JETISH: Journal of Education Technology Information Social Sciences and Health
E-ISSN: 2964-2507 P-ISSN: 2964-819X
Vol. 6 No. 1 March 2027

Feasible but Unproven: A Systematic Review of Generative AI and Agentic Copilots in Administrative Work

Irfan Taufik[1] Rasto[2] Edi Suryadi[3]
Business Administration Study Program, Graduate School, Universitas Pendidikan Indonesia, Bandung City, West Java Province, Indonesia[1,2,3]
Email: irfantaufik@upi.edu[1] rasto@upi.edu[2] edi_suryadi@upi.edu[3]

Abstract
Generative AI copilots and agentic systems are increasingly proposed to reduce administrative work, yet evidence on their effects has not been appraised across sectors. This systematic review asked which administrative tasks and technologies have been studied, which effects on productivity, quality, and errors are reported, which human–AI collaboration mechanisms accompany them, and which risks and governance measures are reported. Following PRISMA 2020 with an unregistered protocol, we searched Scopus on 30 October 2026 for English-language journal articles published in 2022–2026 (314 records), appraised quality with the Mixed Methods Appraisal Tool, and synthesized findings narratively using context–intervention–mechanism–outcome logic. Twenty-nine studies from health care, private organizations, public administration, and higher education were included; none was randomized, and one evaluated an explicitly agentic system. Of 19 rated studies, three were high quality and seven low. The largest reductions in task time (60–90%) were reported only by system-development studies that were not quality-rated, typically single cases or internal tests. Accuracy results were mixed and lowest for medical coding requiring exact matches (ICD-10 concordance 58.1%). The evidence shows feasibility under supervision, not yet organizational benefit; designs that attribute effects and measure work quality, downstream errors, and oversight are needed.
Keywords: Agentic AI; Generative AI Copilots; Administrative Work; Human Oversight; Automation Levels; Evidence Synthesis

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

INTRODUCTION

Administrative work is a large and measurable cost. In one US academic health system, billing and insurance-related activities cost about US$20 for a primary care visit and about US$215 for an inpatient surgical procedure, or 3.1–25.2% of professional revenue depending on the service (Tseng et al., 2018). In public administration, the learning, psychological, and compliance costs of dealing with government constitute an administrative burden whose size reflects organizational and political choices (Moynihan et al., 2015). Software has long targeted this work. Robotic process automation (RPA) uses bots that mimic user actions on structured tasks, although academic research trailed vendor adoption (Syed et al., 2020). RPA is poorly suited, however, to reading a free-text request, drafting a reply, or reconciling an unstructured document. Large language models (LLMs) lift that limit. Analyses of labor-market exposure assessed their potential to affect occupations (Eloundou et al., 2024), and adoption followed: in a survey of about 18,000 Danish workers in 11 occupations, use was widespread, especially among younger and less experienced workers, and women were 16 percentage points less likely to have used ChatGPT for work (Humlum & Vestergaard, 2025). Evidence on performance comes mainly from customer support, where a generative AI assistant raised issues resolved per hour by 14% (Brynjolfsson et al., 2025), and from randomized writing tasks, where completion time fell by about 40% (Noy & Zhang, 2023). These settings resemble administrative work but are not the same, and whether such effects transfer to coding, procurement, human-resources, or policy-information tasks is unknown.

The technology is changing as well. "Copilot"-type assistants draft, summarize, or retrieve at a user's request, whereas agentic AI systems plan multistep tasks, use tools, and act with limited human oversight (Acharya et al., 2025; Kumar et al., 2026; Wang et al., 2024). Where the human sits along this spectrum is a design decision with consequences. The levels-of-automation framework separates degrees of human involvement across stages of information processing (Parasuraman et al., 2000), and the automation–augmentation paradox holds that the two cannot be cleanly separated in organizations (Raisch & Krakowski, 2021). Labels such as "copilot" or "agent" therefore say little about actual autonomy, which must be established from study design. Reliability and oversight remain open problems. Larger, instruction-tuned models more often give plausible but wrong answers instead of declining (Zhou et al., 2024), hallucination persists in retrieval-augmented professional tools (Magesh et al., 2025), and exact medical coding remains unreliable (Soroush et al., 2024). The usual safeguard is human review, yet human–AI combinations do not reliably outperform the better of human or AI alone (Vaccaro et al., 2024), and policies mandating human oversight of government algorithms rest on an untested assumption that people can perform it (Green, 2022). Acceptance and governance also constrain deployment. Public servants favor AI in supportive rather than decisive roles (Haesevoets et al., 2025); public-sector adoption involves a tension between building AI capability apart from the organization and integrating it (Selten & Klievink, 2024); generative AI raises distinct questions of performance, accountability, and democratic governance (Porumbescu et al., 2025); and governance maturity varies across public organizations (de Almeida & dos Santos Júnior, 2025).

Evidence syntheses have not yet established whether generative AI works in administrative work. A systematic review of health-care LLM research found administrative tasks among the least evaluated applications (Bedi et al., 2025). A review of real-world LLM use in clinical workflows identified only four eligible studies (Artsi et al., 2025). A review of 125 articles on public-sector digital transformation described a widening gap between technological developments and scholarly inquiry on generative AI (Jonathan, 2026). Because this literature centers on clinicians and patients, it offers limited guidance on the staff who perform administrative work, whose tasks, error costs, and accountability structures differ. Primary studies in health-care documentation also diverge by design: an uncontrolled evaluation reports benefits (Olson et al., 2025), whereas a randomized evaluation found no reduction in reply time (Tai-Seale et al., 2024). To our knowledge, no synthesis has appraised empirical evidence on generative AI across administrative work in health care, public administration, higher education, and private organizations, or distinguished assistive from agentic designs. This systematic review addresses four questions: which administrative tasks and generative-AI technologies have been studied (RQ1); which effects on productivity, quality, and errors are reported (RQ2); which human–AI collaboration mechanisms accompany these effects (RQ3); and which risks and governance measures are reported (RQ4). Following PRISMA 2020 (Page et al., 2021) and a context–intervention–mechanism–outcome logic (Denyer et al., 2008), we synthesize empirical journal articles indexed in Scopus for 2022–2026, appraise their methodological quality, and relate the pattern of evidence to the degree of automation. The review contributes an evidence map of tasks and technologies, a quality-weighted reading of reported effects, and an agenda for the designs needed to move from demonstrations of feasibility to evidence of organizational benefit.

RESEARCH METHODS

Review Design and Reporting

We conducted a systematic review reported in accordance with PRISMA 2020 (Page et al., 2021) and, for the search, PRISMA-S, the extension for reporting literature searches (Rethlefsen

et al., 2021). Review questions were structured by context–intervention–mechanism–outcome (CIMO) logic (Denyer et al., 2008): which administrative tasks and generative-AI technologies have been studied (RQ1); which effects on productivity, quality, and errors are reported (RQ2); which human–AI collaboration mechanisms accompany these effects (RQ3); and which risks and governance measures are reported (RQ4). A protocol (v0.2) was written before searching but, contrary to its plan, was not registered; it is available from the corresponding author on request. Departures from it were: a second, stricter screening pass; Scopus as the only database, without citation chasing or benchmark validation of the search string; no pilot round, so the protocol's agreement threshold ($\kappa \geq 0.61$) was not applied; direct exclusion of inaccessible reports without contacting authors; and Synthesis Without Meta-analysis (SWiM) and Thomas and Harden (2008) for synthesis in place of the frameworks named in the protocol. The use of AI-assisted tools in this study is described in the Declaration of generative AI and AI-assisted technologies in the writing process.

Information sources and search

Scopus, chosen for its curated coverage (Baas et al., 2020), was searched on 30 October 2026. A Boolean string combined generative-AI and agent terms with administrative-work terms in titles, abstracts, and keywords, and was limited to final journal articles in English published in 2022–2026 (Box 1). No other databases, gray literature, or citation chasing were used. Records were exported as CSV and checked for duplicates by DOI, EID, and normalized title (Bramer et al., 2016).

Box 1. Scopus search string (Advanced search; executed 30 October 2026; 314 records)

TITLE-ABS-KEY (("agentic AI" OR "agentic artificial intelligence" OR "agentic system*" OR "AI agent*" OR "LLM agent*" OR "autonomous agent*" OR "intelligent agent*" OR "generative AI" OR "generative artificial intelligence" OR "large language model*" OR "AI assistant*" OR "AI copilot*" OR copilot* OR co-pilot* OR "virtual assistant*" OR chatgpt) AND ("office work*" OR "office task*" OR "office automation" OR "office productivity" OR "office suite*" OR "administrative work*" OR "administrative task*" OR "administrative process*" OR "administrative support" OR "administrative professional*" OR "office administrat*" OR clerical OR "back-office" OR "back office" OR "business process automation" OR "workflow automation" OR "intelligent process automation" OR hyperautomation OR "Microsoft 365" OR "Office 365" OR "Google Workspace")) AND PUBYEAR > 2021 AND PUBYEAR < 2027 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (SRCTYPE , "j")) AND (LIMIT-TO (PUBSTAGE , "final"))

The first block covers generative-AI and agent terms, the second administrative-work terms; outcome terms were left out to preserve sensitivity and were applied at screening.

Eligibility criteria and study selection

Eligibility criteria are summarized in Table 1. Selection proceeded in three stages (Figure 1): title–abstract screening in an inclusive pass (doubt resolved toward inclusion), a second, stricter title–abstract pass restricted to records retained in the first (doubt resolved toward exclusion), and full-text assessment. Full texts were obtained by the authors; reports that could not be accessed were excluded and counted separately.

Table 1. Eligibility criteria applied at the second title–abstract pass and full-text stages

Domain	Inclusion	Exclusion
Context	Administrative or office work in organizations (e.g., drafting, email, minutes, data entry, records, reporting, back-office processes, office suites) as an explicit focus	Clinical, educational, manufacturing, software-engineering, or customer-service settings without administrative processes; administrative use as

		one of several uses; patients, students, or teachers as end users
Technology	Generative AI or large language model (LLM), AI assistant or copilot, or LLM-based agent	Non-generative AI, rule-based automation, predictive models
Publication	Peer-reviewed journal article, final stage, English, 2022–2026	All other

Each record was screened by two parallel reviewers working independently: R1 applied the criteria to the record text, and R2, a rule-based lexical classifier, did not use R1 output. A third reviewer (R3) resolved disagreements. Decisions were binary and carried a reason code. Agreement between R1 and R2 was quantified with Cohen's kappa (Cohen, 1960); kappa indicates agreement between R1 and a lexical benchmark rather than inter-rater reliability.

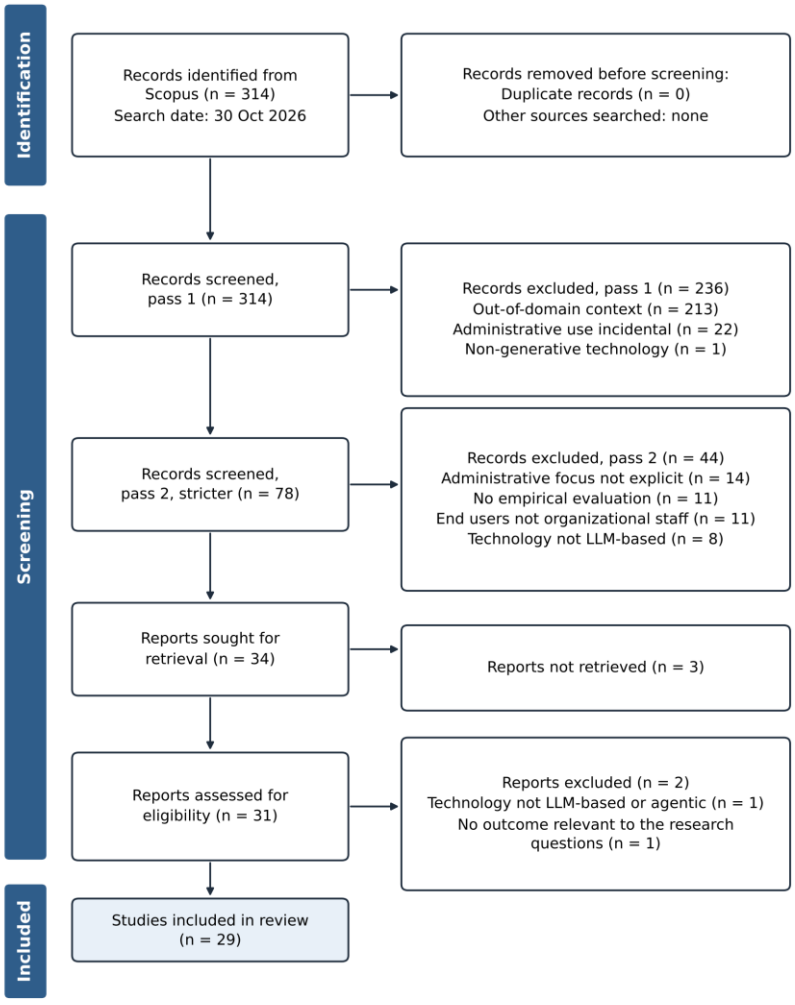


Figure 1. PRISMA 2020 flow diagram of the identification, screening, and inclusion of studies

Data extraction

Data were extracted from full texts with a structured form covering country, design, sample, sector, task category, AI technology, autonomy and oversight model, CIMO elements, outcome measures, main findings, direction of effect, risks, governance measures, limitations, and funding. Missing information was recorded as "not reported", not imputed. Where study authors did not state an automation level, the extracting reviewer assigned one using the levels of Parasuraman et al. (2000) and flagged it as inferred. Extraction was performed by one

reviewer, and a second pass re-checked all extracted values against the full texts; discrepancies would have been resolved by consensus.

Quality appraisal

Empirical studies were appraised with the Mixed Methods Appraisal Tool (MMAT) version 2018 (Hong et al., 2018), using the qualitative, non-randomized, and quantitative descriptive categories. Each criterion was rated Yes, No, or Cannot tell; for interpretation only, five Yes ratings were classed as high, three or four as moderate, and two or fewer as low. Studies reporting system development or technical evaluation fall outside MMAT's design categories and were left unrated and analyzed as a separate stratum. Appraisal was performed by a single reviewer, followed by a second appraisal pass on all rated studies; agreement was summarized with Cohen's kappa, and disagreements were resolved by the rule that unreported information is rated "cannot tell". No study was excluded on quality.

Synthesis

For RQ1, studies were assigned to mutually exclusive categories by primary focus (task, technology class, sector, region, design) and tabulated as frequencies. For RQ2, we conducted a narrative synthesis structured by the SWiM guideline (Campbell et al., 2020), grouping studies by outcome type and direction of effect. Effects were taken as reported by the original authors and not recalculated. Meta-analysis was prespecified only for three or more studies with comparable effect sizes; this was not met, given heterogeneous designs and outcomes. A narrative sensitivity analysis excluded low-rated and unrated studies. Strength of evidence was judged narratively from design, appraisal rating, consistency, and directness, without formal GRADE; reporting bias was not tested because too few studies were comparable. For RQ3 and RQ4, reported mechanisms, risks, and governance measures were grouped thematically, adapting Thomas and Harden (2008), and mapped to a CIMO matrix. Counts and agreement statistics were computed in Python 3.13 (pandas 3.0.5, scikit-learn 1.9.1).

RESEARCH RESULTS AND DISCUSSION

Results

Study selection

The Scopus search returned 314 records, with no duplicates (Figure 1). The inclusive title–abstract pass excluded 236 records and retained 78; the stricter second pass excluded a further 44 (administrative focus not explicit, 14; no empirical evaluation, 11; end users not organizational staff, 11; non-generative technology, 8), leaving 34 reports for retrieval. Three could not be retrieved and were excluded as inaccessible; 31 were assessed in full text. Two were excluded: one used OCR-based computer-vision "agents" without a language model, and one evaluated generative-AI literacy training without reporting work tasks or outcomes. Twenty-nine studies were included. Agreement between R1 and the lexical benchmark R2 was low to moderate at title–abstract screening (round 1: $\kappa = 0.356$, 81 conflicts; round 2: $\kappa = 0.457$, 20 conflicts) and low at full text ($\kappa = 0.075$, 74.2% raw agreement, 8 conflicts). The adjudicator (R3) upheld R1 in all 20 round-2 conflicts and in 7 of 8 full-text conflicts. These values describe agreement between R1 and a lexical benchmark, not inter-rater reliability.

Characteristics and methodological quality of included studies

Table 2 lists the 29 studies. Twenty were published in 2026, five in 2025, and four in 2024; none dated from 2022 or 2023. Europe (n = 9), Asia (n = 9, including one study from the United Arab Emirates), and North America (n = 8) were represented almost equally, South America

once (Peru), and the setting was not reported in two studies (regions follow the study setting where stated; for five further studies marked † in Table 2, only the authors' affiliation country was available). Studies came from health care (n = 10), private organizations (n = 9), public administration (n = 6), and higher education (n = 4). Designs were system development or evaluation (n = 10), evaluation of LLM output against a reference standard (n = 6), quantitative descriptive studies (surveys, one chat-log analysis, one expert-panel assessment; n = 6), qualitative studies (n = 4), and non-randomized quasi-experiments (n = 3). No randomized trial was found. Design and sector were partly confounded. All six evaluations against a reference standard were set in health care, which also contributed one qualitative, two descriptive, and one system study; the 10 system-development studies came mostly from private organizations (n = 5) and higher education (n = 3); and two of the three quasi-experiments were in public administration (Kim, 2026; Rodriguez Saavedra et al., 2026), the third in a university (López-Galisteo, 2026). Sample sizes spanned several orders of magnitude, from 29 simulated scenarios (Aiumtrakul et al., 2026), 31 administrators (Chen et al., 2025), and 31 interviews with 26 participants (Tahtali et al., 2026) to 1,377 hospital administrative staff (Du et al., 2026), 428 civil servants (Rodriguez Saavedra et al., 2026), 2,738 radiology reports (Ruhwedel et al., 2026), 32,950 queries (Klang et al., 2026), and 30,503 chat threads (Black et al., 2026). One ethnography accumulated 476 observation hours over 2.5 years (Choroszewicz, 2026). Most studies were conducted in a single organization.

Table 2. Characteristics of the 29 included studies

Study	Country	Sector	Design	Primary task	Tech.[a]	MMAT[b]
Chen et al. (2025)	China	Health	Qualitative interviews	Drafting/communication	A	Y Y Y Y Y (High)
Wilson et al. (2025)	USA	Higher education	Case study (proof of concept)	Policy information	C	n/a
Aiumtrakul et al. (2026)	NR	Health	LLM evaluation vs. reference	Payer documents/coding	A	CT N Y CT Y (Low)
Jo & Park (2024)	South Korea	Private	Survey (PLS-SEM)	General office use	A	Y CT Y CT Y (Moderate)
Jo (2024)	South Korea	Private	Survey (SEM)	General office use	A	Y CT Y CT Y (Moderate)
Mitroulias & Sioutas (2026)	Greece†	Private	System development and evaluation (RAG)	Policy information	C	n/a
Rodriguez Saavedra et al. (2026)	Peru	Public	Quasi-experiment (DiD + PSM)	Workflow automation	D	CT N CT Y CT (Low)
López-Galisteo (2026)	Spain	Higher education	Quasi-experiment (longitudinal case)	Drafting/communication	C	CT Y Y N CT (Low)
Kim (2026)	South Korea	Public	Quasi-experiment (DiD + PSM)	Drafting/communication	C	CT Y CT Y CT (Low)
Tahtali et al. (2026)	Netherlands	Public	Qualitative interviews	Drafting/communication	A	Y Y Y Y Y (High)

Study	Country	Sector	Design	Primary task	Tech.[a]	MMAT[b]
Jo & Park (2025)	South Korea	Private	Survey (PLS-SEM)	General office use	A	Y CT Y CT Y (Moderate)
Wang et al. (2026)	China†	Health	System evaluation (proof of concept)	Payer documents/coding	D	n/a
Zaki et al. (2026)	UAE	Higher education	Multi-case framework evaluation	Workflow automation	D	n/a
Kiser et al. (2024)	USA	Health	LLM evaluation (blinded raters)	Payer documents/coding	A	CT N Y CT Y (Low)
Nai et al. (2025)	Europe (multi-country)	Public	Case study (process mining)	Workflow automation	D	n/a
Klang et al. (2026)	USA	Health	LLM evaluation (EHR data)	Administrative data	A	Y Y Y CT Y (Moderate)
Doropoulos et al. (2026)	Greece†	Private	System development and benchmark	Workflow automation	E	n/a
Carroll et al. (2026)	USA	Health	LLM + RAG vs. expert	Drafting/communication	C	Y Y Y CT Y (Moderate)
Callari & Puppione (2025)	Europe (multinational firm)	Private	Qualitative (open-ended survey)	Drafting/communication	B	Y CT Y Y Y (Moderate)
Kalasapudi et al. (2026)	USA†	Private	Perspective and tool case	Workflow automation	C	n/a
Menano de Carvalho et al. (2026)	Portugal†	Private	Framework and two applications	Workflow automation	D	n/a
Choroszewicz (2026)	Finland	Public	Ethnography	General office use	B	Y Y Y Y Y (High)
Ruhwedel et al. (2026)	Germany	Health	LLM evaluation (ICD-10)	Payer documents/coding	A	Y CT Y CT Y (Moderate)
Soma (2026)	NR	Private	System evaluation (three deployments)	Workflow automation	D	n/a
Latario & Fowler (2026)	USA	Health	Agreement study	Payer documents/coding	A	CT CT Y CT Y (Low)
Black et al. (2026)	USA	Health	Retrospective chat-log analysis	General office use	A	Y CT Y CT Y (Moderate)
Goyal et al. (2026)	India	Public	Expert-panel assessment (FBWM–DEMATEL)	General office use	A	CT CT Y CT Y (Low)
Paek et al. (2024)	USA	Higher education	System design and simulation	Workflow automation	D	n/a

Study	Country	Sector	Design	Primary task	Tech.[a]	MMAT[b]
Du et al. (2026)	China	Health	Survey (PLS-SEM)	General office use	A	Y Y Y CT Y (Moderate)

Note. NR = not reported; † country of authors' affiliation (study setting not stated). [a] Technology class: A = general-purpose chat LLM; B = office-suite copilot; C = custom assistant, retrieval-augmented generation (RAG), or small model; D = LLM embedded in a workflow or pipeline; E = multi-agent or explicitly agentic. [b] Mixed Methods Appraisal Tool (MMAT 2018): ratings of criteria 1–5 for the applicable design category (Y = yes, N = no, CT = cannot tell) and the overall rating; n/a = system development or technical evaluation, not rated. DiD = difference-in-differences; PSM = propensity score matching; PLS-SEM = partial least squares structural equation modeling; SEM = structural equation modeling; FBWM–DEMATEL = fuzzy best-worst method with decision-making trial and evaluation laboratory; EHR = electronic health record; ICD-10 = International Classification of Diseases, 10th Revision; UAE = United Arab Emirates.

MMAT appraisal applied to 19 studies; 10 system-development or technical-evaluation studies were left unrated. Of the 19 rated studies, three were high (Chen et al., 2025; Tahtali et al., 2026; Choroszewicz, 2026), nine moderate, and seven low (Aiumtrakul et al., 2026; Rodriguez Saavedra et al., 2026; Kiser et al., 2024; Latario & Fowler, 2026; Goyal et al., 2026; Kim, 2026; López-Galisteo, 2026). Recurrent weaknesses were perception-based outcomes (Rodriguez Saavedra et al., 2026), no control of confounders in a cross-period comparison (López-Galisteo, 2026), small simulated samples in three evaluations, unreported survey non-response in four surveys, and single-site data in most studies; the longest observation period was 12 months (Rodriguez Saavedra et al., 2026). A second appraisal pass agreed on the overall rating for 17 of 19 studies ($\kappa = 0.83$) and on 84 of 95 criterion ratings (88.4%; $\kappa = 0.77$); because it was not independent of the first pass, these values are likely optimistic. The two disagreements were quasi-experiments that the first pass rated moderate and the second low (Kim, 2026; López-Galisteo, 2026), because representativeness, completeness of outcome data, or delivery of the intervention was not reported. Disagreements were resolved by applying the protocol rule that unreported information is rated "cannot tell"; both studies are therefore rated low in Table 2 and throughout. Criterion-level ratings are given in Table 2. The second extraction pass found no discrepancy in country, design, sample, sector, task, or technology class, and all quantitative values matched the source texts.

RQ1: Which tasks and technologies have been studied

Eight studies addressed workflow or document-processing automation, seven general office use or adoption without a specific task, six drafting or communication, five payer documents, coding, or billing in health care, two policy or procedure information assistants, and one an administrative data task (Figure 2). Workflow automation covered invoice and customer-service processing (Soma, 2026), customs and shop-floor documents (Menano de Carvalho et al., 2026), human-resources processes (Zaki et al., 2026), construction submittals (Kalasapudi et al., 2026), procurement event logs (Nai et al., 2025), project-management conversations (Doropoulos et al., 2026), group administrative tasks (Paek et al., 2024), and the classification of requests (Rodriguez Saavedra et al., 2026). Drafting and communication ranged from daily documents (Chen et al., 2025), email inquiries (López-Galisteo, 2026), and written answers to citizen questions (Kim, 2026) to tender documents (Tahtali et al., 2026), support-ticket replies (Carroll et al., 2026), and meeting minutes and presentations (Callari & Puppione, 2025). Payer-related work comprised prior-authorization letters (Aiumtrakul et al.,

2026), appeals against denied radiotherapy services (Kiser et al., 2024), and diagnostic or billing coding (Latario & Fowler, 2026; Ruhwedel et al., 2026; Wang et al., 2026). The two policy-information assistants served research administration (Wilson et al., 2025) and payroll (Mitroulias & Sioutas, 2026); the remaining task was structured queries on electronic health records (Klang et al., 2026).

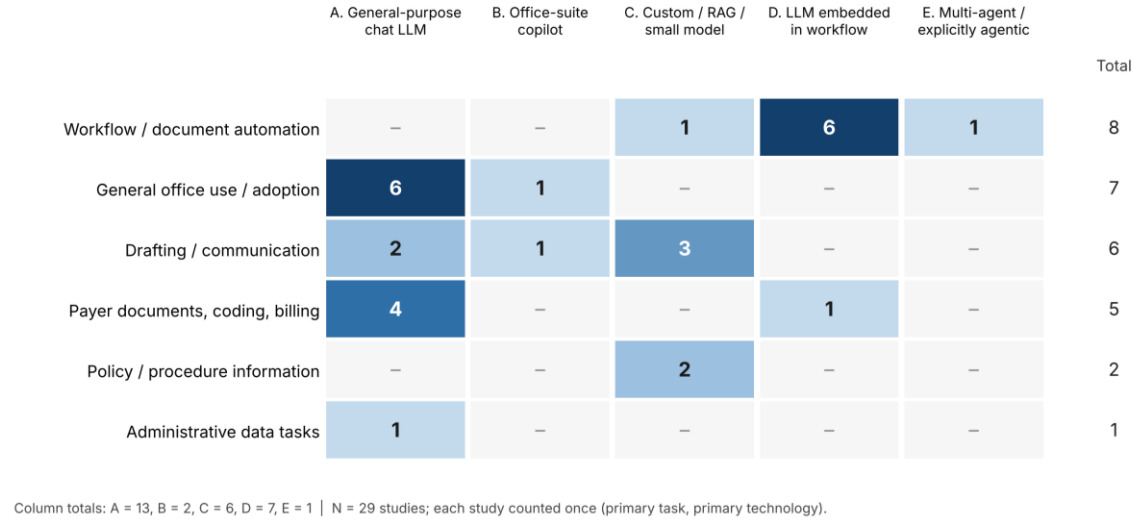


Figure 2. Evidence map of the 29 included studies by primary administrative task and AI technology class

General-purpose chat LLMs used as supplied (e.g., ChatGPT, Gemini) accounted for 13 studies and dominated general office use (6 of 7) and payer documentation (4 of 5). LLMs embedded in automation pipelines (n = 7) were concentrated in workflow automation (6 of 8). Custom assistants, RAG, or small models appeared in six studies, and a Microsoft 365 Copilot-type tool in two (Callari & Puppione, 2025; Choroszewicz, 2026). Models named included ChatGPT-5 (Aiumtrakul et al., 2026; Ruhwedel et al., 2026), GPT-4 or GPT-4o (Carroll et al., 2026; Rodriguez Saavedra et al., 2026; Soma, 2026), GPT-3.5 variants (Kiser et al., 2024; Paek et al., 2024), Gemini (Latario & Fowler, 2026), GPT-4o with Claude Sonnet 4 (Zaki et al., 2026), and nine models compared in one evaluation (Klang et al., 2026). Only one study evaluated an explicitly multi-agent, agentic system, a benchmark on synthetic project-management conversations (Doropoulos et al., 2026). Interaction mode was described by the study authors in eight studies, inferred by the extracting reviewer in 16, and not assessable in five adoption surveys. Where it could be judged, the AI drafted, suggested, or processed content while a person decided or validated, corresponding to intermediate levels of automation (Parasuraman et al., 2000). Higher automation appeared in embedded workflows: full human review of all outputs for the first three months and 10% sampling thereafter (Rodriguez Saavedra et al., 2026), and end-to-end processing with rule-based or expert validation (Soma, 2026; Wang et al., 2026; Zaki et al., 2026). No study reported fully autonomous execution without human review.

RQ2: Effects on productivity, quality, and errors

Table 3 summarizes outcomes. No outcome was pooled; designs and metrics were not comparable.

Table 3. Reported effects of generative AI on efficiency and accuracy, by study quality

Outcome	Study	MMAT	Reported result
Time/ efficiency (objective)	Kim (2026)	Low	Answer drafting 3.81 min faster (difference-in-differences); 4.05 min after regression adjustment; newer staff benefited most

	López-Galisteo (2026)	Low	Email inquiries down by 115 (−21.4%) vs. same period of prior year
	Wang et al. (2026)	n/a	Documentation time 25.0 → 3.2 min (−88%); single-case proof of concept
	Zaki et al. (2026)	n/a	Process steps −59.6%, execution time −89.6%, ROI 359% (five processes)
	Soma (2026)	n/a	Responses 62% faster; cost per transaction −53% (internal benchmark)
Efficiency (perceived)	Rodriguez Saavedra et al. (2026)	Low	Absenteeism −9.4%, processing time −8.7%, administrative cost −18.2% (civil-servant perceptions)
	Kalasapudi et al. (2026)	n/a	"Hours to minutes" (practitioner statement; no measurement)
Accuracy: drafting	Aiumtrakul et al. (2026)	Low	93.1% of letters cited valid references; 89.7% rated strong clinical reasoning; code and citation weaknesses
	Kiser et al. (2024)	Low	97–100% of scores judged letters likely to speed clerical work; citations unreliable
	Carroll et al. (2026)	Moderate	Ticket categorization κ ≈ 0.85; about two-thirds of replies correct for easy–moderate tickets
Accuracy: coding/classification	Ruhwedel et al. (2026)	Moderate	Exact ICD-10 concordance 58.1% (1,590/2,738 reports)
	Latario & Fowler (2026)	Low	68% agreement, κ = 0.586; billing level over-assigned 31 times, under-assigned 10 times
Accuracy: data tasks	Klang et al. (2026)	Moderate	Direct prompting low accuracy; tool-based (code) execution much higher; accuracy fell with large tables
Accuracy: systems	Doropoulos et al. (2026); Paek et al. (2024); Wang et al. (2026); Soma (2026)	n/a	81.3% (F1 = 0.845); 87% task success and 89% language understanding; 98.3% clinical and 96.7% administrative entity extraction; 92% accuracy

Note. MMAT = Mixed Methods Appraisal Tool; n/a = not rated (system development or technical evaluation). Results are as reported by the original authors and were not recalculated. ROI = return on investment.

The three quasi-experiments differed in scope: 80 civil servants in one central-government agency (Kim, 2026), 428 civil servants in 20 public organizations, 10 treated and 10 control (Rodriguez Saavedra et al., 2026), and four months of email volume in one university program (López-Galisteo, 2026). Seven studies reported time or efficiency outcomes, all favorable: six with quantified results and one practitioner statement. Objective evidence from rated designs came from two quasi-experiments rated low (Kim, 2026; López-Galisteo, 2026); the largest reductions (60–90%) came from unrated system studies, several by the developers themselves and reported as single cases or internal benchmarks (Soma, 2026; Wang et al., 2026; Zaki et al., 2026). The only rated study reporting organizational outcomes relied on perceptions and was rated low (Rodriguez Saavedra et al., 2026). Accuracy results were mixed and task-dependent. Drafting of payer letters and ticket replies was judged useful, whereas precise coding was weaker: exact ICD-10 concordance was 58.1% and billing-level agreement 68% (Latario & Fowler, 2026; Ruhwedel et al., 2026). Implementation mattered: tool-assisted or retrieval-supported designs outperformed direct prompting (Carroll et al., 2026; Klang et al., 2026). Errors in codes, billing levels, and citations were the most frequent failure reported. Quality of finished work, cost of human oversight, and downstream error rates were not measured in any study. Overall strength of evidence is low to moderate: no randomized design, three short quasi-experiments, mostly single-site studies, and favorable results concentrated in unrated or developer-reported work. Excluding low-rated and unrated studies left no objective efficiency study, because both quasi-experiments were rated low, and three accuracy studies,

all mixed (Carroll et al., 2026; Klang et al., 2026; Ruhwedel et al., 2026); the mixed direction for accuracy did not change. Publication bias could not be assessed with so few comparable studies. Studies of adoption and perceptions reported mostly favorable but heterogeneous patterns. Among office workers in South Korea, perceived ChatGPT characteristics were associated with knowledge acquisition, knowledge application, satisfaction, and word-of-mouth (Jo & Park, 2024), and satisfaction and self-efficacy with subscription intention (Jo, 2024); perceived personalization and anthropomorphism heightened anxiety about replacement, moderated by skepticism (Jo & Park, 2025). Employees of a multinational firm perceived that Microsoft 365 Copilot reduced administrative burden (Callari & Puppione, 2025). Adoption among Chinese hospital administrators was low and varied across sites, with drafting the main use (Chen et al., 2025); at one academic medical center, email and document writing accounted for 53.9% of 30,503 chat threads (Black et al., 2026). These outcomes describe attitudes and usage, not performance.

RQ3: Mechanisms of human–AI collaboration

Table 4 presents the mechanisms identified. Statistical tests of mechanisms were limited to moderation in one quasi-experiment (Rodriguez Saavedra et al., 2026) and structural-equation surveys (Du et al., 2026; Jo, 2024; Jo & Park, 2024, 2025); the remainder are qualitative findings or authors' arguments.

Table 4. Mechanisms linking generative AI use to outcomes

Mechanism	Studies	Finding
Human oversight as precondition	Aiumtrakul et al. (2026); Rodriguez Saavedra et al. (2026); Tahtali et al. (2026); Choroszewicz (2026); Kalasapudi et al. (2026); Menano de Carvalho et al. (2026)	Review or validation checkpoints were treated as essential; in one ethnography, oversight was assumed and liability fell on users
Trust	Chen et al. (2025); Tahtali et al. (2026); Du et al. (2026)	Distrust of accuracy hindered adoption; trust predicted intention and behavior; no evidence that trust changes output quality
Task–technology fit	Du et al. (2026); Menano de Carvalho et al. (2026)	Fit was the strongest predictor of perceived usefulness in a survey of 1,377 staff; gains were larger for bounded, repetitive, structured work
User experience and skills	Kim (2026); Tahtali et al. (2026); Chen et al. (2025)	Newer staff gained most; concern about eroded critical thinking in novices (not tested); low prompting skill limited use
Organizational conditions	Rodriguez Saavedra et al. (2026); Callari & Puppione (2025); Tahtali et al. (2026)	Digital infrastructure moderated effects and resistance to change reduced them; culture and unclear policy shaped use
System design	Carroll et al. (2026); Klang et al. (2026); Mitroulias & Sioutas (2026)	Retrieval, tools, and response validation improved accuracy compared with direct prompting

Human oversight was usually framed as a precondition rather than studied as a variable. Two studies described how it works in practice: a staged reduction from full review to 10% sampling as confidence grew (Rodriguez Saavedra et al., 2026), and, in contrast, an ethnography of public-sector practitioners using Microsoft 365 Copilot and two in-house assistants that identified five mechanisms by which the nominal duty to verify outputs shifted ethical and operational burden onto individual users, including opacity of the tools, fluent but misleading output, and the vigilance required to catch errors (Choroszewicz, 2026). Trust operated chiefly on adoption: distrust of accuracy was a barrier among hospital administrators (Chen et al., 2025), whereas in a large survey trust predicted intention and behavior (Du et al.,

2026). Oversight was thus described as indispensable but rarely evaluated; its cost to users appeared only in qualitative accounts (Choroszewicz, 2026).

RQ4: Risks, governance, and research gaps

Table 5 counts risks and governance measures as reported. Accuracy-related problems were most common (eight studies): hallucination, miscoding, unreliable citations, unsupported content, execution errors, or distrust of accuracy. Others reported accountability burdens (three), integration, regulatory, or data constraints (three), unspecified ethical concerns (two), data security (one), prompt sensitivity or model variability (two; in one, ±3.8% variation and Fleiss κ = 0.84, Rodriguez Saavedra et al., 2026), possible over-reliance and skill erosion (one), anxiety about job replacement (one), and off-label use: 5.9% of 30,503 chat threads at a medical center concerned clinical decision-making (Black et al., 2026).

Table 5. Risks and governance measures reported in the included studies (number of studies)

Category	n	Studies
Risks		
Accuracy, hallucination, citation or coding error	8	Chen et al. (2025); Wilson et al. (2025); Aiumtrakul et al. (2026); Mitroulias & Sioutas (2026); Kiser et al. (2024); Klang et al. (2026); Ruhwedel et al. (2026); Latario & Fowler (2026)
Accountability, liability, discretion	3	Kim (2026); Choroszewicz (2026); Menano de Carvalho et al. (2026)
Integration, regulatory, or data constraints	3	Zaki et al. (2026); Goyal et al. (2026); Tahtali et al. (2026)
Ethical concerns (unspecified)	2	López-Galisteo (2026); Callari & Puppione (2025)
Data security	1	Wilson et al. (2025)
Prompt sensitivity, model variability	2	Rodriguez Saavedra et al. (2026); Paek et al. (2024)
Over-reliance, skill erosion (not tested)	1	Tahtali et al. (2026)
Job-replacement anxiety	1	Jo & Park (2025)
Off-label use	1	Black et al. (2026)
Governance measures		
Human review or validation checkpoints	5	Aiumtrakul et al. (2026); Rodriguez Saavedra et al. (2026); Nai et al. (2025); Kalasapudi et al. (2026); Menano de Carvalho et al. (2026)
Organizational policy, standards, quality control	4	Kim (2026); Tahtali et al. (2026); Black et al. (2026); Goyal et al. (2026)
Training	3	Chen et al. (2025); Kim (2026); Black et al. (2026)
Rule-based validation or compliance controls	3	Mitroulias & Sioutas (2026); Wang et al. (2026); Zaki et al. (2026)
Secure or private platform	2	Wilson et al. (2025); Black et al. (2026)
Logging and audit sampling	1	Rodriguez Saavedra et al. (2026)

Note. A study can appear in more than one category. Nine studies reported no risk and 13 no governance measure; two further studies only recommended a structured service framework or stronger accountability (López-Galisteo, 2026; Choroszewicz, 2026).

Governance measures were mostly recommendations or features of the studied system; none was evaluated as an intervention. Six research gaps follow from the evidence map: (a) agentic AI, with one explicit study and none in a real office; (b) causal designs, with no randomized trial and only three quasi-experiments of 12 months or less; (c) outcomes beyond speed and reference accuracy, such as work quality, error cost, and oversight effort; (d) skill erosion and accountability, so far only qualitative concerns; (e) geographic and sectoral coverage, with a cluster in health care and little from Africa, Latin America outside Peru, or South and Southeast Asia, and only English-language studies; and (f) governance tested as an intervention.

Synthesis across contexts

Mapping the findings onto the CIMO structure showed four recurring configurations, each with a different evidence base. First, in drafting and communication work by public-sector and university staff, specialized assistants or chat LLMs were associated with shorter task time and lower email volume (Kim, 2026; López-Galisteo, 2026), while adoption was limited by distrust and unclear policy (Chen et al., 2025; Tahtali et al., 2026); the evidence here combined low-rated quasi-experiments with high-quality qualitative studies. Second, in payer-related documentation in health care, chat LLMs used as supplied produced usable drafts but only partly correct codes (58.1–68% agreement), and every study assumed human review; the evidence consisted of low- to moderate-quality evaluations (Aiumtrakul et al., 2026; Kiser et al., 2024; Latario & Fowler, 2026; Ruhwedel et al., 2026). Third, in structured document processing and workflow automation, embedded LLMs reported time reductions of about 60–90% and accuracy of 81–98%, all from unrated system studies (Doropoulos et al., 2026; Soma, 2026; Wang et al., 2026; Zaki et al., 2026). Fourth, in general office work with Copilot or ChatGPT, surveys and qualitative studies of moderate to high quality addressed acceptance, task support, and ethical burden rather than performance (Callari & Puppione, 2025; Choroszewicz, 2026; Du et al., 2026; Jo & Park, 2024). Across configurations, the mechanisms named most often were task structure, human validation, and organizational conditions, whereas the outcomes measured were predominantly speed and agreement with a reference.

Discussion

Feasibility is not yet benefit: the efficiency evidence

Generative AI has been studied across many administrative tasks, but the evidence supports feasibility under supervision more than it supports organizational benefit. The strongest speed claims come from the weakest designs. Reductions of 60–90% were reported by unrated system studies, several by their developers (Soma, 2026; Wang et al., 2026; Zaki et al., 2026), whereas better-controlled evidence outside our corpus points to smaller, conditional effects: a 14% increase in issues resolved per hour in a staggered rollout among 5,179 customer-support agents (Brynjolfsson et al., 2025) and about 40% less time on writing tasks in a randomized online experiment (Noy & Zhang, 2023). The 60–90% figures are therefore better read as upper bounds for structured, single-case tasks without a counterfactual than as expected gains. Adjacent randomized evidence also cautions against assuming that drafting assistance saves time: AI-generated draft replies to patient messages did not shorten reply time and increased reading time by 21.8% (Tai-Seale et al., 2024). By contrast, the favorable burnout and after-hours results for ambient scribes came from an uncontrolled pre–post study, two of whose authors reported vendor affiliations (Olson et al., 2025). Our corpus shows the same pattern: favorable results cluster where control and independence are weakest. One pattern converges across bodies of evidence: gains concentrated among less experienced workers (Brynjolfsson et al., 2025; Kim, 2026), although in our corpus it rests on a single low-rated quasi-experiment.

Accuracy is task-dependent, and plausible errors are the hazard

Free-text drafting was usually judged useful, whereas exact classification was not (ICD-10 concordance of 58.1% and billing-level agreement of 68%). Benchmarks show the same gradient: no model exceeded 50% exact match on medical code querying, and GPT-4 reached 33.9% for ICD-10-CM (Soroush et al., 2024). Tasks and model generations differ, so the figures are not interchangeable, but both indicate that unsupervised exact coding remains unreliable. Retrieval and tool use improved results in the included studies (Carroll et al., 2026; Klang et al.,

2026), yet retrieval-based legal research tools still hallucinated in 17–33% of queries (Magesh et al., 2025); retrieval narrows the gap but does not close it. What kind of error occurs matters more than how often. Scaled-up, instruction-tuned models tend to give plausible but wrong answers instead of declining, and errors on seemingly easy items can escape human supervisors (Zhou et al., 2024). The failures the included studies report most often (wrong codes, billing levels, and citations) are plausible-looking outputs whose correctness can be established only against a reference. Reported accuracy is also a dated snapshot. The behavior of the same hosted model changed substantially within months (Chen et al., 2024), so values for named models such as ChatGPT-5 or Gemini should not be assumed to hold after updates. The wider health-LLM literature shows a similar measurement imbalance: only 5% of 519 evaluations used real patient data, 95.4% assessed accuracy, and administrative tasks such as billing codes appeared in 0.2% (Bedi et al., 2025). Applying reporting guidance for LLM studies (Gallifant et al., 2025), including model version and date, would make workplace evaluations comparable.

Human oversight is assumed rather than evaluated

Included studies treated human review as a precondition and almost never as a variable. Only one varied it, moving from full review to 10% sampling (Rodriguez Saavedra et al., 2026), and the ethnography suggested that the nominal duty to verify moved burden to individual users (Choroszewicz, 2026). Evidence from neighboring fields indicates that oversight cannot be taken for granted. Green (2022) examined 41 policies requiring human oversight of government algorithms and concluded that they rest on an untested assumption that people can perform it, offering false reassurance. Complacency and automation bias are well documented in human-factors research (Parasuraman & Manzey, 2010), yet three survey experiments on public-sector decision making, the last with civil servants, found no automation bias, although selective adherence to stereotype-congruent advice appeared in an earlier experiment but not in the civil-servant sample (Alon-Barkat & Busuioc, 2023). Oversight failure thus appears contingent on task, cues, and context, which are the dimensions the included studies did not vary. Meta-analytic evidence yields a testable prediction. Human–AI combinations performed on average worse than the better of human or AI alone, with losses in decision tasks and gains in content-creation tasks (Vaccaro et al., 2024). Drafting and communication resemble content creation; coding, classification, and approval resemble decision tasks. If so, the favorable pattern for drafting and the weaker pattern for coding in our corpus may reflect task type interacting with oversight rather than technology alone. This is a hypothesis, which no included study tested. Two further mechanisms bear on design. First, human–AI teams outperformed AI alone only when the AI delegated work to humans, because people lacked metaknowledge about their own competence (Fügener et al., 2022); a rule such as 10% sampling after three months is a delegation decision that was never compared with alternatives, such as risk-based sampling. Second, when AI output conflicted with radiologists' judgment without explanation, uncertainty rose, and only the department that practiced "AI interrogation" incorporated the output productively (Lebovitz et al., 2022). This fits the finding that fluent but misleading output makes vigilance costly (Choroszewicz, 2026). Acceptance also shapes what is observed. Public servants prefer AI in supportive or advisory roles over co-decisive roles or full autonomy (Haesevoets et al., 2025), and their attitudes become less favorable as AI's role in policymaking grows (Haesevoets et al., 2026). The absence of fully autonomous execution in our corpus may therefore reflect accountability and acceptance limits as much as technical ones, in a field where algorithmic systems raise accountability questions (Busuioc, 2021) and generative AI raises questions of performance, accountability, and democratic governance (Porumbescu et al., 2025).

Agentic AI: the conceptual literature runs ahead of the evidence

Surveys and frameworks on agentic AI and LLM-based agents have multiplied (Acharya et al., 2025; Kumar et al., 2026; Wang et al., 2024), and information-systems scholars warn that agentic systems that plan and act with little human oversight raise concerns about labor displacement, transparency, governance, and equity (Kumar et al., 2026). Against this, our corpus contains one explicitly agentic study, a benchmark on synthetic conversations (Doropoulos et al., 2026), and none in a real office. Statements that agentic copilots will transform administrative work are thus extrapolations from assistive and pipeline-embedded LLMs. The extrapolation is not neutral. Choosing a level of automation means weighing the consequences of error (Parasuraman et al., 2000); the hazards identified here, plausible errors and the effort of catching them, would bear on actions rather than drafts, an inference that requires empirical testing. Governance evidence lags similarly. Training was recommended in three studies and evaluated in none. In a comparative study of 28 public organizations, those that trained decision-makers, developers, and users were more advanced in AI governance (de Almeida & dos Santos Júnior, 2025), which describes governance maturity rather than outcomes. The tension between separating AI expertise from the organization and integrating it into daily work (Selten & Klievink, 2024) is consistent with the moderating role of digital infrastructure found in one included quasi-experiment (Rodriguez Saavedra et al., 2026), but no included study tested an organizational design directly.

Learning and skill: gains and risks may accumulate in the same workers

Evidence on skills comes mostly from other settings, and its transfer to office work is untested. In a field experiment with nearly a thousand students, unrestricted GPT-4 access improved practice performance but lowered later unaided performance by 17%, whereas a tutor with safeguards largely mitigated the harm (Bastani et al., 2025). A review of deskilling in medicine calls for longitudinal research (Natali et al., 2025), and a survey of 666 participants linked frequent AI use to lower critical thinking, mediated by cognitive offloading (Gerlich, 2025), although a cross-sectional correlation cannot separate an effect of AI from self-selection. In our corpus, newer staff benefited most (Kim, 2026) while concern about eroded critical thinking in novices was voiced but not tested (Tahtali et al., 2026). The implication is a conditional hypothesis: if the workers who gain most are those still acquiring expertise, short-term productivity measures may overstate net benefit unless unaided performance is also tracked.

Implications for research and practice

For research, the evidence calls for designs that can attribute effects, such as randomized or stepped-wedge rollouts within administrative units, measuring work quality, rework, oversight time, and downstream errors alongside speed; for treating the oversight protocol (full review, sampling, or risk-based review) as the intervention; for follow-up long enough to observe skill change; and for reporting model version, date, and prompts (Gallifant et al., 2025). For practice, the evidence supports piloting first in bounded content-creation tasks with measured baselines, auditing exact-match tasks such as coding by sampling, and not equating a nominal duty to verify with accountability. These are inferences from the evidence, not tested recommendations.

Limitations of the review

Several features of the review process limit confidence in its findings. Only Scopus was searched, without gray literature, citation chasing, or validation of the search string against

benchmark articles, and the search was restricted to English-language journal articles; relevant studies indexed elsewhere or published in other forms may therefore be missing. Administrative work has no settled vocabulary, and because the string required explicit administrative-work terms, studies describing the same tasks in other words may not have been retrieved. Three reports could not be accessed. The protocol was not registered, and its departures (a second screening pass, no pilot round, and a different synthesis framework) could have shaped which studies were included. Screening agreement was measured against a lexical benchmark rather than a second independent reviewer, and the second passes for extraction and appraisal were not independent of the first, so no decision was verified independently and the agreement statistics should not be read as inter-rater reliability. The count-based MMAT classification was used for interpretation only and is not prescribed by the tool. Finally, designs and outcomes were too heterogeneous to pool, so the synthesis is narrative, and accuracy results for named models describe versions that may since have changed. The conclusions are accordingly provisional.

CONCLUSIONS

This review synthesized 29 empirical studies of generative AI in administrative work and found that the evidence shows feasibility under human supervision but not yet organizational benefit. The studies covered drafting, workflow automation, and payer documentation, mostly with general-purpose chat models or models embedded in pipelines. None used a randomized design, one evaluated an explicitly agentic system, and none reported fully autonomous execution. The largest time savings came from unrated, often developer-reported system studies, whereas accuracy was mixed and task-dependent, and weakest where exact outputs such as medical codes were required. Accuracy-related problems were the most frequently reported risk, and governance measures were recommended or built into systems but not evaluated as interventions. Three points stand out. First, human oversight is the central unexamined mechanism: it was assumed as the safeguard but rarely varied or measured, so its effectiveness and cost remain unknown. Second, claims about agentic copilots in office work currently exceed the evidence and are best treated as hypotheses to be tested. Third, outcomes beyond speed and agreement with a reference, including work quality, downstream errors, oversight effort, and skill development, are essentially unmeasured. For research, the priority is attributable designs, such as randomized or stepped-wedge rollouts, that treat the oversight protocol as the intervention and report model version and date. For practice, organizations should pilot in bounded content-creation tasks with measured baselines and audit exact-match tasks by sampling, rather than equate a nominal duty to verify with accountability. These conclusions rest on a young, fast-changing, and mostly single-site evidence base and should be revisited as controlled studies accumulate.

Novelty of this study lies in providing a cross-sector systematic evidence synthesis that critically distinguishes the technical feasibility of generative AI copilots and agentic systems from their demonstrated organizational effectiveness in administrative work. Unlike previous studies that predominantly emphasize automation potential, task efficiency, or isolated AI implementation outcomes, this review integrates evidence on productivity, work quality, accuracy, error occurrence, human–AI collaboration mechanisms, and governance requirements through a context–intervention–mechanism–outcome perspective. A particularly distinctive contribution is the identification of an evidence asymmetry between reported efficiency gains and the robustness of their empirical validation, revealing that substantial task-time reductions (60–90%) are largely derived from system-development studies and internal testing rather than rigorous evaluations of organizational performance.

Furthermore, by identifying the limited representation of explicitly agentic systems, the absence of randomized studies, and the mixed accuracy of AI-assisted administrative tasks, this review establishes an evidence-based distinction between supervised task-level feasibility and validated organizational benefits. Consequently, the study advances the understanding of AI-enabled administrative transformation by proposing a more critical evaluative perspective in which productivity improvements must be assessed alongside output quality, downstream error risks, human oversight, and governance capacity. This contribution provides a foundation for developing more rigorous research designs and responsible implementation strategies before generative AI copilots and agentic systems can be considered reliable instruments of organizational administrative transformation.

BIBLIOGRAPHY

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey. IEEE Access, 13, 18912–18936. https://doi.org/10.1109/ACCESS.2025.3532853

Aiumtrakul, N., Thongprayoon, C., Kookanok, C., Poochanasri, M., Phichedwanichskul, K., & Cheungpasitporn, W. (2026). Quality assessment of large language model–generated prior authorization letters in nephrology. Frontiers in Digital Health, 8, Article 1767648. https://doi.org/10.3389/fdgth.2026.1767648

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: "Automation bias" and "selective adherence" to algorithmic advice. Journal of Public Administration Research and Theory, 33(1), 153–169. https://doi.org/10.1093/jopart/muac007

Artsi, Y., Sorin, V., Glicksberg, B. S., Korfiatis, P., Nadkarni, G. N., & Klang, E. (2025). Large language models in real-world clinical workflows: A systematic review of applications and implementation. Frontiers in Digital Health, 7, Article 1659134. https://doi.org/10.3389/fdgth.2025.1659134

Baas, J., Schotten, M., Plume, A. M., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies, 1(1), 377–386. https://doi.org/10.1162/qss_a_00019

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J. A., Wornow, M., Swaminathan, A., Lehmann, L. S., Hong, H. J., Kashyap, M., Chaurasia, A. R., Shah, N. R., Singh, K., Tazbaz, T., Milstein, A., Pfeffer, M. A., & Shah, N. H. (2025). Testing and evaluation of health care applications of large language models: A systematic review. JAMA, 333(4), 319–328. https://doi.org/10.1001/jama.2024.21700

Black, K. C., Haberkorn, W. J., Ma, S. P., Kiani, H., Bhasin, A., Chen, J. H., Shah, N. H., & Morse, K. (2026). Uses of generative AI by non-clinician staff at an academic medical center. npj Health Systems, 3(1), Article 13. https://doi.org/10.1038/s44401-025-00063-y

Bramer, W. M., Giustini, D., de Jonge, G. B., Holland, L., & Bekhuis, T. (2016). De-duplication of database search results for systematic reviews in EndNote. Journal of the Medical Library Association, 104(3), 240–243. https://doi.org/10.3163/1536-5050.104.3.014

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

Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. Public Administration Review, 81(5), 825–836. https://doi.org/10.1111/puar.13293

Callari, T. C., & Puppione, L. (2025). Can generative artificial intelligence productivity tools support workplace learning? A qualitative study on employee perceptions in a multinational corporation. Journal of Workplace Learning, 37(3), 266–283. https://doi.org/10.1108/JWL-11-2024-0258

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ, 368, Article l6890. https://doi.org/10.1136/bmj.l6890

Carroll, A., Brown, D., Silvio, D., Maszatics, J., & Hanauer, D. A. (2026). Evaluating large language models for automated REDCap support ticket triage and response. JAMIA Open, 9(3), Article ooag073. https://doi.org/10.1093/jamiaopen/ooag073

Chen, J., Liu, Y., Liu, P., Zhao, Y., Zuo, Y., & Duan, H. (2025). Adoption of large language model AI tools in everyday tasks: Multisite cross-sectional qualitative study of Chinese hospital administrators. Journal of Medical Internet Research, 27, Article e70789. https://doi.org/10.2196/70789

Chen, L., Zaharia, M., & Zou, J. (2024). How is ChatGPT's behavior changing over time? Harvard Data Science Review, 6(2). https://doi.org/10.1162/99608f92.5317da47

Choroszewicz, M. (2026). Positioning public sector practitioners as 'moral crumple zones': Mechanisms in the early use of generative AI work support tools. Government Information Quarterly, 43(2), Article 102128. https://doi.org/10.1016/j.giq.2026.102128

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

de Almeida, P. G. R., & dos Santos Júnior, C. D. (2025). Artificial intelligence governance: Understanding how public organizations implement it. Government Information Quarterly, 42(1), Article 102003. https://doi.org/10.1016/j.giq.2024.102003

Denyer, D., Tranfield, D., & van Aken, J. E. (2008). Developing design propositions through research synthesis. Organization Studies, 29, 393–413. https://doi.org/10.1177/0170840607088020

Doropoulos, S., Vologiannidis, S., & Magnisalis, I. (2026). DevNous: An LLM-based multi-agent system for grounding IT project management in unstructured conversation. Information and Software Technology, 195, Article 108078. https://doi.org/10.1016/j.infsof.2026.108078

Du, X., Ling, C., Yang, L., & Wei, L. (2026). Factors shaping the adoption of large language models among hospital administrative staff: A cross-sectional survey study. Digital Health, 12. https://doi.org/10.1177/20552076261450815

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. Science, 384(6702), 1306–1308. https://doi.org/10.1126/science.adj0998

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. Information Systems Research, 33(2), 678–696. https://doi.org/10.1287/isre.2021.1079

Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S., Cacciamani, G., Demner-Fushman, D., Dligach, D., Daneshjou, R., Fernandes, C., Hansen, L., Landman, A., Lehmann, L., McCoy, L., Miller, T., Moreno, A., Munch, N., Restrepo, D., Savova, G., . . . Bitterman, D.

(2025). The TRIPOD-LLM reporting guideline for studies using large language models. Nature Medicine, 31(1), 60–69. https://doi.org/10.1038/s41591-024-03425-5

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), Article 6. https://doi.org/10.3390/soc15010006

Goyal, R., Deshmukh, S. G., & Bolia, N. (2026). Navigating barriers to GenAI adoption in public administration: A systematic evaluation and policy roadmap. Socio-Economic Planning Sciences, 105, Article 102461. https://doi.org/10.1016/j.seps.2026.102461

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review, 45, Article 105681. https://doi.org/10.1016/j.clsr.2022.105681

Haesevoets, T., Verschuere, B., & Roets, A. (2025). AI adoption in public administration: Perspectives of public sector managers and public sector non-managerial employees. Government Information Quarterly, 42(2), Article 102029. https://doi.org/10.1016/j.giq.2025.102029

Haesevoets, T., Verschuere, B., Van Hiel, A., & Roets, A. (2026). Public servants' attitudes toward the use of artificial intelligence in policymaking: An experimental study. Government Information Quarterly, 43(2), Article 102134. https://doi.org/10.1016/j.giq.2026.102134

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Education for Information, 34(4), 285–291. https://doi.org/10.3233/EFI-180221

Humlum, A., & Vestergaard, E. (2025). The unequal adoption of ChatGPT exacerbates existing inequalities among workers. Proceedings of the National Academy of Sciences, 122(1), Article e2414972121. https://doi.org/10.1073/pnas.2414972121

Jo, H. (2024). Subscription intentions for ChatGPT Plus: A look at user satisfaction and self-efficacy. Marketing Intelligence & Planning, 42(6), 1052–1073. https://doi.org/10.1108/MIP-08-2023-0411

Jo, H., & Park, D.-H. (2024). Effects of ChatGPT's AI capabilities and human-like traits on spreading information in work environments. Scientific Reports, 14(1), Article 7806. https://doi.org/10.1038/s41598-024-57977-0

Jo, H., & Park, D.-H. (2025). The fear of being replaced by generative AI: An examination of influential factors among office workers. Technological Forecasting and Social Change, 220, Article 124326. https://doi.org/10.1016/j.techfore.2025.124326

Jonathan, G. M. (2026). Public-sector digital transformation in the age of generative AI. Discover Artificial Intelligence, 6(1), Article 968. https://doi.org/10.1007/s44163-026-02077-3

Kalasapudi, V. S., Seelam, A., Ganupa, S., & Tofferi, C. (2026). The overlooked frontier of AI in construction: Conversational, document-native automation for administrative workflows. Frontiers in Built Environment, 11, Article 1713342. https://doi.org/10.3389/fbuil.2025.1713342

Kim, E. (2026). Generative AI in public administration: A quasi-experimental analysis of bureaucratic productivity. Government Information Quarterly, 43(1), Article 102108. https://doi.org/10.1016/j.giq.2026.102108

Kiser, K. J., Waters, M., Reckford, J., Lundeberg, C., & Abraham, C. D. (2024). Large language models to help appeal denied radiotherapy services. JCO Clinical Cancer Informatics, 8, Article e2400129. https://doi.org/10.1200/CCI.24.00129

Klang, E., Sorin, V., Korfiatis, P., Sawant, A. S., Freeman, R., Charney, A. W., Nadkarni, G. N., & Glicksberg, B. S. (2026). Large language models are poor clinical administrators: An evaluation of structured queries in real-world electronic health records. PLOS Digital Health, 5, Article e0001326. https://doi.org/10.1371/journal.pdig.0001326

Kumar, N., Wei, X., & Zhang, H. (2026). Agentic artificial intelligence as a new frontier in information systems: Promise, peril, and research opportunities. Information & Management, 63(3), Article 104317. https://doi.org/10.1016/j.im.2026.104317

Latario, L. D., & Fowler, J. R. (2026). Chatbot demonstrates moderate interrater reliability in billing for hand surgery clinic encounters. Hand, 21(1), 68–72. https://doi.org/10.1177/15589447241295328

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. Organization Science, 33(1), 126–148. https://doi.org/10.1287/orsc.2021.1549

López-Galisteo, A. J. (2026). Optimizing university administrative services with generative AI: Evidence from email inquiry reduction and assistant performance. Information, 17(6), Article 587. https://doi.org/10.3390/info17060587

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. Journal of Empirical Legal Studies, 22(2), 216–242. https://doi.org/10.1111/jels.12413

Menano de Carvalho, L., Peças, P., & Jorge, D. (2026). AI-enabled process improvement in information-intensive administrative work: Real-case applications of LLMs in a Lean Six Sigma context. Sustainability, 18(10), Article 4787. https://doi.org/10.3390/su18104787

Mitroulias, D., & Sioutas, S. (2026). Empirical evaluation of a domain-specific RAG-based conversational assistant for payroll management: A mixed-methods study of accuracy, satisfaction, and usability. Big Data and Cognitive Computing, 10(6), Article 193. https://doi.org/10.3390/bdcc10060193

Moynihan, D., Herd, P., & Harvey, H. (2015). Administrative burden: Learning, psychological, and compliance costs in citizen-state interactions. Journal of Public Administration Research and Theory, 25(1), 43–69. https://doi.org/10.1093/jopart/muu009

Nai, R., Sulis, E., Audrito, D., Trifiletti, V. M. S., Meo, R., & Genga, L. (2025). Leveraging process mining and event log enrichment in European public procurement analysis: A case study. Computer Law & Security Review, 57, Article 106144. https://doi.org/10.1016/j.clsr.2025.106144

Natali, C., Marconi, L., Dias Duran, L. D., & Cabitza, F. (2025). AI-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond. Artificial Intelligence Review, 58(11), Article 356. https://doi.org/10.1007/s10462-025-11352-1

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

Olson, K. D., Meeker, D., Troup, M., Barker, T. D., Nguyen, V. H., Manders, J. B., Stults, C. D., Jones, V. G., Shah, S. D., Shah, T., & Schwamm, L. H. (2025). Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Network Open, 8(10), Article e2534976. https://doi.org/10.1001/jamanetworkopen.2025.34976

Paek, E. S., Fan, T., Finch, J. D., & Choi, J. D. (2024). Enhancing task-oriented dialogue systems through synchronous multi-party interaction and multi-group virtual simulation. Information, 15(9), Article 580. https://doi.org/10.3390/info15090580

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10.1109/3468.844354

Porumbescu, G., Young, M., & Yang, K. (2025). Introduction to special issue on generative AI and public administration: Performance, accountability, and democratic governance. Public Performance & Management Review, 48(6), 1263–1267. https://doi.org/10.1080/15309576.2025.2570315

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. Academy of Management Review, 46(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: An extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Systematic Reviews, 10, Article 39. https://doi.org/10.1186/s13643-020-01542-z

Rodriguez Saavedra, M. O., Lopez, J. M. B., Nina, W. Q., Morales Gonzales, A. V., Galindo, I. C., Ascuña, L. M. C., Colana, A. S. S., Almanza Cabe, R. B., Tito, P. G. L., & Teran, S. V. L. (2026). Machine learning and generative AI in administrative processes in Peru: Administrative efficiency in the national public sector. Informatics, 13(3), Article 44. https://doi.org/10.3390/informatics13030044

Ruhwedel, T., Rogasch, J. M. M., Dahlke, P. M., Shnayien, S., Furth, C., Wetz, C., Amthauer, H., Schatka, I., & Beetz, N. L. (2026). Less time coding, more time caring: Performance evaluation of ChatGPT-5 for ICD-10 coding of radiology reports. International Journal of Medical Informatics, 210, Article 106296. https://doi.org/10.1016/j.ijmedinf.2026.106296

Selten, F., & Klievink, B. (2024). Organizing public sector AI adoption: Navigating between separation and integration. Government Information Quarterly, 41(1), Article 101885. https://doi.org/10.1016/j.giq.2023.101885

Soma, R. (2026). End-to-end architecture for LLM-based business process automation: A low-code integration framework with performance benchmarking. IEEE Access, 14, 105921–105940. https://doi.org/10.1109/ACCESS.2026.3696852

Soroush, A., Glicksberg, B. S., Zimlichman, E., Barash, Y., Freeman, R., Charney, A. W., Nadkarni, G. N., & Klang, E. (2024). Large language models are poor medical coders—Benchmarking of medical code querying. NEJM AI, 1(5). https://doi.org/10.1056/AIdbp2300040

Syed, R., Suriadi, S., Adams, M., Bandara, W., Leemans, S. J. J., Ouyang, C., ter Hofstede, A. H. M., van de Weerd, I., Wynn, M. T., & Reijers, H. A. (2020). Robotic process automation: Contemporary themes and challenges. Computers in Industry, 115, Article 103162. https://doi.org/10.1016/j.compind.2019.103162

Tahtali, M. A., Le Blanc, P., Dirne, C., & Snijders, C. (2026). Why public procurement professionals accept (or resist) generative AI: A qualitative study. Journal of Public Procurement, 26(4), 727–761. https://doi.org/10.1108/JOPP-09-2025-0110

Tai-Seale, M., Baxter, S. L., Vaida, F., Walker, A., Sitapati, A. M., Osborne, C., Diaz, J., Desai, N., Webb, S., Polston, G., Helsten, T., Gross, E., Thackaberry, J., Mandvi, A., Lillie, D., Li, S., Gin, G., Achar, S., Hofflich, H., . . . Longhurst, C. A. (2024). AI-generated draft replies integrated into health records and physicians' electronic communication. JAMA Network Open, 7(4), Article e246565. https://doi.org/10.1001/jamanetworkopen.2024.6565

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8, Article 45. https://doi.org/10.1186/1471-2288-8-45

Tseng, P., Kaplan, R. S., Richman, B. D., Shah, M. A., & Schulman, K. A. (2018). Administrative costs associated with physician billing and insurance-related activities at an academic health care system. JAMA, 319(7), 691–697. https://doi.org/10.1001/jama.2017.19148

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), Article 186345. https://doi.org/10.1007/s11704-024-40231-1

Wang, M. H., Lee, S. M. Y., Hou, G., Wang, Y., Alves, J. C., Xie, R., He, Y., Liu, J., Fang, X., Yang, Y., Cai, X., Zheng, S., Yu, Z., Lin, E. Z., Cheang, C., Ian, K. K., & Qin, S. (2026). Explainable multimodal AI and neuro-symbolic clinical decision support system for chronic eye disease management: A digital health implementation study. Frontiers in Digital Health, 8, Article 1836890. https://doi.org/10.3389/fdgth.2026.1836890

Wilson, L. A., Konsynski, B., & Yisrael, T. (2025). Advancing research administration with AI: A case study from Emory University. Journal of Research Administration, 56(2), 52–74.

Zaki, N., Alneyadi, A., Alamoodi, M., Alshaali, S., Alkaabi, F., & Alhashmi, A. (2026). An AI-enabled framework for reducing administrative bureaucracy in higher education while maintaining governance and compliance. Discover Artificial Intelligence, 6(1), Article 879. https://doi.org/10.1007/s44163-026-01565-w

Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. Nature, 634(8032), 61–68. https://doi.org/10.1038/s41586-024-07930-y