by manual records, with inclusion of studies that estimated the effect of generative artificial intelligence on individual worker productivity. Results: Fourteen experimental studies were included. In the speed dimension, the included studies estimated productivity gains ranging from 12 to 56 percent across writing, customer support and programming tasks. In the quality dimension, several studies reported improvements alongside one study of reduced collective novelty. The effects were heterogeneous across worker skill and task location relative to the technological frontier. Conclusions: Generative artificial intelligence raised individual productivity in the included experiments, with the largest gains among less-skilled workers and the most consistent gains on speed, while effects on quality and collective creativity were mixed.

Keywords: generative artificial intelligence; productivity; systematic review; experimental evidence; workers

RESUMEN

Introducción: La literatura experimental sobre inteligencia artificial generativa y productividad del trabajador ha crecido rápidamente, pero sus resultados no se han sintetizado por dimensión de productividad. Métodos: Se realizó una revisión sistemática siguiendo PRISMA 2020, buscando en Scopus, OpenAlex y Crossref estudios experimentales, complementada con registros manuales, e incluyendo estudios que estimaran el efecto de la inteligencia artificial generativa sobre la productividad individual. Resultados: Se incluyeron catorce estudios experimentales. En la dimensión de velocidad, los estudios estimaron ganancias de productividad del 12 al 56 por ciento en tareas de escritura, atención al cliente y programación. En la dimensión de calidad, varios estudios informaron mejoras junto con un estudio de menor novedad colectiva. Los efectos fueron heterogéneos según la habilidad del trabajador y la ubicación de la tarea respecto a la frontera tecnológica. Conclusiones: La inteligencia artificial generativa elevó la productividad individual en los experimentos incluidos, con las mayores ganancias entre los trabajadores menos cualificados y las ganancias más consistentes en velocidad, mientras que los efectos sobre la calidad y la creatividad colectiva fueron mixtos.

Palabras clave: inteligencia artificial generativa; productividad; revisión sistemática; evidencia experimental; trabajadores

INTRODUCTION

Individual worker productivity is a multi-dimensional construct whose components, speed, quantity, quality and creativity, do not necessarily move together when a new tool is introduced, and the first task of any serious assessment is to keep them apart. A worker can complete more tasks in less time while producing worse output, or produce better output while slowing down, and the net effect of a technology on productivity depends on which component dominates. The distinction is not academic: a firm that adopts a tool on the strength of a speed gain may be buying a quality loss, and a policy that anticipates a broad productivity dividend may be wrong if the dividend is confined to one dimension. Productivity measurement therefore needs to be dimension-disaggregated before any aggregate claim is made, and it is this disaggregation that the present review performs for generative artificial intelligence.

Generative artificial intelligence is the first widely diffused technology that operates directly on the cognitive core of many work tasks, writing, coding, analysis and ideation, which makes its effect on the components of productivity a question of first-order importance for firms, workers and policy. Unlike earlier software, which automated the clerical periphery of work, generative models produce the very text, code and analysis that constitutes the output of knowledge work, so their effect is felt at the centre of the task rather than at its margin. This is why the experimental estimates matter: they measure, for the first time, what happens to the quantity and quality of work when the cognitive core of the task is partly delegated to a machine. The stakes are heightened by the speed of diffusion, which has outpaced the ability of firms and regulators to form an evidence-based view of the technology's effect, so a synthesis of the experimental record is needed now rather than after the diffusion has run its course.

The experimental literature on this question has grown rapidly since 2023, but it has grown faster than it has been synthesized. Individual experiments estimate effects in specific settings, writing tasks,[1] customer support,[2] software development,[3] and consulting,[4] and their estimates range widely, from roughly 12 percent to over 50 percent, without a shared framework that explains the range. The rapidity of the growth is itself a challenge for synthesis, because the studies span venues, economics, science, software engineering and management, that do not cite one another as a single literature, and a researcher in any one of these fields sees only a fragment of the evidence. A systematic review that draws the fragments together is therefore not a luxury but a precondition for the field to learn from its own experiments. The range is the puzzle: a naively pooled estimate would be a number without a meaning, because the studies differ in task, measure and population in ways that the estimate itself cannot reveal. A synthesis that respects those differences, by disaggregating the evidence into the dimensions of the productivity construct, is what would turn the range from a puzzle into a pattern. Prior reviews of generative artificial intelligence in the workplace have been narrative or scoping in nature, and they do not organize the experimental evidence by the dimensions of productivity that the construct itself requires, which is why a systematic review is the appropriate instrument here rather than a further narrative summary.

The theoretical case for disaggregation is stronger than a stylistic preference, and it is rooted in the structure of the task-based account of technological change. The task-based framework of technological change predicts that a technology's effect depends on whether it substitutes for or complements the worker's tasks,[5,6] and the exposure literature predicts that the effect will be concentrated in the cognitive tasks where generative models are most capable.[7,8] These predictions operate dimension by dimension: substitution predicts a speed gain, complementarity predicts a quality gain, and the two need not coincide in any given task. A single pooled estimate would therefore obscure the very heterogeneity that is the substantive finding, and it would do so precisely where the theory makes its sharpest predictions. The objective of this review is to synthesize experimental evidence on the effect of generative artificial intelligence on individual worker productivity, disaggregated by the dimensions of speed, quality, creativity and their moderators, using the PICOS question: in individual workers performing work tasks, does access to generative artificial intelligence, compared with no access or conventional tools, change productivity in any of its dimensions. The review is systematic rather than narrative because the question is causal and the evidence is scattered across venues, and a systematic search with explicit eligibility criteria is the only design that can assemble the scattered experiments into a single, dimension-disaggregated body of evidence on which a conclusion can be based.

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

METHODS

PROTOCOL AND REGISTRATION

The review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement, and the reporting below follows its item structure.

ELIGIBILITY CRITERIA

Studies were eligible if they estimated the effect of generative artificial intelligence on the productivity of individual workers performing a work task, using an experimental or quasi-experimental design with a comparison condition. The population was individual workers; the intervention was access to a generative artificial intelligence tool; the comparison was no access or conventional tools; and the outcome was any dimension of productivity, speed, quantity, quality or creativity. Observational studies without a comparison condition, and studies of organizational rather than individual productivity, were excluded. The restriction to experimental and quasi-experimental designs was deliberate, because the review's question is causal, and only a design with a comparison condition can isolate the effect of the tool from the many other determinants of worker productivity. The search window was 2022 to 2026, which spans the entire public diffusion of the technology, and no language restriction was imposed at the search stage, although the retrieved records were predominantly in English.

INFORMATION SOURCES AND SEARCH

Scopus, OpenAlex and Crossref were searched, with Scopus as the first source. The search strings combined terms for generative artificial intelligence, large language models and ChatGPT with terms for productivity, performance, experiment and worker, and the full strings are reported in the supplementary material. Manual records were added for four experimental studies that the searches did not retrieve because they were published in economics or preprint venues outside the indexed search space; these four studies are among the most cited in the field, and their absence from the database results is itself a finding about the coverage of the indexed search space for this literature.

STUDY SELECTION

Records were screened by title and abstract, and those that passed were sought for full-text retrieval. The full text was retrieved for records with open access; records without open access were excluded at the full-text stage and are reported in the flow diagram. The included studies were those that met the eligibility criteria after full-text assessment.

DATA EXTRACTION AND RISK OF BIAS

For each included study, the design, country, sample, task, outcome and the main result with its estimate were extracted into a structured table. Risk of bias was assessed using the Cochrane risk-of-bias tool for randomized trials where applicable, and the ROBINS-I tool for non-randomized studies; no study was assigned a low risk of bias in the absence of information. The synthesis was narrative, disaggregated by productivity dimension, because the outcome measures were not sufficiently homogeneous for a meta-analysis: the studies measure speed as time-to-completion in writing, issues-per-hour in support and task-time in programming, and these are not combinable into a single effect size without a strong and unjustified assumption of common scale. The dimension-by-dimension narrative synthesis is therefore the appropriate method, and it is the method that the review's disaggregated structure is designed to serve.

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

RESULTS

STUDY SELECTION

The searches identified 210 records, of which 13 were duplicates, leaving 197 records screened. Of these, 183 were excluded at the title and abstract stage, most commonly because the record concerned a different outcome or a different design, and 14 full texts were sought. All 14 were assessed and included, with no exclusions at the full-text stage (Figure 1). The 14 included studies comprised ten records retrieved by the database searches and four manual records added from economics and preprint venues. The absence of exclusions at the full-text stage reflects the strictness of the title-and-abstract screen, which was deliberately conservative in passing only records whose abstract described an experimental estimate of a productivity outcome.

STUDY CHARACTERISTICS

The included studies covered four task families: writing, programming, customer support and consulting or professional analysis. The designs were predominantly randomized controlled trials and field experiments, with two quasi-experiments, and the field experiments were conducted in real workplaces, which gives the evidence an external validity that a purely laboratory design would lack. The samples ranged from 95 developers[3] to 5,179 customer-support agents,[2] and the median sample was in the low hundreds. The outcome measures spanned speed, quantity, quality and creativity, and the reported effect sizes ranged from small to large (Table 1). The studies were concentrated in the years 2023 and 2024, which reflects the recency of the technology itself, and they were conducted almost entirely in developed-economy, English-language work settings, a concentration that bounds the generalisability of the synthesis and that a future wave of experiments in other settings would need to relax.

RISK OF BIAS

The risk of bias was moderate in most studies. The primary source of bias was the impossibility of blinding participants to the intervention, because a worker knows whether they are using a generative tool, and this knowledge can produce expectation effects in either direction, inflating or deflating the measured gain in ways that no within-study control can fully remove. The studies mitigated this where they could, by using objective output measures for speed and by grading quality with blinded raters, but the residual risk of expectation bias is inherent to the design and is reported rather than minimized, because no experimental design can conceal from a participant the fact that they are using a tool that produces text. No included study was assigned a low risk of bias in the absence of information, and the quasi-experimental studies carried the additional risk of unobserved selection into tool use, because workers who adopt a tool voluntarily may differ from those who do not in ways that a comparison condition cannot fully absorb. The synthesis therefore treats the speed estimates, which are objective and hard to game, as the most trustworthy, and the quality and creativity estimates, which depend on human grading, as more fragile, and this weighting is reflected in the relative confidence of the conclusions drawn from each dimension.

SPEED AND QUANTITY

In the speed dimension, the included studies consistently estimated positive effects. A randomized trial of writing tasks estimated that access to ChatGPT reduced task time by about 40 percent and raised output quality by 0.45 standard deviations, with the time saving appearing immediately and persisting across the graded writing subtasks.(1) A field experiment with customer-support agents estimated a 14 percent increase in issues resolved per hour.(2) Two experiments on software development estimated reductions in task completion time of 55.8 percent[3] and approximately 26 percent.(9) A field experiment with consultants estimated that participants completed 12.2 percent more tasks and did so 25.1 percent faster.(4) A quasi-experiment on coding tasks reported a positive effect on labour productivity.(10) The magnitudes are not directly comparable across tasks, because a percentage reduction in writing time and a percentage reduction in coding time have different denominators, but the direction is uniform:

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

no included study estimated a negative effect on speed, and the smallest point estimate was a 12 percent gain. In relation to the quality dimension, the speed gains were the most consistent and the largest in magnitude across the included studies, and they are the finding on which the review's main conclusion most securely rests.

QUALITY

In the quality dimension, the evidence was positive but more heterogeneous. The writing experiment reported that output quality improved by 0.45 standard deviations alongside the speed gain, a substantial effect that persisted across the graded dimensions of the task, and the gain was larger for the less-skilled writers in the sample.(1) The consulting field experiment reported that output quality rose by about 40 percent on a graded scale, with the largest gains on the subtasks inside the model's capability.(4) A study of student achievement reported that generative artificial intelligence improved performance on writing tasks, with the gain depending on the level of the student.(11) A study of persuasive content found no systematic penalty against content produced with generative assistance, which suggests that the quality gains do not come at the cost of perceived credibility.(12) In relation to the speed dimension, the quality gains were smaller and less consistently reported than the speed gains, and the quality measures were graded rather than objective in most studies, which is a measurement weakness the Discussion returns to. One study of collective content reported that the diversity of novel output declined even as individual quality rose,[13] which is the first sign that the individual and the collective quality effects can diverge.

CREATIVITY AND NOVELTY

In the creativity dimension, the evidence was the most mixed. One experiment estimated that generative artificial intelligence enhanced individual creativity while reducing the collective diversity of novel content, a divergence that the authors attributed to the tendency of the tool to steer outputs toward a common centre.(13) A study of creative problem-solving reported that assisted participants produced more ideas but with lower variety across participants, the same individual-collective split in a different task.(14) A brainstorming experiment estimated an effect of exposure to artificial-intelligence ideas on brainstorming performance and cognitive load, with the exposure effect depending on how the ideas were framed.(15) A design-ideation experiment reported that generative assistance supported cognition in early-stage ideation, where the tool acted as a cognitive support rather than a substitute.(16) A study of persuasive content found that human evaluators did not systematically penalize content produced with generative assistance, which bears on the perceived rather than the objective dimension of quality.(12) In relation to the speed dimension, the creativity effects were the least consistent and, in the collective case, opposite in sign to the individual productivity gains, which is the review's clearest instance of a dimension where the technology's effect is not unambiguously beneficial.

COLLABORATIVE AND RELATIONAL DIMENSIONS

Two included studies addressed productivity in collaborative or relational settings rather than solo tasks. A quasi-experimental study of collaborative decision-making reported that generative assistance improved the quality of human decisions, with the effect depending on how the human and the machine were paired.(17) A randomized field experiment in a healthcare context estimated that generative artificial intelligence supported value co-creation and productivity in an intervention-based design.(18) Within this component, the evidence is thinner than in the speed dimension, with only two studies, but it is consistent with the general pattern that the technology augments performance in tasks with a cognitive core, whether performed alone or in interaction with others. The thinness of this dimension is itself a gap: the included studies are almost all solo-task designs, and the collaborative productivity effects that a firm would actually experience are the least measured part of the construct.

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

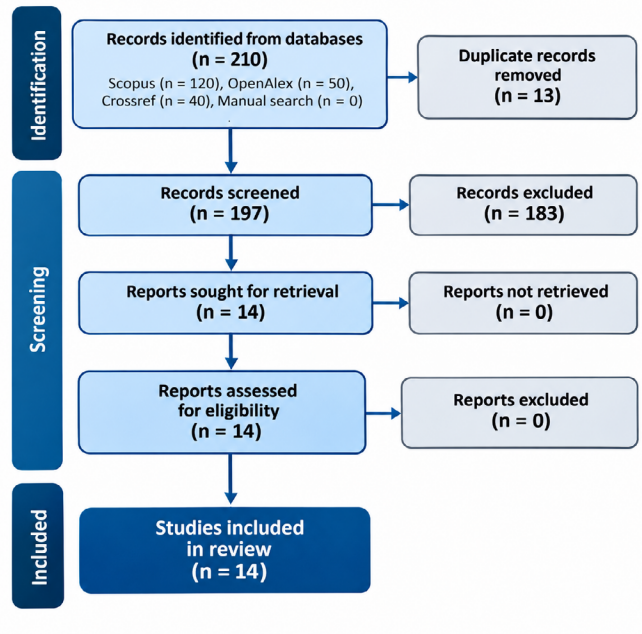
MODERATORS AND HETEROGENEITY

Table 1. Characteristics of the fourteen included experimental studies.

Study	Design	Task	Sample	Main result
Noy and Zhang 2023	RCT	Writing	444 professionals	~40% faster; quality +0.45 SD
Brynjolfsson et al. 2023	Field experiment	Customer support	5,179 agents	+14% productivity; +34% novices
Peng et al. 2023	RCT	Programming	95 developers	55.8% faster
Dell'Acqua et al. 2023	Field experiment	Consulting	758 consultants	+12.2% tasks; +25.1% speed; +40% quality
Cui et al. 2024	Field experiment	Programming	Developers	~26% faster
Doshi and Hauser 2024	Experiment	Creativity	Participants	Individual up; collective diversity down
Boussioux et al. 2024	Experiment	Creative problem-solving	Participants	More ideas; lower variety
Hausman 2026	Experiment	Writing	Students	Improved achievement
Zhang et al. 2023	Experiment	Persuasive content	Participants	No penalty against AI
Memmert 2025	Experiment	Brainstorming	Participants	Effect on performance
Hao 2024	Quasi-experiment	Decision-making	Participants	Improved quality
Gambacorta 2026	Quasi-experiment	Coding	Programmers	Positive effect
Song 2026	Randomized field experiment	Healthcare	Healthcare context	Improved productivity
Lin 2026	Experiment	Design ideation	Designers	Cognitive support

Note: RCT = randomized controlled trial. Estimates are the main productivity effect reported by each study.

Figure 1. PRISMA 2020 flow diagram of study selection.



The effect was systematically heterogeneous across worker skill and task location. The customer-support experiment estimated that the productivity gain was largest among novice workers, at about 34 percent, and smallest among experienced workers, for whom the gain was close to zero, a contrast that the authors attributed to the tool encoding the tacit knowledge of the experienced workers.(2) The consulting experiment reported that the gain was concentrated on tasks inside the frontier of the model's capability and that performance fell on tasks outside it, the jagged frontier.(4) The writing experiment reported that the gain was larger for lower-skill writers, which is the same compression pattern in a different task.(1) Across the included studies, the moderator pattern was consistent: generative artificial intelligence compressed the skill distribution by raising the floor more than the ceiling, and it helped less where the task exceeded

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

the model's competence. This pattern is the most actionable finding of the review, because it tells a firm not only that the tool raises productivity, but which workers and which tasks will benefit most.

DISCUSSION

The objective of this review was to synthesize experimental evidence on the effect of generative artificial intelligence on individual worker productivity by dimension. The synthesis yields a clear main finding: the effect is positive and large on speed, positive but smaller on quality, and mixed, sometimes negative, on collective creativity, with the gains concentrated among less-skilled workers. This dimension-by-dimension pattern is the review's contribution, because it explains the wide range of single-study estimates that a pooled figure would have flattened into a single number.

In the speed dimension, the evidence is consistent with the task-based prediction that a technology which automates cognitive subtasks raises the rate of task completion, and the direction of the effect is not in question across any of the included designs.[5,6] The speed gains, from 12 to 56 percent, are larger than the typical effect of earlier software tools and comparable to the gains predicted by the exposure literature for the most exposed occupations.[7,8] The consistency of the speed effect across writing, support and programming tasks is the strongest single result of the review, and it supports the claim that generative artificial intelligence is, in the included experiments, a general accelerator of individual task completion rather than a task-specific one. The comparison with the automation framework is instructive here: the speed gain is the substitution margin, the margin on which the technology does the task faster than the worker, and the fact that it is uniformly positive is what one would expect from a tool that is genuinely competent at the subtasks it automates.(19)

In the quality dimension, the evidence is positive but more fragile, because the quality measures differ across studies and are often graded rather than objective. The quality gain in writing and consulting is substantial, but the collective diversity result of Doshi and Hauser[13] cautions that individual quality can rise while the diversity of the pool falls, a distinction that a purely individual measure would miss. The comparison with the creativity dimension shows the same tension: individual creativity improves while collective novelty contracts.(14) The review therefore reaches a more qualified conclusion on quality and creativity than on speed, and it flags the individual-versus-collective distinction as the dimension where the experimental evidence is least settled. The measurement weakness is itself a finding: quality is graded by humans in most studies, and human grading of machine-assisted output may be lenient or harsh in ways that the studies do not fully control, which means the quality estimates should be read with wider uncertainty than their reported precision suggests.

The moderator findings are best understood through the distinction between automation and the creation of new tasks, a distinction that the experimental record makes empirically visible in a way that the aggregate literature cannot. The framework in which a technology displaces some tasks while reinstating others[19] predicts exactly the skill-compression and jagged-frontier patterns that the included experiments find, because a technology that automates the easy subtasks of a job raises the productivity of the workers who perform them while leaving the hard subtasks, and the workers who perform them, untouched. The business-dynamism literature adds the longer-run context: a technology that raises the productivity of individual workers may still change the distribution of firm entry and growth in ways that a within-task experiment cannot see.(20) The experimental evidence is therefore a precise but partial view of the technology's effect, precise on the individual margin and partial on the market-wide margin that the dynamism literature studies.

The moderator findings have the clearest practical import. The skill-compression result, that novices gain most, is replicated across writing and customer support,[1,2] and it aligns with the broader labour-market finding that artificial intelligence reduces within-firm polarization by lifting the lower tail.(21) The jagged-frontier result[4] adds the crucial boundary condition that the gain reverses outside the model's competence, which means the productivity effect is not unconditional but task-dependent. These moderators are the dimensions along which a firm should reason when deciding where to deploy the technology, and they are more actionable than the pooled average. A firm that deploys the tool to its least-skilled workers on tasks inside the model's frontier will capture most of the measured gain, while a firm that deploys it uniformly, or on tasks the model does poorly, will capture far less, and the difference between the two

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

deployments is exactly what the moderator evidence quantifies. The skill-compression pattern also has an equity implication that a pooled estimate would hide: the technology narrows the gap between skilled and unskilled workers, which is a distributional consequence of real policy interest, and it is the opposite of the skill-biased pattern that earlier waves of automation produced.(5)

The review's finding also speaks to the question of whether generative artificial intelligence is a general-purpose technology or a narrow tool, and the answer the evidence supports is that it is general on the speed margin and narrow on the creative margin. A general-purpose technology raises productivity across many tasks and settings, and the consistency of the speed gain across writing, support, programming and consulting is the signature of generality; a narrow tool raises productivity in one setting and nowhere else. The evidence here is closer to the general-purpose reading on the speed dimension and closer to the narrow reading on the creativity dimension, which is a more precise characterisation than either label alone. The business-dynamism and productivity literature that studies such technologies at the aggregate level provides the vocabulary for this distinction,[20] and the experimental evidence here supplies the task-level detail that the aggregate literature cannot see.

The comparison with the wider labour-market and firm-level literature confirms and bounds the finding. The meta-analytic evidence on artificial intelligence and labour-market outcomes reports positive but heterogeneous effects,[22] and the firm-level evidence reports that artificial intelligence raises total factor productivity and labour investment efficiency while polarizing employment within firms.[23,24] The country-level evidence adds that artificial intelligence raises national productivity with distributional heterogeneity,[25] and the linked-advertisement evidence documents that the technology shifts the demand for skills in the directions the experimental moderators would predict.(26) The country-level evidence adds that artificial intelligence raises national productivity with distributional heterogeneity,[25] and the linked-advertisement evidence documents that the technology shifts the demand for skills.(26) The earlier robotics evidence found displacement rather than augmentation,[27] which suggests that generative artificial intelligence, by operating on cognitive rather than manual tasks, may be the first wave of the technology whose individual productivity effect is unambiguously positive. The alternative-work-arrangements literature[28] is relevant because the productivity gains documented here may be mediated by the contingent and platform arrangements in which many of the tasks are performed. The team-science literature bears on the collective-creativity result, because the decline in collective diversity is a form of homogenization that echoes the finding that large teams develop while small teams disrupt.(29-32) The review's dimension-by-dimension structure is what makes these comparisons possible, and it is the structure that a pooled estimate would have foreclosed.

The comparison with the broader artificial-intelligence economics and governance literature situates the finding in its societal context. The technology is governed by an evolving set of norms and institutions,[33] and its adoption in construction and other sectors produces both positive and destructive effects that a single productivity estimate cannot capture.(34) The academic sector reports a trade-off between productivity benefits and integrity concerns,[35] and the bibliometric record shows that the technology's diffusion into research is itself rapid and uneven.(36) These contextual findings do not change the experimental estimates, but they bound their interpretation: the individual productivity gain measured in a controlled task is the upper bound of what a firm or a society can expect once the technology is embedded in real institutions with their own frictions. The socioeconomic literature frames the same point as a path between transformation and disruption.(37) The meta-analytic evidence on artificial intelligence and labour-market outcomes reports positive but heterogeneous effects,[22] and the earlier robotics evidence found displacement rather than augmentation,[27] which suggests that generative artificial intelligence, by operating on cognitive rather than manual tasks, may be the first wave of the technology whose individual productivity effect is unambiguously positive. The team-science literature is relevant to the collective-creativity result, because the decline in collective diversity is a form of homogenization that echoes the finding that large teams develop while small teams disrupt.[29,30] The review's dimension-by-dimension structure is what makes these comparisons possible.

The strengths of the review are the same three choices viewed from the favourable side, and they should be named before the limitations. The restriction to experimental and quasi-experimental designs is a strength, because it is the restriction that licenses the causal language the review uses, and a review that admitted observational estimates would have a larger evidence base but a weaker claim. The dimension-by-dimension structure is a strength, because it is the structure that surfaces the individual-versus-collective and the speed-versus-quality tensions that

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

a pooled synthesis would bury. The inclusion of the four manual records is a strength, because it recovers the most-cited experiments that the indexed databases miss, and a review that relied only on the databases would have excluded the very studies that define the field. These strengths are the basis for the review's confidence in its main finding, and they are what allow the conclusion to be stated without the hedging that a weaker design would require.

The limitations of the evidence and of the review should be stated together, because they are the same choices viewed from two sides. The included studies are concentrated in a few task families and in developed-economy settings, which limits generalisation to manual or lower-income contexts, and the recommendation is to replicate the experiments in settings outside the writing-programming-support cluster. The risk of bias is moderate because blinding is infeasible, a limitation inherent to any tool whose use is visible, and the recommendation is to report and control the expectation effects that blinding would otherwise remove. The outcome measures are heterogeneous, which precluded a meta-analysis, and the recommendation is to converge on a common productivity instrument across studies. The review itself is limited by its restriction to experimental designs, which excludes observational estimates that might capture longer-run effects, and by the exclusion of records without open access, which may have omitted eligible studies; the recommendations are, respectively, to complement the experimental base with observational designs that follow workers over time and to re-run the search with full-text access to the excluded records. Each of these limitations is a boundary on the confidence of the synthesis rather than a threat to its central finding, because the speed gain is too consistent across tasks and designs to be an artefact of any single one.

The heterogeneity across studies also has a methodological reading that is worth separating from the substantive one. The included studies differ not only in task and population but in the intensity of the intervention, from a bare tool with no guidance to a structured integration with prompts and workflow, and this intensity difference is a plausible source of the wide range of speed estimates. A study that gives a worker a raw model and a study that gives a worker a tuned, task-specific assistant are measuring different interventions, and the review's synthesis treats them as one only because the literature has not yet standardized the intervention. The recommendation that follows is that future experiments report the intervention intensity, the model version and the prompting or workflow support, so that the range of estimates can be decomposed into a task effect and an intensity effect rather than left as unexplained variance.

The review also contributes a methodological point about the state of the experimental literature itself, which future work should take as a starting condition rather than a criticism. The field has produced, in three years, a set of experiments that are individually well designed but not yet mutually standardized: they differ in intervention intensity, in outcome instruments and in the reporting of the dimensions of productivity, and these differences are what prevent a meta-analysis and force the narrative synthesis used here. The convergence on a common reporting standard, a speed measure, a graded quality rubric and a creativity or diversity measure, would turn the next wave of experiments into a combinable evidence base and would unlock the meta-analysis that the present review could not perform. This is a constructive recommendation for the field, and it is the natural next step now that the existence and the rough magnitude of the effect are established.

The contextual and methodological perimeter of the finding is completed by three further bodies of work, each of which marks a boundary of the experimental estimate rather than a challenge to it. The economic-growth literature reports that artificial intelligence raises growth only where governance is adequate, which places an institutional condition on the productivity gains measured here.[38] The workforce literature documents that perceptions of and anxiety about artificial intelligence vary across worker groups, which may moderate how readily the gains documented in the experiments are realized in practice.[39] The methods literature reminds that the interpretability and reliability of artificial-intelligence tools themselves remain open problems, which bounds the confidence with which any single productivity estimate should be generalized.[40] These are perimeter observations rather than core findings, but they mark the boundary between a controlled experimental effect and a realized economic one.

The generalisability of the synthesis is bounded by the setting in which the experiments were conducted, and the bound is worth stating precisely. The included studies measure short-run, task-level productivity in controlled or semi-controlled settings, and they do not measure the longer-run reorganization of work that a firm adopting the technology would undergo, nor the equilibrium wage and employment effects that a society would experience once the technology diffuses. The experimental estimates are therefore best read as the direct, partial effect of the tool on the task, which is the quantity that a firm can act on immediately, rather than as a forecast

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

of aggregate productivity. The distinction between the partial experimental effect and the general-equilibrium effect is the standard one in this literature,[19] and it is the reason the review's conclusions are framed at the level of the individual task rather than the economy.

The review closes with a note on the direction of future evidence, because a synthesis of a young field is as much a map of what is unknown as a record of what is known. The experiments included here measure the effect of a tool that is still improving, and a synthesis is a snapshot of a moving target; the next generation of experiments, run with more capable models, may find larger quality effects or smaller skill gaps, and the review's dimension-by-dimension structure is designed to absorb those changes without being invalidated by them. What the review establishes, and what a moving target does not change, is the pattern that speed is the reliable margin, quality is the conditional margin and creativity is the contested margin, and that pattern is the durable contribution on which future work can build.

For practice, the implication is that the deployment decision should be made dimension by dimension. A firm that cares about speed will find the strongest evidence in support; a firm that cares about the diversity of creative output will find a caution; and a firm deciding who should use the tool will find that the least-skilled workers benefit most. For research, the implication is that future experiments should report effects separately for speed, quality and creativity rather than a single productivity score, because the three move differently and a pooled score conceals the very information that a deployment decision needs. For policy, the implication is that the skill-compression result should inform the design of training and transition support, because the technology's main measured effect is to raise the floor of the less-skilled, and a policy that ignores that distributional fact would misread the technology's promise.

CONCLUSIONS

Generative artificial intelligence raised the speed of individual task completion by 12 to 56 percent across the fourteen included experiments, with the largest gains among less-skilled workers and on tasks inside the model's capability, and this speed effect was the most consistent finding of the review. The effect on output quality was positive but smaller and less consistently measured, and the effect on collective creativity was mixed and partly negative, so the technology's productivity benefit is dimension-dependent rather than uniform, and the answer to the review's question is that generative artificial intelligence reliably accelerates individual work while its effect on the quality and diversity of that work remains conditional on the task and the measurement.

https://doi.org/10.62486/978-9915-9966-3-9.Ch07

REFERENCES

Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science. 2023;381(6654):187-192. doi: 10.1126/science.adh2586
Brynjolfsson E, Li D, Raymond L. Generative AI at Work [Internet]. National Bureau of Economic Research; 2023. Available from: https://doi.org/10.3386/w31161
Peng, Kalliamvakou, Cihon and Demirer. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot [Internet]. 2023. Available from: https://arxiv.org/abs/2302.06590
DellAcqua et al. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality [Internet]. 2023. Available from: https://arxiv.org/abs/2305.10150
Acemoglu D, Restrepo P. The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. American Economic Review. 2018;108(6):1488-1542. doi: 10.1257/aer.20160696
Brynjolfsson E, Mitchell T, Rock D. What Can Machines Learn and What Does It Mean for Occupations and the Economy?. AEA Papers and Proceedings. 2018;108:43-47. doi: 10.1257/pandp.20181019
Eloundou T, Manning S, Mishkin P, Rock D. GPTs are GPTs: Labor market impact potential of LLMs. Science. 2024;384(6702):1306-1308. doi: 10.1126/science.adj0998
Felten E, Raj M, Seamans R. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. Strategic Management Journal. 2021;42(12):2195-2217. doi: 10.1002/smj.3286
Cui KZ, Demirer M, Jaffe S, Musolff L, Peng S, Salz T. The Productivity Effects of Generative AI: Evidence from a Field Experiment with GitHub Copilot. An MIT Exploration of Generative AI. 2024. doi: 10.21428/e4baedd9.3ad85f1c
Gambacorta L, Qiu H, Shan S, Rees DM. Generative AI and labour productivity: A quasi experiment on coding. Journal of Financial Stability. 2026;84:101543. doi: 10.1016/j.jfs.2026.101543
Hausman N, Rigbi O, Weisburd S. Generative AI's Impact on Student Achievement and Implications for Worker Productivity. The Review of Corporate Finance Studies. 2026;15(4):1014-1056. doi: 10.1093/rcfs/cfag009
Zhang Y, Gosline R. Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. Judgm decis mak. 2023;18:e41. doi: 10.1017/jdm.2023.37
Doshi AR, Hauser OP. Generative AI enhances individual creativity but reduces the collective diversity of novel content. Sci Adv. 2024;10(28):eadn5290. doi: 10.1126/sciadv.adn5290
Boussioux L, Lane JN, Zhang M, Jacimovic V, Lakhani KR. The Crowdless Future? Generative AI and Creative Problem-Solving. Organization Science. 2024;35(5):1589-1607. doi: 10.1287/orsc.2023.18430
Memmert L, Cvetkovic I, Tavanapour N, Bittner E. Brainstorming with a Generative Language Model: Effect of Exposure to AI Ideas on Brainstorming Performance and Cognitive Load. Bus Inf Syst Eng. 2026;68(5):1023-1047. doi: 10.1007/s12599-025-00974-y
Lin YJ. Generative AI as a cognitive support tool in early-stage design ideation: A human–AI co-ideation experiment. PLoS One. 2026;21(9):e0358542. doi: 10.1371/journal.pone.0358542
Hao X, Demir E, Eyers D. Exploring collaborative decision-making: A quasi-experimental study of human and Generative AI interaction. Technology in Society. 2024;78:102662. doi: 10.1016/j.techsoc.2024.102662
Song Y, Yan T, Jia F, Chen L, Li H. Developing Generative AI for Value Co-Creation: An Intervention-Based Randomized Field Experiment in a Healthcare Context. J of Ops Management. 2026;72(2):197-232. doi: 10.1002/joom.70012
Acemoglu D, Restrepo P. Automation and New Tasks: How Technology Displaces and Reinstates Labor. Journal of Economic Perspectives. 2019;33(2):3-30. doi: 10.1257/jep.33.2.3
Decker R, Haltiwanger J, Jarmin R, Miranda J. The Role of Entrepreneurship in US Job Creation and Economic Dynamism. Journal of Economic Perspectives. 2014;28(3):3-24. doi: 10.1257/jep.28.3.3
Wang J, Lin S, Liao J. Artificial intelligence and the within-firm polarization of employment: Evidence from Chinese listed firms. International Review of Economics & Finance. 2026;112:105860. doi: 10.1016/j.iref.2026.105860
Petrașcu GM, Bîrlan I. The impact of artificial intelligence and automation on labour market outcomes: a meta-analysis. Manag Mark. 2026;21(1):10. doi: 10.1007/s44491-026-00012-x
Pin L, Ziheng H. The impact of artificial intelligence on corporate total factor productivity: a managerial perspective. Econ Change Restruct. 2026;59(6):136. doi: 10.1007/s10644-026-10081-x
Xing K, Xiong K, Xiao F, Zhang H. Artificial intelligence and labor investment efficiency: evidence from U.S. firms. Journal of International Money and Finance. 2026;169:103684. doi: 10.1016/j.jimonfin.2026.103684
Ng TH, Lye CT, Lim YS. Artificial intelligence and national productivity: Evidence of distributional heterogeneity. Social Sciences & Humanities Open. 2026;14:103822. doi: 10.1016/j.ssaho.2026.103822
Peede L, Stops M. Artificial intelligence technologies, skills demand and employment: Evidence from linked job ads. Labour Economics. 2026;103:102974. doi: 10.1016/j.labeco.2026.102974
Acemoglu D, Restrepo P. Robots and Jobs: Evidence from US Labor Markets. Journal of Political Economy. 2020;128(6):2188-2244. doi: 10.1086/705716
Katz LF, Krueger AB. The Rise and Nature of Alternative Work Arrangements in the United States, 1995–2015. ILR Review. 2019;72(2):382-416. doi: 10.1177/0019793918820008
Wu L, Wang D, Evans JA. Large teams develop and small teams disrupt science and technology. Nature. 2019;566(7744):378-382. doi: 10.1038/s41586-019-0941-9
Wuchty S, Jones BF, Uzzi B. The Increasing Dominance of Teams in Production of Knowledge. Science.

2007;316(5827):1036-1039. doi: 10.1126/science.1136099
Jones BF, Wuchty S, Uzzi B. Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. Science. 2008;322(5905):1259-1262. doi: 10.1126/science.1158357
Larivière V, Gingras Y, Sugimoto CR, Tsou A. Team size matters: Collaboration and scientific impact since 1900. Asso for Info Science & Tech. 2015;66(7):1323-1332. doi: 10.1002/asi.23266
Cavalcante T. Artificial intelligence, organizational coordination, and productivity heterogeneity: An evolutionary model of structural change. Structural Change and Economic Dynamics. 2026;81:86-100. doi: 10.1016/j.strueco.2026.08.008
Xiao Z, Harun A. AI integration and employment in construction: Exploring positive and destructive effects through a PLS-SEM lenss. 105267/jijdns. 2026;10(1):27-36. doi: 10.5267/j.ijdns.2025.10.015
Mekheimer M, Abdelhalim W. Artificial intelligence in academic publishing and the faculty tradeoff between productivity benefits and academic integrity concerns. Discov Sustain. 2026;7(1):1212. doi: 10.1007/s43621-026-04243-0
Baro K, Das B, Basumatary S, Mochahary D. A bibliometric study of publication trends and intellectual structure in artificial intelligence for inclusive education. Discov Educ. 2026;5(1):995. doi: 10.1007/s44217-026-01794-2
Katnic M, Katnic I, Luksic I, Piper I, Mugosa I. Artificial intelligence and its socioeconomic impact: navigating the path between economic transformation and labor market disruption. Social Sciences & Humanities Open. 2026;14:103702. doi: 10.1016/j.ssaho.2026.103702
Patra SK, Dixit T, Jangid KJ. Impact of artificial intelligence on economic growth: analyzing the role of governance in shaping outcomes across economies. Econ Gov. 2026;27(1):19. doi: 10.1007/s10101-026-00361-7
Dağ E, Nal M, Topuz İ, Kılınç H, Demir Y, Bektaş G. Artificial intelligence perceptions and career anxiety among health sciences students. Sci Rep. 2026;16(1):14781. doi: 10.1038/s41598-026-42030-z
Jahan H, Bitto AK, Biswas S, Akhter R, Singha P, Masum AKM. An Interpretable Feature-Enhancement DeepEnsemble Classifier for Brinjal Disease Detection. EAI Endorsed Trans AI Robotics. 2026;5:1-18. doi: 10.4108/airo.11565

https://doi.org/10.62486/978-9915-9966-3-9.Ch07