Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Short-Term Efficiency, Long-Term Cognitive Cost? A Systematic Literature Review of Large Language Model Use in Professional Work

Chetna Khandare

Charlie@inookey.co.uk

University of Birmingham https://orcid.org/0009-0009-1128-3046

Systematic Review

Keywords: Large language models, generative AI, cognitive offloading, skill decay, automation bias, human-AI interaction, knowledge work, systematic review

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11845791/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Short-Term Efficiency, Long-Term Cognitive Cost? A Systematic Literature Review of Large Language Model Use in Professional Work

CHETNA KHANDARE
Inookey AI Technology Private Limited
Birmingham, United Kingdom
ORCID 0009-0009-1128-3046
Corresponding author: chetnakhandare2000@gmail.com

ABSTRACT Many professionals now use large language models (LLMs) at work, and experiments show that the tools make many tasks faster and improve the immediate output. The worry is that these gains may cost workers some of the skill, judgement and self-knowledge they would otherwise build or keep. We systematically reviewed empirical research on the cognitive and performance consequences of LLM use among workers and professionals, published between January 2020 and November 2026. Following the PRISMA 2020 guidelines, we identified 1,559 records in OpenAlex and Europe PMC and five further studies through citation searching. After screening, 157 studies remained. Two-thirds were published in 2026, and most come from software development, healthcare and general knowledge work. Randomised and field experiments consistently report short-term gains in speed, output and quality, typically largest for less experienced workers. These gains are task-dependent, and LLM-assisted professionals frequently fail to match the model working alone. Fewer than ten studies measured unaided performance, memory or skill after LLM use. Those that did reported reduced recall of AI-assisted work, impaired skill formation and fragility when access was disrupted. Cross-sectional studies link reliance and confidence in AI with lower self-reported critical thinking and professional self-efficacy, but self-reported capability loss did not track objective performance. Experience, task fit, mode of use, AI literacy and verification behaviour emerged as moderators. We conclude that the short-term efficiency gain is well supported, while the long-term cognitive cost remains plausible but largely untested in professional settings. We propose a framework and research agenda centred on longitudinal and withdrawal-based designs.

INDEX TERMS *Large language models, generative AI, cognitive offloading, skill decay, automation bias, human-AI interaction, knowledge work, systematic review.*

I. INTRODUCTION

Within three years of public release, LLM-based assistants such as ChatGPT, GitHub Copilot, Claude and Gemini have entered writing, coding, analysis and clinical work. Experiments have found large immediate benefits. Access to ChatGPT reduced completion time and raised quality in professional writing tasks (Noy and Zhang, 2023). An LLM-based assistant increased the productivity of customer-support agents, especially novices (Brynjolfsson, Li and Raymond, 2025). Consultants using GPT-4 completed more work, of higher quality, on tasks within the model's capabilities (Dell'Acqua et al., 2023).

These studies measured performance with the tool. They say less about what happens to a worker's own ability when the tool is used repeatedly and then fails, errs subtly or is withdrawn. In a field experiment with students, unrestricted GPT-4 access improved practice performance but harmed later unaided performance (Bastani et al., 2025). Developers learning a new library with AI assistance formed weaker skills (Shen and Tamkin, 2026). Knowledge workers with greater confidence in generative AI reported less critical thinking (Lee et al., 2025). Macnamara et al. (2024) argue that AI assistance may speed up skill decay in experts and slow skill development in novices, often without the user noticing.

Several systematic reviews have addressed related questions, but they concentrate on students in higher education and on critical thinking (Alubthane, 2026; Suazo Galdames et al., 2026). These reviews note that most evidence is cross-sectional and self-reported, and cannot establish long-term decline (Suazo Galdames et al., 2026). That leaves three gaps. First, evidence from professional settings, where skill erosion carries safety and oversight implications, has not been synthesised. Second, prior reviews rarely separate performance during assistance from unaided performance afterwards, a distinction central to learning research (Soderstrom and Bjork, 2015). Third, the temporal structure of exposure (single session, repeated or longitudinal) is seldom coded.

This review addresses these gaps with four research questions:

1) *What short-term effects of LLM assistance on professionals' task performance are reported?*
2) *What evidence exists for longer-term cognitive consequences of repeated LLM use at work?*
3) *Which user, task and interaction factors moderate these effects?*
4) *Through which mechanisms do these effects arise, and which interventions mitigate them?*

The contributions are: (a) the first PRISMA-based synthesis of this evidence restricted to workers and professionals; (b) a coding of studies by exposure structure and by whether unaided outcomes were measured; and (c) a provisional framework linking offloading mechanisms and moderators to longer-term outcomes, with a research agenda.

The remainder of the paper is organised as follows. Section II introduces the conceptual background and related reviews. Section III describes the methodology. Section IV reports results. Section V discusses them by research question. Section VI concludes.

II. BACKGROUND

A. COGNITIVE OFFLOADING

Cognitive offloading is the use of physical action or external tools to reduce the cognitive demands of a task (Risko and Gilbert, 2016). Offloading frees limited internal resources and often improves immediate performance, but it can reduce memory for the offloaded information (Grinschgl, Papenmeier and Meyerhoff, 2021), and the expectation of later access lowers recall of content itself (Sparrow, Liu and Wegner, 2011). LLMs extend offloading from storage and calculation to generation, reasoning and evaluation, the activities through which expertise is built.

B. PERFORMANCE VERSUS LEARNING

Conditions that make practice easier often raise performance during training while impairing retention and transfer (Soderstrom and Bjork, 2015). A study that measures only assisted performance therefore cannot speak to cognitive cost. The outcome that matters here is unaided performance after a delay.

C. AUTOMATION, RELIANCE AND COGNITIVE LOAD

Human-factors research anticipated these concerns. Automating a task can leave operators less practised and less able to intervene when automation fails (Bainbridge, 1983). Reliance can drift into complacency and automation bias (Parasuraman and Manzey, 2010). Appropriate reliance requires trust calibrated to actual system reliability (Lee and See, 2004). Cognitive load theory distinguishes extraneous load, which assistance can usefully remove, from germane processing that builds schemas, which it may also remove (Sweller, 1988). LLMs reduced students' mental effort but also their depth of inquiry (Stadler, Bannert and Sailer, 2024), and raised performance without improving metacognitive accuracy (Fernandes et al., 2026). A meta-analysis found that human-AI combinations did not, on average, outperform the better of human or AI alone (Vaccaro, Almaatouq and Malone, 2024). Cognitive forcing functions can reduce overreliance, though users like them less (Buçinca, Malaya and Gajos, 2021).

D. RELATED REVIEWS

Table 1 summarises the closest prior syntheses. None restricts its population to workers, and none codes whether unaided outcomes were measured.

TABLE 1. Related reviews and how this review differs.

Review	Population	Focus
Suazo Galdames et al. (2026)	Higher education	Offloading vs critical autonomy
Alubthane (2026)	University students	Higher-order cognitive skills; amplifier vs substitute model (89 studies)
Vaccaro, Almaatouq and Malone (2024)	Mixed	Human-AI team performance (meta-analysis)
Macnamara et al. (2024)	Experts and learners	AI-induced skill decay (conceptual)
This review	Workers and professionals	Short-term vs long-term; unaided outcomes; exposure duration

III. METHODOLOGY

This review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement (Page et al., 2021) and the PRISMA-S extension for reporting searches (Rethlefsen et al., 2021). The completed PRISMA checklist, search strings, screening log and extraction form are provided as supplementary material (workbook SLR_PRISMA_workbook.xlsx).

A. SYSTEMATIC REVIEW PLANNING

As far as we are aware, no prior review has synthesised empirical evidence on the cognitive consequences of LLM use specifically among workers. Eligibility followed a PICOS structure (Table 2). Eligible studies reported primary empirical data from adults in paid work or professional practice using an LLM-based tool, with at least one cognitive or performance outcome. Peer-reviewed articles, conference papers and preprints were eligible, with preprints flagged for sensitivity analysis.

Search strategies: OpenAlex, a multidisciplinary index covering journals, proceedings and preprint servers (including SSRN, arXiv and medRxiv), and Europe PMC (including PubMed/MEDLINE) were searched on 22 November 2026. In a second cycle, citation searching from a seed set of known studies identified further records. Scopus, Web of Science, APA PsycINFO and the ACM Digital Library are planned to be searched with the strings in the supplementary workbook and merged in a later version.

Quality assessment: methodological quality is assessed with the Mixed Methods Appraisal Tool (MMAT) (Hong et al., 2018), which covers qualitative, quantitative and mixed-methods designs in one instrument. Appraisal requires full texts and is pending.

TABLE 2. Inclusion and exclusion criteria.

	Include	Exclude
P	Workers, professionals, clinicians incl. residents	Students; general public
I	LLM-based tools	Non-LLM AI or automation
O	Cognitive or performance outcome	Attitudes or adoption only; model benchmarks
S	Primary empirical, any design	Reviews, commentary, protocols
Other	2020 to 2026; English	No abstract or full text obtainable

B. SEARCHING STRINGS

The search combined four concept blocks with Boolean operators: LLM tools; cognitive and performance outcomes; occupational populations; and empirical study designs. To keep the corpus screenable by a single reviewer, LLM and occupational terms were required in titles, and outcome and design terms in titles or abstracts. The OpenAlex string was:

TITLE: ("large language model*" OR LLM* OR ChatGPT OR "GPT-4" OR "generative AI" OR GenAI OR "AI assistant*" OR "AI chatbot" OR Copilot OR "conversational AI") AND (worker* OR employee* OR professional* OR practitioner* OR developer* OR programmer* OR clinician* OR physician* OR resident* OR nurse* OR consultant* OR analyst* OR writer* OR workplace OR work OR job* OR occupational OR organi?ational OR enterprise OR industry …) AND TITLE-ABS: (cognitive offloading OR critical thinking OR metacognit* OR deskilling OR skill decay OR overreliance OR automation bias OR reliance OR cognitive load OR productivity OR task performance OR decision making OR reasoning OR creativity OR expertise OR verification) AND (experiment* OR trial OR randomi* OR user study OR survey OR interview* OR longitudinal OR participants OR mixed methods OR qualitative)

The same logic was adapted to each source's syntax. Full strings for all sources are in the supplementary workbook (sheet 2). Results were limited to articles, conference papers and preprints

published from 1 January 2020. A recall check against six known workforce studies showed that four were missed by the database search (Noy and Zhang, 2023; Brynjolfsson, Li and Raymond, 2025; Dell'Acqua et al., 2023; Shen and Tamkin, 2026). These, and Perry et al. (2023), were added through citation searching. Missing four of six known studies is the main reason we plan to add the subscription databases.

C. SYSTEMATIC REVIEW EXECUTION

Fig. 1 outlines the main phases of the review. The database searches returned 1,559 records (OpenAlex 1,402; Europe PMC 157). After 180 duplicates were removed, three automated rules (listed in the workbook) marked 341 records as ineligible: 43 non-empirical titles, 279 abstracts with no human participants (mostly model benchmarks), and 19 student-only samples. The remaining 1,038 records were screened on title, and 798 were excluded.

Of 240 reports sought, 20 could not be retrieved. The 220 assessed on abstract yielded 152 eligible studies. The 68 exclusions were for wrong outcome (29), non-empirical work (14), wrong population (13), non-LLM exposure (6), no human outcome (5) and duplicate report (1). With five studies from citation searching, the final corpus comprised 157 studies.

Selection process and automation: title and abstract screening was AI-assisted, with a single human reviewer applying the eligibility criteria. Missing abstracts were sought from Crossref. All included and excluded decisions, plus a random 10% of title-stage and automated exclusions, are to be re-screened by the author, with intra-rater κ reported. Data were extracted from titles and abstracts with a structured form. Study design and outcome domains were coded by rule and will be refined at full-text stage. Findings were synthesised narratively following the SWiM guideline (Campbell et al., 2020), as heterogeneity in tasks and measures precluded meta-analysis.

IV. RESULTS

A. PUBLICATION TRENDS AND SOURCES

Most of the evidence is very recent (Fig. 2). One study was published in 2022, 8 in 2023, 10 in 2024 and 35 in 2025. A further 103 (66%) appeared between January and November 2026. Most studies were journal articles (101), followed by preprints (29) and conference papers (27). So almost one in five included studies has not yet been peer reviewed.

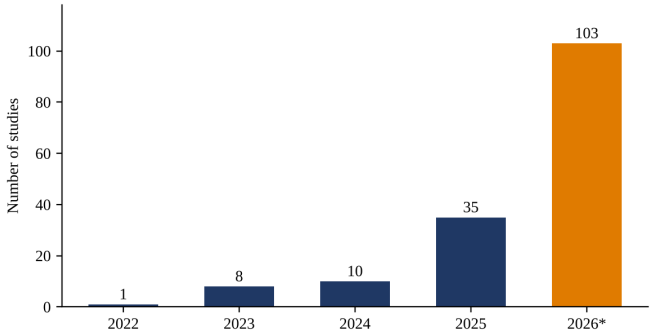


FIGURE 2. Included studies by publication year (n = 157). *January to November.

B. OCCUPATIONAL CONTEXTS

General or mixed knowledge-worker samples were most common (54), followed by software development (39), business, finance and consulting (24), healthcare (23), and creative, writing and design work (13). Customer service and education professionals were each represented by two studies (Fig. 3). The concentration in software and medicine reflects the availability of objective outcomes in those fields: completed tasks, code quality and diagnostic accuracy.

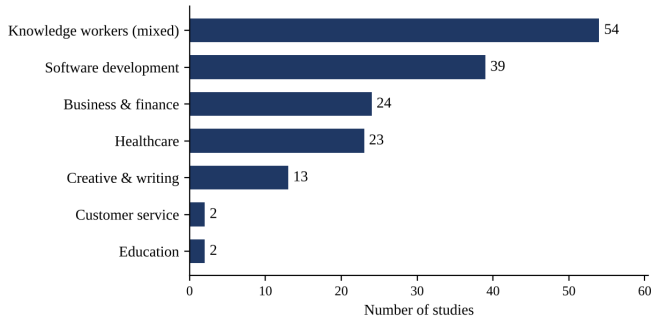


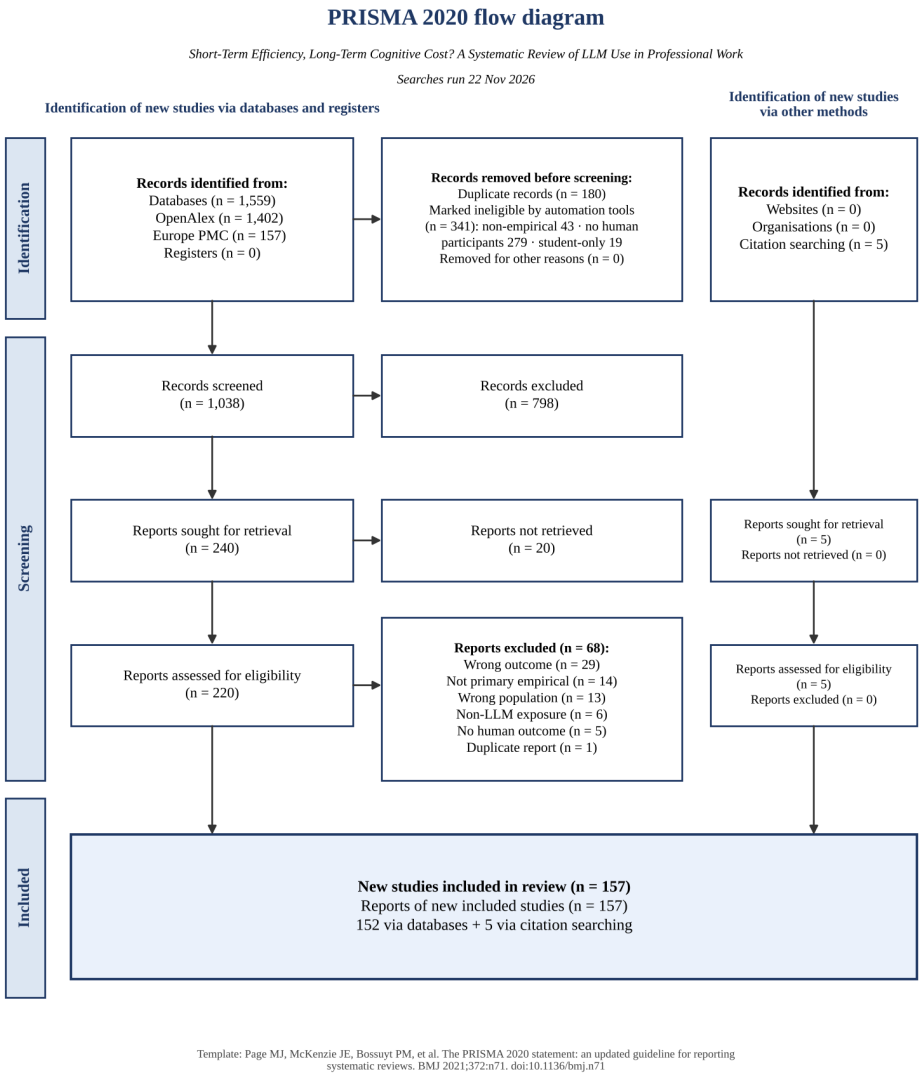
FIGURE 3. Included studies by occupational group.

C. STUDY DESIGNS

Cross-sectional surveys were the largest group (52), typically modelling self-reported productivity or performance with structural equation models. There were 29 randomised experiments and 40 other experiments or user studies. Qualitative (17), mixed-methods (10) and longitudinal or multi-wave designs (8) were less common. Twenty-one studies involved some form of repeated or extended exposure. Very few tested participants without the AI after a period of use.

D. LLM TOOLS

Most abstracts (114) referred to "generative AI" or "LLMs" without naming a product, which limits replication. Where tools were named, ChatGPT or GPT models dominated (35), followed by GitHub Copilot (9). Claude, Gemini and open-weight models appeared once each. Few studies reported model versions, which matters because the models change quickly.



Automation tools: transparent rule-based pre-screening, then AI-assisted title and abstract screening with a single human reviewer. All decisions are logged in the workbook (sheet 5_Screening_log). Subscription-database searches (Scopus, Web of Science, PsycINFO, ACM DL) are pending; counts will be updated after they are merged.

FIGURE 1. Systematic literature review PRISMA 2020 chart.

E. OUTCOME DOMAINS

Productivity or efficiency was the most frequently measured outcome (107 studies), followed by quality or accuracy (67), reliance, trust or verification (56), critical thinking or judgement (46) and creativity (39). Cognitive load (21), metacognition or self-efficacy (23) and memory (4) were rarely measured (Fig. 4). Researchers have measured what LLMs do to the work far more often than what they do to the worker.

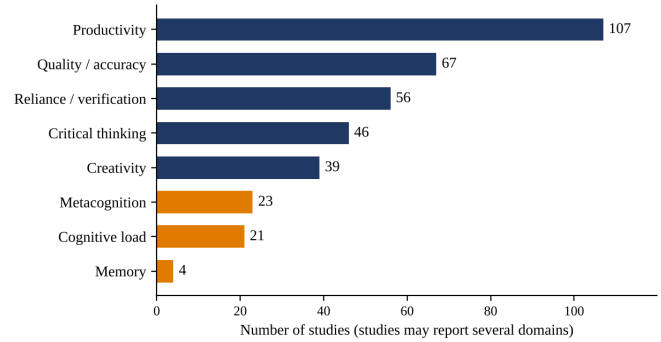


FIGURE 4. Outcome domains measured (studies may report several). Orange marks domains most relevant to long-term cognitive cost.

TABLE 3. Studies with the strongest designs or most direct evidence on cognitive outcomes.

Study	Population	Design	Main finding
Cui et al. (2026)	Developers, three firms	3 field RCTs, N = 4,867	+26% completed tasks; larger gains for less experienced developers
Noy and Zhang (2023)	Writing professionals	Preregistered experiment	Less time, higher quality
Dell'Acqua et al. (2023)	Consultants	Field experiment, N = 758	Gains inside the capability frontier; lower correctness outside it
Goh et al. (2025)	Physicians	RCT	+6.5 points on management reasoning; no better than LLM alone
Qazi et al. (2026a)	Physicians, lower-middle-income country	RCT, N = 58	LLM alone outperformed LLM-assisted physicians by 11.5 points
Stray et al. (2026)	Public-sector developers	Longitudinal case study	No significant change in commit activity after adoption
Shen and Tamkin (2026)	Developers	RCT	Reduced skill formation, except for engaged interaction patterns
Sarfraz et al. (2026)	Knowledge workers	Randomised experiment, N = 90	Free GenAI users could not recall most of what they had written
Lee et al. (2025)	Knowledge workers	Survey, N = 319	Confidence in GenAI associated with less critical thinking
Huang et al. (2026)	Health professionals	Survey with objective test	Perceived capability loss unrelated to objective performance
Qazi et al. (2026b)	AI-trained physicians	RCT, N = 72	Behavioural nudges raised accuracy by 7.6 points on error-seeded cases
Perry et al. (2023)	Developers	User study	Less secure code, greater belief it was secure

V. DISCUSSION

A. ANALYSIS OF RESULTS REGARDING THE FIRST RESEARCH QUESTION

Short-term gains are the best-supported finding in the corpus. Three field RCTs with 4,867 developers found a 26% increase in completed tasks (Cui et al., 2026). Comparable gains were reported for professional writing (Noy and Zhang, 2023), customer support (Brynjolfsson, Li and Raymond, 2025), HR tasks (Katsavou et al., 2026) and a range of UK occupations (Haslberger et al., 2023). In medicine, LLM access improved management reasoning (Goh et al., 2025) and residents' diagnostic accuracy in complex cases (Wu et al., 2025). GenAI use also allowed developers to produce higher-quality code with less cognitive effort (Shan et al., 2026).

The gains have limits. Performance fell on tasks beyond the model's capability (Dell'Acqua et al., 2023), specialised agents added little over a general model (Bekker et al., 2026), and routine telemetry showed no significant change after Copilot adoption (Stray et al., 2026). LLM-assisted physicians frequently failed to match the LLM alone (Goh et al., 2025; Qazi et al., 2026a). This is consistent with meta-analytic evidence on human-AI teams (Vaccaro, Almaatouq and Malone, 2024), and suggests that the bottleneck is the human step of checking and integrating the AI's output.

B. ANALYSIS OF RESULTS REGARDING THE SECOND RESEARCH QUESTION

Direct evidence of long-term cognitive cost is scarce. Fewer than ten studies measured any outcome without the AI after use. None combined repeated workplace exposure with a delayed unaided test. The studies that did test it found costs. Participants who freely used GenAI could not recall most of what they had written (Sarfraz et al., 2026). LLM users showed lower long-term recall despite faster problem solving (Jamshaid, 2025). Developers learning with AI formed weaker skills (Shen and Tamkin, 2026). Disruption of AI access was framed as a source of cognitive fragility (Xu et al., 2026).

Cross-sectional evidence is consistent but weaker. Confidence in GenAI predicted less critical thinking (Lee et al., 2025). Reliance and offloading were negatively associated with workplace critical thinking (De La Rosa Betito et al., 2026). Reliance was associated with lower professional self-efficacy (Siddqiue and Sultan, 2026). However, perceived capability loss among health professionals did not track objective performance (Huang et al., 2026), which cautions against inferring skill loss from self-report. Qualitative accounts describe development, maintenance, erosion and revaluation of skills occurring together (Oviedo-Trespalacios et al., 2026). Longitudinal interviews found perceived accomplishment rising with prolonged use (Kobiella et al., 2024; Kobiella et al., 2025).

C. ANALYSIS OF RESULTS REGARDING THE THIRD RESEARCH QUESTION

Five moderators recur. Gains are largest for novices (Brynjolfsson, Li and Raymond, 2025; Cui et al., 2026), the same group for whom skill-formation costs are most plausible (Macnamara et al., 2024; Shen and Tamkin, 2026). Benefits also depend on whether the task lies within the model's capability and on how complex and ambiguous it is (Dell'Acqua et al., 2023; Haslberger et al., 2023). How people use the tool matters too. Iterative editing produced different results from raw adoption (Katsavou et al., 2026), engaged interaction patterns preserved learning (Shen and Tamkin, 2026), and AI-supported project teams produced more structured but less context-sensitive work (Pompei et al., 2026). AI literacy buffered the link between reliance and self-efficacy (Siddqiue and Sultan, 2026). Finally, confidence in AI and confidence in oneself predicted critical engagement in opposite directions (Lee et al., 2025).

D. ANALYSIS OF RESULTS REGARDING THE FOURTH RESEARCH QUESTION

Three mechanisms have empirical support. The first is reduced encoding of AI-assisted work (Sarfraz et al., 2026; Jamshaid, 2025). The second is miscalibrated confidence: professionals overrated GenAI-assisted outputs (Mayer et al., 2026), and developers believed insecure code was secure (Perry et al., 2023). The third is incomplete verification: many clinicians did not consistently check GenAI-drafted documentation (Stoiacovici et al., 2026). Automation bias was examined in randomised trials with AI-trained physicians (Qazi et al., 2025). In a tax-research setting, errors flowed from humans as well as machines, which complicates simple "human checks AI" rules (Barrick et al., 2026). Some interventions helped. Behavioural nudges improved physicians' accuracy on error-seeded cases (Qazi et al., 2026b). Visible verification cues increased perceived responsible behaviour

in enterprise assistants (Demirtaş and Emre, 2026). Cognitive forcing functions, which reduced overreliance in earlier AI-assisted decision tasks, are another option (Buçinca, Malaya and Gajos, 2021).

E. A PROVISIONAL FRAMEWORK

We propose a framework with four parts. Short-term gains (speed, quality, reduced effort) arise when workers offload generation, reasoning or evaluation. Offloading operates through mechanisms (reduced encoding, reduced effortful generation, reduced verification) whose strength depends on moderators (experience, task fit, mode of use, AI literacy, confidence). Over repeated exposure, these mechanisms may produce longer-term outcomes: skill decay, impaired skill acquisition, weakened evaluative judgement and miscalibrated confidence. The evidence for the first link is good. The links to longer-term outcomes remain untested in workplaces.

F. RECENT ADVANCES AND EMERGING TRENDS

Two shifts are visible in the 2026 literature. First, research is moving from whether LLMs raise productivity to how humans and LLMs combine, with growing attention to verification, automation bias and calibrated reliance (Qazi et al., 2026a; Qazi et al., 2025; Qazi et al., 2026b). Second, intervention studies are emerging: nudges, verification cues, training and interface friction. These studies begin to test how tools could be designed to protect workers' skills. Agentic systems, which act with less human input, will probably make both questions more pressing.

G. LIMITATIONS

The evidence has four main weaknesses: few longitudinal or withdrawal designs, heavy reliance on self-report, frequent omission of model versions, and a large share of preprints. The review process has its own. Searches were limited to two open databases with title-restricted terms, and a recall check showed that relevant studies were missed. Screening was AI-assisted with a single reviewer, and data were extracted from abstracts. Design and outcome categories were rule-coded. These limitations will be addressed by the planned subscription-database searches, full-text extraction, MMAT appraisal and intra-rater reliability checks.

VI. CONCLUSION

This review synthesised 157 empirical studies of LLM use among workers and professionals. The short-term efficiency gain is well supported across occupations, particularly for less experienced workers and for tasks within model capability. A long-term cognitive cost is plausible. A few experiments on memory and skill formation support it, and so do consistent cross-sectional associations, but it has hardly been tested in professional settings. The studies that could settle the question, longitudinal designs with unaided tests, have not yet been done.

Future research should prioritise: (a) longitudinal field studies with repeated, unaided performance tests after a retention interval; (b) designs that separate assisted from unaided outcomes; (c) objective measures of metacognitive calibration alongside self-report; (d) reporting of model versions and modes of use; and (e) trials of interventions that preserve verification and skill without sacrificing productivity.

DECLARATIONS

Funding: none.

Competing interests: the author is co-founder of Inookey, which develops AI agents for organisational workflows.

Use of AI tools: AI assistance was used for search execution, first-pass screening, rule-based coding and drafting. Screening and extraction decisions were made by a single human reviewer working with AI assistance, as described in Section III-C.

Data availability: search strings, screening log, included-study list and extraction form are in the supplementary workbook (SLR_PRISMA_workbook.xlsx).

REFERENCES

Alubthane, F.O. (2026) 'Amplifier or substitute? A systematic review of generative AI's impact on higher-order cognitive skills among university students', *Frontiers in Psychology*. Available at: https://doi.org/10.3389/fpsyg.2026.1863931.

Bainbridge, L. (1983) 'Ironies of automation', *Automatica*, 19(6), pp. 775–779.

Barrick, J.A. et al. (2026) 'Who verifies whom? Generative AI, human oversight, and the production of professional research memoranda'. *SSRN*. Available at: https://doi.org/10.2139/ssrn.7260822.

Bastani, H. et al. (2025) 'Generative AI without guardrails can harm learning: Evidence from high school mathematics', *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. Available at: https://doi.org/10.1073/pnas.2422633122.

Bekker, P. et al. (2026) 'Man, machine, and management consulting: The differential impact of general-purpose GenAI and specialised AI agents on knowledge-work performance'. *Gesellschaft für Informatik*. Available at: https://doi.org/10.18420/akwi2026_01.

Brynjolfsson, E., Li, D. and Raymond, L. (2025) 'Generative AI at work', *The Quarterly Journal of Economics*, 140(2), pp. 889–942. Available at: https://doi.org/10.1093/qje/qjae044.

Buçinca, Z., Malaya, M.B. and Gajos, K.Z. (2021) 'To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making', *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), article 188.

Campbell, M. et al. (2020) 'Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline', *BMJ*, 368, l6890.

Cui, K.Z. et al. (2026) 'The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers', *Management Science*. Available at: https://doi.org/10.1287/mnsc.2025.00535.

De La Rosa Betito, A. et al. (2026) 'ChatGPT reliance and cognitive offloading as predictors of critical thinking in the workplace', *International Journal of Latest Technology in Engineering, Management & Applied Science*. Available at: https://doi.org/10.51583/ijltemas.2026.150900075.

Dell'Acqua, F. et al. (2023) *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality*. Harvard Business School Working Paper 24-013.

Demirtaş, İ. and Emre, İ.E. (2026) 'Responsible response behavior in enterprise AI assistants: The effect of verification cues on trust and reliance intention', in *Proceedings of UBMK 2026*. Available at: https://doi.org/10.1109/ubmk71847.2026.11745308.

Fernandes, D. et al. (2026) 'AI makes you smarter but none the wiser: The disconnect between performance and metacognition', *Computers in Human Behavior*, 175, 108779.

Goh, E. et al. (2025) 'GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial', *Nature Medicine*. Available at: https://doi.org/10.1038/s41591-024-03456-y.

Grinschgl, S., Papenmeier, F. and Meyerhoff, H.S. (2021) 'Consequences of cognitive offloading: Boosting performance but diminishing memory', *Quarterly Journal of Experimental Psychology*, 74(9), pp. 1477–1496.

Haslberger, M. et al. (2023) 'No great equalizer: Experimental evidence on productivity effects of generative AI use in the UK labor market'. *SSRN*. Available at: https://doi.org/10.2139/ssrn.4594466.

Hong, Q.N. et al. (2018) ‘The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers’, *Education for Information*, 34(4), pp. 285–291.

Huang, X. et al. (2026) ‘Large language model use, performance anxiety, and objective medical literature retrieval performance among health care professionals’. *SSRN*. Available at: https://doi.org/10.2139/ssrn.6885430.

Jamshaid, S. (2025) ‘Cognitive effects of large language models: An empirical study of memory retention, reasoning strategies, and critical thinking among students and professionals’, *ACADEMIA International Journal for Social Sciences*. Available at: https://doi.org/10.63056/acad.004.03.0893.

Katsavou, D. et al. (2026) ‘ChatGPT and the HR professional: Experimental evidence on productivity gains in HR tasks’, *Journal of Organizational Effectiveness: People and Performance*. Available at: https://doi.org/10.1108/joepp-07-2025-0642.

Kobiella, C. et al. (2024) ‘“If the machine is as good as me, then what use am I?” How the use of ChatGPT changes young professionals’ perception of productivity and accomplishment’, in *Proceedings of CHI 2024*. Available at: https://doi.org/10.1145/3613904.3641964.

Kobiella, C. et al. (2025) ‘When efficiency meets fulfillment: Understanding long-term LLM integration in knowledge work’, in *Proceedings of CHIWORK 2025*. Available at: https://doi.org/10.1145/3729176.3729180.

Lee, H.-P. et al. (2025) ‘The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers’, in *Proceedings of CHI 2025*. Available at: https://doi.org/10.1145/3706598.3713778.

Lee, J.D. and See, K.A. (2004) ‘Trust in automation: Designing for appropriate reliance’, *Human Factors*, 46(1), pp. 50–80.

Macnamara, B.N. et al. (2024) ‘Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers’ awareness?’, *Cognitive Research: Principles and Implications*, 9, article 46.

Mayer, M. et al. (2026) ‘Generative AI in knowledge work – A breeding ground for the illusion of expertise and the erosion of responsibility?’. *AIS Electronic Library*.

Noy, S. and Zhang, W. (2023) ‘Experimental evidence on the productivity effects of generative artificial intelligence’, *Science*, 381(6654), pp. 187–192. Available at: https://doi.org/10.1126/science.adh2586.

Oviedo-Trespalacios, Ó. et al. (2026) ‘Skill development, maintenance, erosion, and revaluation: How knowledge workers experience generative AI’. *AHFE International*. Available at: https://doi.org/10.54941/ahfe1007458.

Page, M.J. et al. (2021) ‘The PRISMA 2020 statement: An updated guideline for reporting systematic reviews’, *BMJ*, 372, n71. Available at: https://doi.org/10.1136/bmj.n71.

Parasuraman, R. and Manzey, D.H. (2010) ‘Complacency and bias in human use of automation: An attentional integration’, *Human Factors*, 52(3), pp. 381–410.

Perry, N. et al. (2023) ‘Do users write more insecure code with AI assistants?’, in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. Available at: https://doi.org/10.1145/3576915.3623157.

Pompei, A. et al. (2026) ‘How generative AI supports novice project managers in creative reasoning: Evidence from a pilot experiment in risk management’, *International Journal of Managing Projects in Business*. Available at: https://doi.org/10.1108/ijmpb-10-2025-0451.

Qazi, I.A. et al. (2025) ‘Automation bias in large language model assisted diagnostic reasoning among AI-trained physicians’. *medRxiv*. Available at: https://doi.org/10.1101/2025.08.23.25334280.

Qazi, I.A. et al. (2026a) ‘Large language model diagnostic assistance for physicians in a lower-middle-income country: A randomized controlled trial’, *Nature Health*. Available at: https://doi.org/10.1038/s44360-025-00007-8.

Qazi, I.A. et al. (2026b) ‘Mitigating automation bias in physician-LLM diagnostic reasoning using behavioral nudges: A randomized controlled trial’. *medRxiv*.

Rethlefsen, M.L. et al. (2021) ‘PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews’, *Systematic Reviews*, 10, article 39.

Risko, E.F. and Gilbert, S.J. (2016) ‘Cognitive offloading’, *Trends in Cognitive Sciences*, 20(9), pp. 676–688.

Sarfraz, Z. et al. (2026) ‘Thinking outsourced: How generative AI reduces cognitive engagement, memory, and originality in knowledge work’, *Frontiers in Engineering Science & Management*. Available at: https://doi.org/10.68907/fesm.1.1.2026.19.

Shan, G. et al. (2026) ‘Mind and the machine: How does generative artificial intelligence usage affect the coding performance of developers?’, *Journal of Management Information Systems*. Available at: https://doi.org/10.1080/07421222.2026.2692282.

Shen, J.H. and Tamkin, A. (2026) ‘How AI impacts skill formation’. *arXiv:2601.20245*.

Siddqiue, M.U. and Sultan, S.A. (2026) ‘When AI helps but humans feel less capable: The impact of generative AI reliance on professional self-efficacy and meaningful work’, *Journal of Excellence in Social Sciences*. Available at: https://doi.org/10.69565/jess.v5i1.395.

Soderstrom, N.C. and Bjork, R.A. (2015) ‘Learning versus performance: An integrative review’, *Perspectives on Psychological Science*, 10(2), pp. 176–199.

Sparrow, B., Liu, J. and Wegner, D.M. (2011) ‘Google effects on memory: Cognitive consequences of having information at our fingertips’, *Science*, 333(6043), pp. 776–778.

Stadler, M., Bannert, M. and Sailer, M. (2024) ‘Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry’, *Computers in Human Behavior*, 160, 108386.

Stoiacovici, C.A. et al. (2026) ‘Trusted but not fully verified: Generative AI use and exposure to errors in clinical documentation among healthcare professionals’, *Healthcare*. Available at: https://doi.org/10.3390/healthcare14213660.

Stray, V.G. et al. (2026) ‘Developer productivity with and without GitHub Copilot: A longitudinal mixed-methods case study’, in *Proceedings of HICSS 2026*. Available at: https://doi.org/10.24251/hicss.2026.880.

Suazo Galdames, I.C. et al. (2026) ‘Between cognitive offloading and critical autonomy: A systematic review of the epistemic implications of generative AI in higher education’, *Frontiers in Education*. Available at: https://doi.org/10.3389/feduc.2026.1878665.

Sweller, J. (1988) ‘Cognitive load during problem solving: Effects on learning’, *Cognitive Science*, 12(2), pp. 257–285.

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(12), pp. 2293–2303.

Wu, X.C. et al. (2025) ‘A large language model improves clinicians’ diagnostic performance in complex critical illness cases’, *Critical Care*. Available at: https://doi.org/10.1186/s13054-025-05468-7.

Xu, L.Z. et al. (2026) ‘When ChatGPT is down, so are our brains: Cognitive fragility following AI disruption in knowledge work’. *AIS Electronic Library*.



CHETNA KHANDARE is the co-founder of Inookey AI Technology Private Limited, Birmingham, U.K., which develops AI agents for organisational workflows. Her research interests include large language models, human–AI interaction, and the cognitive and performance effects of generative AI in professional work.