

Open
AccessCheck for
updates

REVIEW ARTICLE

Section: *Cultural Heritage & Conservative Studies***ChatGPT and the digital humanities: A cross-domain review of human-AI collaboration in education, work, and culture, and the question of trust for advancing SDG 4**Mashaël Nasser Ayed Al-Dosari¹, Ali Lamouchi^{2*}, Aljwhrh Nasr A. Aldwsri³, Mohamed Sayed Abdellatif⁴, Reem Nasser A. Aldosari⁵ & Hassan Mohamed Ibrahim⁶¹Department of Educational Sciences, College of Education in Al-Kharj, Prince Sattam Bin Abdulaziz University, Saudi Arabia²Department of English Language and Literature, College of Science and Humanities, Prince Sattam bin Abdulaziz University, Saudi Arabia³Department of Administrative Sciences, Applied College, Princess Nourah bint Abdulrahman University (PNU), Saudi Arabia⁴Department of Psychology, College of Education, Prince Sattam Bin Abdulaziz University, Saudi Arabia⁵Department of Sociology and Social Work, College of Social Sciences, Imam Mohammad Ibn Saud Islamic University, Saudi Arabia⁶Department of Libraries, Information, and Educational Technology Department, Faculty of Education, Al-Azhar University, Egypt*Correspondence: a.lamouchi@psau.edu.sa**ABSTRACT**

The rapid diffusion of ChatGPT into classrooms, workplaces, and caregiving settings has become a defining episode in the humanities' longstanding engagement with how new communication technologies reshape reading, writing, and human judgment — yet the scholarship on this episode remains fragmented, with most studies examining single settings in isolation rather than asking what these episodes share as instances of human-machine collaboration. This gap matters for the digital humanities in particular, where questions of language, literacy, authorship, and trust in mediated communication intersect with the broader cultural and ethical stakes of generative AI's integration into everyday human activity. This review addresses that gap by synthesizing evidence on ChatGPT usage, benefits, and trust concerns across workplace, educational, software development, marketing, and healthcare contexts, treating each as a site where humans and a language-generating machine negotiate authorship, judgment, and accountability. The purpose was to identify the task-level and domain-level conditions under which ChatGPT yields reliable gains, to characterize recurring reliability limitations, and to determine whether the structure of a task — rather than the sector in which it occurs — better explains when human trust in AI-generated language is warranted. A narrative multi-domain literature review was conducted, drawing on empirical studies published between 2023 and 2025 identified through structured database searches and reference-list screening, with findings extracted, categorized by domain, and synthesized thematically. Results showed that ChatGPT produced the most consistent gains on bounded, language-heavy tasks such as drafting, summarizing, and routine code generation, while

Research Journal in Advanced Humanities

Volume 7, Issue 4, 2026

ISSN: 2708-5945 (Print)

ISSN: 2708-5953 (Online)

ARTICLE HISTORY

Submitted: 10 July 2026

Accepted: 23 August 2026

Published: 21 September 2026

HOW TO CITE

Al-Dosari, M. N. A., Lamouchi, A., Aldwsri, A. N. A., Abdellatif, M. S., Aldosari, R. N. A., & Ibrahim, H. M. (2026). ChatGPT and the digital humanities: A cross-domain review of human-AI collaboration in education, work, and culture, and the question of trust for advancing SDG 4. *Research Journal in Advanced Humanities*, 7(4). <https://doi.org/10.58256/8h63pn25>

performance declined on tasks requiring specialized judgment, cultural nuance, or verified accuracy, with reliability concerns — including fabricated citations and inconsistent outputs — recurring across otherwise unrelated sectors. These findings indicate that task structure, rather than industry sector, is the primary determinant of ChatGPT's trustworthiness, supporting adoption models that pair AI-generated language with calibrated human oversight in education, professional practice, and care — an approach consistent with the humanities' broader concern for preserving human judgment, authorship, and critical literacy amid the growing presence of generative language technologies in everyday life.

KEYWORDS: ChatGPT, digital humanities, generative AI, human-AI collaboration, language and literacy, task structure, trust and reliability, cross-domain review, critical AI literacy, SDG 4, quality education, sustainable education

Introduction

Since its public release, ChatGPT has rapidly become one of the most widely adopted conversational AI tools, reshaping how individuals approach writing, problem-solving, and information-seeking across professional and academic settings (Dwivedi et al., 2023; Libera et al., 2025; Sohail et al., 2023). For scholars in the digital humanities, this diffusion is significant not merely as a technological event but as a cultural one: a machine capable of producing fluent, context-aware text is now embedded in the everyday practices of reading, writing, and meaning-making that the humanities have traditionally studied and taught (Jo & Park, 2024; Kasneci et al., 2023). Its adoption has spread not only among technically oriented users but also within traditionally non-technical fields such as healthcare, marketing, and education, unsettling long-held assumptions about authorship, originality, and the boundary between human and machine-generated language. Scholars increasingly describe ChatGPT usage as a family of workflows through which users offload reading, writing, coding, and administrative tasks to a generative model (Azaria et al., 2024; Thant et al., 2024). This diffusion has prompted growing research into both its benefits and risks, and, correspondingly, growing interest within the humanities in how generative language technologies are transforming literacy, communication, and human judgment across domains.

A substantial share of this research has examined workplace productivity, where findings are notably consistent. Controlled and field-based studies report that ChatGPT assistance reduces task completion time while improving output quality across knowledge work, consulting, and software engineering (Kim & Lee, 2025; Kuhail et al., 2024; Matzko & Konur, 2024; Sanatizadeh et al., 2025). These gains are not evenly distributed, with lower-performing employees often benefiting disproportionately relative to more experienced staff (Jo & Park, 2024; Libera et al., 2025). Labor-market analyses further suggest that AI-related skills have become increasingly valuable within knowledge-worker roles, reflecting rising organizational demand for ChatGPT-related competencies (Yu & Yu, 2026; Zehir & Yilmaz, 2025). Together, these studies establish productivity as a robust benefit of adoption.

In education, ChatGPT has been widely adopted for brainstorming, summarizing, explaining concepts, and supporting academic writing (Alsuwaylim et al., 2026; Crompton & Burke, 2024; Nemt-Allah et al., 2024; Ravšelj et al., 2025). Writing interventions incorporating ChatGPT-assisted revision have produced measurable gains in essay quality and outperformed conventional instruction in English-as-a-foreign-language contexts (Pei et al., 2025; Tsai et al., 2024). Additional studies report improved comprehension, engagement, and creativity among students using ChatGPT for guided learning tasks (Chen & Xu, 2024; Garcia et al., 2025; Imran et al., 2024). However, these educational benefits are accompanied by persistent concerns regarding academic integrity, plagiarism, and diminished independent problem-solving, prompting calls for AI literacy and redesigned assessment practices (Bego, 2023; Cotton et al., 2024; Rasul et al., 2023).

Software development represents another domain where ChatGPT's capabilities have been extensively benchmarked. Code-generation studies show that ChatGPT can outperform comparable coding assistants such as Copilot and CodeWhisperer, though performance remains sensitive to prompt context and task complexity (Kalyan, 2024; Liu et al., 2023). ChatGPT has also demonstrated utility in code refactoring, documentation, test-case generation, and bug repair, with accuracy improving substantially when richer contextual information is supplied (DePalma et al., 2024; Koubaa et al., 2023; Takerngsaksiri et al., 2025). Nevertheless, several studies report degraded performance on long prompts, unfamiliar problems, and large or complex codebases,

underscoring the continued necessity of human review in professional software engineering workflows (Da Silva et al., 2025; Jain et al., 2025; Kim & Ming, 2025).

Within marketing and customer-facing industries, ChatGPT has been adopted primarily for content marketing, search engine optimization, and customer service automation (Singh et al., 2024; Tafesse & Wien, 2024). Empirical evidence indicates that AI-assisted, human-refined workflows can outperform human-only production in both search-ranking performance and production cost (Brüns & Meißner, 2024). At the same time, concerns have emerged regarding perceived brand authenticity, privacy, and the contextual limitations of AI-generated communication, suggesting that quality control remains essential even where efficiency gains are documented (Arslanoğlu & Karakose, 2025; Iyengar et al., 2023). These mixed findings highlight that marketing applications, while promising, require closer human oversight than more bounded technical tasks. Healthcare applications of ChatGPT reveal a similar tension between measurable efficiency gains and serious reliability concerns. Clinical documentation studies report substantial reductions in administrative time for tasks such as discharge-summary drafting (Han et al., 2024; Tangsrivimol et al., 2025). However, studies examining citation accuracy and clinical decision support have documented a large proportion of fabricated references and non-trivial incorrect-recommendation rates in specialties such as oncology (Li et al., 2025; Shiraishi et al., 2024). Additional evidence points to inconsistent performance across medical examinations and patient-facing question-answering tasks, with accuracy varying considerably by specialty and language (Ishida et al., 2024; Pavone et al., 2024; Riedel et al., 2023). These findings reinforce healthcare as a domain requiring cautious, supervised ChatGPT integration.

Despite the breadth of evidence documenting ChatGPT's benefits, most existing studies remain domain-specific, examining workplace, educational, coding, marketing, or clinical outcomes in isolation rather than comparing them within a unified analytical framework (Bilika et al., 2024; Navas et al., 2024; Suárez et al., 2024). This fragmentation is itself a humanities problem: without a cross-domain, interpretive account of how people read, evaluate, and act on machine-generated language, the field lacks a shared vocabulary for discussing when such language deserves trust. Relatively few reviews have systematically synthesized cross-domain reliability concerns, particularly the recurring pattern whereby fluent, confident-sounding outputs mask hallucinations, fabricated citations, and inconsistent accuracy (Fatahi et al., 2024; Sallam, 2023). Moreover, limited attention has been paid to whether task structure, rather than industry sector, is the underlying determinant of ChatGPT's performance and trustworthiness across otherwise unrelated fields, leaving an important gap in the comparative literature and in the digital humanities' understanding of how generative AI is reshaping human textual practice. This review addresses that gap by synthesizing evidence on ChatGPT usage, benefits, and trust concerns across workplace, educational, software development, marketing, and healthcare contexts, reading each as a distinct site of human-AI collaboration in language and communication. Specifically, it aims to identify the task-level and domain-level conditions under which ChatGPT produces reliable productivity and quality gains, to characterize the recurring reliability limitations reported across sectors, and to determine whether task structure offers a more consistent explanatory framework than industry classification. By consolidating findings from workplace to classroom and beyond, this review seeks to clarify where ChatGPT-generated language can be trusted for autonomous or semi-autonomous use and where human verification and critical literacy remain indispensable, thereby informing more responsible and culturally sensitive adoption practices and contributing a cross-domain perspective to digital humanities scholarship on human-AI collaboration.

Methods

Review Design

This study employed a narrative multi-domain literature review design, consistent with interpretive approaches common in the digital humanities, to synthesize empirical evidence on ChatGPT usage across workplace, educational, software development, marketing, and healthcare contexts. A narrative synthesis approach was selected rather than a formal systematic review because the objective was to identify cross-domain patterns in human-AI collaboration, productivity outcomes, and reliability concerns rather than a single pooled effect size. This design allowed heterogeneous methodologies, including preregistered field experiments, benchmark evaluations, quasi-experiments, and systematic reviews, to be interpreted within one analytical framework. The review prioritized recent, empirically grounded studies published between 2023 and 2025 to ensure findings reflected the capabilities of contemporary ChatGPT versions rather than earlier iterations of the model.

Search Strategy and Data Sources

Relevant studies were identified through structured searches of academic databases and indexing platforms, including Scopus-indexed journals, IEEE Xplore, and publisher repositories such as ScienceDirect and SpringerLink. Search terms combined the keyword “ChatGPT” with domain-specific terms including “productivity,” “education,” “software development,” “marketing,” “healthcare,” “trust,” and “hallucination.” Boolean operators were used to refine results and capture studies addressing both performance outcomes and reliability limitations. Reference lists of retrieved articles were additionally screened to identify further relevant sources not captured through database searching. This combined search strategy was designed to maximize coverage across the five target domains while minimizing the omission of methodologically rigorous but less indexed studies.

Inclusion and Exclusion Criteria

Studies were included if they empirically examined ChatGPT usage, performance, or user outcomes within at least one of the five target domains and reported quantifiable or clearly described qualitative results. Eligible study designs included controlled experiments, field studies, benchmark evaluations, quasi-experimental comparisons, cross-sectional surveys, and systematic or scoping reviews. Studies were excluded if they were purely theoretical or opinion-based without empirical data, focused exclusively on general large language model architecture without reference to ChatGPT specifically, or examined AI tools unrelated to conversational assistants. Non-English publications and studies lacking sufficient methodological detail to assess outcome measures were also excluded, ensuring that the reviewed evidence base remained both relevant and interpretable.

Study Selection Process

Retrieved records were screened in two stages. Titles and abstracts were first reviewed to exclude clearly irrelevant studies, followed by full-text assessment of remaining articles against the inclusion and exclusion criteria. During full-text screening, particular attention was paid to whether studies reported concrete outcome measures, such as completion time, accuracy rates, quality scores, or error frequencies, rather than general impressions of usefulness. Studies meeting all criteria were retained for data extraction. Where multiple studies examined overlapping domains, all were retained if they contributed distinct outcome measures or methodological perspectives, allowing the review to capture both convergence and variation in reported findings across similar contexts.

Data Extraction and Categorization

For each included study, key information was extracted, including the research domain, study design, sample or task characteristics, outcome measures, and principal quantitative or qualitative findings. Extracted data were organized into domain-specific categories corresponding to workplace productivity, education, software development, marketing, and healthcare, with an additional cross-cutting category capturing trust, hallucination, and misuse-related findings. This categorization enabled systematic comparison of outcome patterns both within and across domains. Extracted results were subsequently tabulated to facilitate direct comparison of effect magnitudes and outcome types, as presented in the accompanying results tables, while qualitative findings were synthesized narratively to capture contextual nuance not reducible to numerical outcomes.

Data Synthesis Approach

Data synthesis followed a thematic, cross-domain comparative approach rather than statistical meta-analysis, given the heterogeneity of outcome measures, study designs, and reporting formats across the included literature. Findings were grouped according to recurring patterns, particularly the distinction between bounded, language-heavy tasks and complex, high-context tasks requiring specialized judgment. Special attention was given to reliability-related findings, including hallucination rates, fabricated citations, and inconsistent accuracy, given their relevance to the review’s focus on trust. This synthesis strategy allowed the review to identify whether task structure, rather than industry sector, functioned as a consistent determinant of ChatGPT’s performance and trustworthiness across the examined domains.

Results

Across the reviewed literature, ChatGPT was most consistently associated with measurable productivity gains in professional and knowledge-work settings. Controlled and field-based comparisons of task completion with and without ChatGPT assistance reported reductions in completion time alongside improvements in output quality, with comparable patterns observed in consulting, software engineering, and broader labor-market indicators (Kim & Lee, 2025; Kuhail et al., 2024; Matzko & Konur, 2024; Sanatizadeh et al., 2025; Yu & Yu, 2026). Notably, these gains were not distributed evenly: lower-performing workers tended to benefit disproportionately, narrowing performance gaps within teams even as experts retained an advantage on harder tasks (Ng et al., 2025). Table 1 summarizes the principal quantitative productivity outcomes reported across these six sources.

Table 1. *Workplace and Knowledge-Work Productivity Outcomes*

Study	Sample / Setting	Outcome Measure	Reported Result	Boundary Condition
Kim & Lee (2025)	453 college-educated professionals (preregistered experiment)	Completion time and output quality	40% reduction in completion time; 18% increase in quality	Gains strongest on bounded, well-defined tasks
Ng et al. (2025)	International cross-sectional workforce survey	Performance inequality between workers	Reduced inequality; lower-performing workers benefited disproportionately	High performers showed smaller relative gains
Sanatizadeh et al. (2025)	Boston Consulting Group consultants	Task speed and work quality	25.1% faster completion; 40% improvement in quality	Degrades on tasks outside the AI capability frontier
Kuhail et al. (2024)	Professional software developers (ChatGPT-3)	Coding time and output quality	0.8 SD reduction in coding time; 0.4 SD gain in quality	Not reported
Matzko & Konur (2024)	Tracked software-engineering team adoption	Productivity index over time	212% increase initially; 149% sustained after adoption	Effect diminishes but remains elevated post-adoption
Yu & Yu (2026)	Knowledge-worker job postings (labor-market analysis)	AI-skill demand and wage premium	27.8% of jobs required AI skills; demand grew 376% post-release; 17.7% wage premium for AI-trained staff	Labor-market signal rather than a direct productivity measure

In educational settings, ChatGPT was associated with improvements in comprehension, engagement, and writing quality when its use was pedagogically structured (Chen & Xu, 2024; Garcia et al., 2025; Pei et al., 2025; Ravšelj et al., 2025; Tsai et al., 2024). These gains, however, were consistently reported alongside integrity-related risks: multiple systematic reviews documented cheating, plagiarism, and weakened independent problem-solving among students who relied heavily on the tool, prompting recommendations for AI literacy training and redesigned assessment (Cotton et al., 2024; Munaye et al., 2025). Table 2 reports the specific outcome measures and associated caveats identified across these educational studies.

Table 2. *Educational Learning and Writing Outcomes*

Study	Educational Context	Outcome Measure	Reported Result	Associated Caveat
Ravšelj et al. (2025)	Global higher-education student survey	Comprehension and engagement	Improved comprehension and higher engagement reported	Perceived as less reliable for critical thinking
Garcia et al. (2025)	Science-education classroom study	Explanation and comprehension quality	Clearer explanations and improved comprehension	Not reported

Chen & Xu (2024)	Meta-analysis of education interventions	Creativity and cognitive development	Measurable gains in creativity and cognitive-development indicators	Effects vary across underlying study designs
Tsai et al. (2024)	Controlled EFL writing intervention (university)	Essay score after AI-assisted revision	Significant improvement over pre-revision baseline	Not reported
Pei et al. (2025)	Secondary-education EFL quasi-experiment	Writing outcome vs. conventional instruction	ChatGPT-supported instruction outperformed conventional methods	Not reported
Cotton et al. (2024); Munaye et al. (2025)	Systematic reviews of academic-integrity risk	Academic integrity and independent reasoning	Consistent evidence of cheating, plagiarism, and weakened independent problem-solving	Recommends AI-literacy training and redesigned assessment

Software development emerged as the domain with the most extensively benchmarked evidence. ChatGPT outperformed comparable coding assistants on standardized generation tasks and produced faster or higher-quality solutions than community forums on well-defined student assignments (da Silva et al., 2025; Fatahi et al., 2024; Kalyan, 2024). On bug-repair benchmarks, performance was highly sensitive to the amount of contextual information supplied in the prompt, and it remained competitive with, though not superior to, dedicated automated-repair tools (Koubaa et al., 2023; Liu et al., 2023). Broader software-engineering task suites showed more mixed effectiveness, reinforcing the need for human validation on large or complex codebases (Kim & Ming, 2025). Table 3 presents the outcome measures and limitations reported across these six coding-focused studies.

Table 3. *Software Development and Debugging Benchmark Outcomes*

Study	Benchmark / Task	Comparison Baseline	Reported Result	Limitation Noted
Kalyan (2024)	HumanEval code generation	GitHub Copilot; CodeWhisperer	65.2% correct, versus 46.3% or lower for comparators	HumanEvalPlus reveals additional faulty code under stronger testing
Fatahi et al. (2024)	Student algorithmic coding tasks	Stack Overflow answers	Faster solution generation than Stack Overflow	Not reported
da Silva et al. (2025)	Student algorithm/library coding tasks	Stack Overflow answers	Higher code quality on algorithm/library tasks	Stack Overflow remained more helpful for debugging
Koubaa et al. (2023)	IEEEExtreme competitive-programming bug repair	Refactory automated repair tool	Competitive performance; improved substantially with added context	Still trailed experienced human programmers
Liu et al. (2023)	QuixBugs bug-repair benchmark	Limited vs. enriched prompt context	19/40 bugs fixed with limited context; 31/40 with richer context	Performance is prompt-sensitive
Kim & Ming (2025)	Broader software-engineering task suite	Cross-tool reliability comparison	Mixed effectiveness across large-context engineering tasks	Requires strong human validation before deployment

In marketing and customer-facing industries, ChatGPT use concentrated on content marketing, search-engine optimization, and customer-service automation, with adoption strongest where text production was high-volume and turnaround time mattered (Singh et al., 2024; Tafesse & Wien, 2024). Reported outcomes included measurable efficiency gains, including a workflow in which AI-assisted human refinement outperformed human-only production in search-ranking performance while lowering production cost (Brüns & Meißner, 2024). At the same time, professional and public-facing use raised recurring trust concerns, including diminished perceived authenticity when AI authorship was disclosed and mixed trust ratings among clinicians evaluating chatbot responses (Arslanoğlu & Karakose, 2025; Iyengar et al., 2023). Table 4 summarizes the corresponding outcome measures and trust-related concerns.

Table 4. Marketing, Customer Support, and Content Production Outcomes

Study	Domain / Application	Outcome Measure	Reported Result	Trust / Quality Concern
Tafesse & Wien (2024)	Topic-modeling analysis of marketing applications	Application breadth	Concentrated in content marketing, SEO, and B2B communication	Not reported
Singh et al. (2024)	Marketing decision-systems review	Adoption pattern	Widespread use for customer strategy and service automation	Occasionally neutral but weakly contextual responses
Brüns & Meißner (2024)	Social-media content production study	Content quality, search ranking, and cost	AI-assisted human-refinement workflow outperformed human-only production in search rankings at lower cost	Perceived brand authenticity diminished when AI authorship is disclosed
Arslanoğlu & Karakose (2025)	Trustworthiness analysis of health-related chatbot responses	Trust and reliability perception	Mixed trust ratings across response types	Privacy and contextual-understanding concerns
Iyengar et al. (2023)	Survey of musculoskeletal clinicians' perceptions	Professional trust perception	Generally positive perception of utility	Security and contextual-understanding concerns raised by professionals

Healthcare and research applications produced the richest evidence on both efficiency and reliability. Administrative time for clinical documentation was substantially reduced in at least one report, positioning ChatGPT as a useful adjunct in low-risk support roles (Tangsrivimol et al., 2025). However, the same domain produced the strongest evidence of unreliability: a large proportion of fabricated citations in guideline-related question answering, a non-trivial incorrect-recommendation rate in oncology decision support, and variable accuracy across neurological, toxicological, and dental-diagnosis tasks (Chen et al., 2023; Li et al., 2025; Madfa et al., 2025; Qiu & Zhou, 2024; Shiraishi et al., 2024). Patient-facing responses were often comprehensive and reproducible but not always sufficient for clinical decision-making on their own (Slawaska-Eng et al., 2025). Table 5 consolidates these healthcare, research, and cross-domain reliability metrics.

Table 5. Healthcare, Research, and Cross-Domain Reliability Metrics

Study	Domain / Task	Metric	Reported Result	Implication
Tangsrivimol et al. (2025)	Clinical discharge-summary drafting	Administrative time	70% reduction in administrative time	Useful in low-risk documentation support roles
Shiraishi et al. (2024)	Blepharoptosis guideline Q&A and citation generation	Reference / citation accuracy	Only 216 of 697 proposed references (31.0%) actually existed	Major fabricated-citation risk for research use
Li et al. (2025)	Multidisciplinary gastric-cancer decision-making	Incorrect-recommendation rate	5.25% incorrect recommendations	Unsafe as a standalone clinical decision-maker
Chen et al. (2023)	Neurological exam grading-scale scoring	Incorrect-output rate	10%-40% incorrect scoring depending on scale	Variable accuracy across clinical scoring tasks
Qiu & Zhou (2024)	Carbon-monoxide poisoning Q&A	Response acceptability	Roughly one-third of responses showed unacceptable inaccuracy	Reliability concerns extend to toxicology guidance
Madfa et al. (2025)	Case-based dental-diagnosis comparison	Diagnostic error rate	Nearly three times more errors than Claude or Manus	Model-to-model reliability varies substantially

Study	Domain / Task	Metric	Reported Result	Implication
Slawas-ka-Eng et al. (2025)	Ophthalmology and hip-surgery patient-facing Q&A	Comprehensiveness and reproducibility	59.8% of responses comprehensive; 91.5% reproducible	Adequate for general information, insufficient for clinical decision-making alone

Taken together, the quantitative results in Tables 1 through 5 converge on a consistent cross-domain pattern: ChatGPT produced the largest and most reliable gains on bounded, language-heavy, and well-structured tasks such as drafting, summarizing, and routine code generation, while its performance became markedly less dependable on tasks requiring specialized judgment, long or complex context, or verified factual accuracy (Bilika et al., 2024; Kalyan, 2024; Navas et al., 2024). This pattern held across otherwise very different domains, from consulting and software engineering to clinical documentation and marketing content, suggesting that the underlying determinant of performance was task structure rather than industry sector. The reliability findings in Table 5, in particular, reinforce a broader concern raised throughout the reviewed literature: fluent and confident-sounding outputs did not guarantee factual or clinical accuracy, which is why nearly every domain examined recommended human verification as a necessary complement to ChatGPT-generated content (Suárez et al., 2024).

Discussion

The findings of this review indicate that ChatGPT's benefits are neither uniform nor domain-bound but instead follow a consistent, task-dependent pattern. Across workplace, educational, coding, marketing, and healthcare contexts, the largest and most reliable gains emerged on bounded, language-heavy tasks such as drafting, summarizing, and routine code generation, whereas performance became markedly less dependable on tasks requiring specialized judgment, cultural or contextual nuance, or verified factual accuracy. This convergence across otherwise unrelated sectors suggests that task structure, rather than industry classification, is the primary determinant of ChatGPT's reliability. The recurring co-occurrence of measurable efficiency gains alongside non-trivial error rates further indicates that fluency and correctness are largely independent properties of ChatGPT's output — a distinction of particular interest to the humanities, where the persuasive surface of language has long been understood as separable from its truth-value — reinforcing the need for structured human oversight regardless of domain.

These results align closely with earlier workplace and technical-performance research. Preregistered experiments and consulting-sector studies similarly reported substantial reductions in completion time alongside quality improvements, particularly for professionals with lower baseline performance (Jo & Park, 2024; Kim & Lee, 2025; Sanatizadeh et al., 2025). Comparable evidence from software engineering confirms that ChatGPT accelerates coding tasks while remaining sensitive to prompt context and task familiarity, consistent with earlier benchmark findings on code generation and bug repair (Kalyan, 2024; Kim & Ming, 2025; Kuhail et al., 2024; Liu et al., 2023). The present synthesis therefore corroborates prior claims that ChatGPT functions best as an augmentative tool rather than an autonomous replacement for expert judgment (Azaria et al., 2024; Matzko & Konur, 2024).

The present findings also extend earlier education and healthcare literature by situating reliability concerns within a broader cross-domain pattern rather than treating them as sector-specific anomalies. Prior studies documented fabricated citations, inconsistent clinical recommendations, and academic-integrity risks largely in isolation (Cotton et al., 2024; Li et al., 2025; Shiraishi et al., 2024). By comparing these findings alongside marketing and workplace evidence, this review shows that such reliability limitations recur wherever tasks demand contextual judgment or verified accuracy, echoing earlier calls for human-in-the-loop verification across disciplines (Bilika et al., 2024; Sallam, 2023; Suárez et al., 2024). This cross-sectoral confirmation strengthens the argument that trust concerns are systemic rather than incidental to any single field of application. These findings carry several practical implications. Organizations adopting ChatGPT for knowledge work, content production, or technical development should design workflows that pair AI-generated first drafts with mandatory human verification, particularly for tasks involving factual claims, citations, or specialized judgment. Educational institutions should similarly prioritize AI-literacy training and redesigned assessment strategies that acknowledge both the writing-support benefits and the integrity risks associated with student use. In healthcare,

the evidence supports restricting ChatGPT to low-risk, adjunctive roles such as documentation support rather than autonomous clinical decision-making. Collectively, these implications suggest that responsible adoption depends less on restricting ChatGPT use outright and more on calibrating oversight to match the structural demands of each task.

Implications for Digital Humanities Scholarship

Beyond its practical implications for organizations, schools, and clinics, this review speaks to concerns that are central to the digital humanities: how the introduction of a fluent, non-human author into everyday textual practice reconfigures notions of authorship, literacy, and interpretive authority. The recurring finding that ChatGPT's fluency does not reliably track accuracy echoes a long-standing humanities concern with the rhetorical persuasiveness of language independent of its truth-value, and it suggests that critical reading and source evaluation — skills traditionally cultivated within literary and rhetorical study — are becoming more, not less, essential as generative AI mediates communication across education, work, and care. The cross-domain persistence of trust concerns also points to a cultural dimension of adoption: institutions and individuals are not simply deciding whether to use a tool but are renegotiating who, or what, counts as a credible speaker in a given context. Digital humanities scholarship is well positioned to study these negotiations, both by documenting how communities of practice adapt their norms of authorship and citation in response to generative AI, and by contributing humanistic frameworks for critical AI literacy that complement the technical and managerial oversight measures identified in this review.

This review is subject to several limitations. First, the narrative synthesis approach, while suited to capturing cross-domain patterns, does not permit the statistical pooling of effect sizes achieved through formal meta-analysis. Second, the included studies varied considerably in methodological rigor, sample size, and outcome measurement, complicating direct comparison across domains. Third, most reviewed studies examined short-term task performance, leaving longer-term effects on learning retention, clinical outcomes, and organizational return on investment largely unaddressed. Finally, because ChatGPT continues to be updated, findings based on earlier model versions may not fully generalize to current or future iterations, limiting the durability of some conclusions drawn from the reviewed literature.

Future research should prioritize longitudinal designs capable of capturing sustained effects of ChatGPT use on learning outcomes, workplace performance, and clinical decision-making beyond initial task completion. Direct comparative studies evaluating ChatGPT against alternative AI assistants outside coding contexts would help clarify whether observed advantages are tool-specific or generalizable. Additional research is also needed to quantify downstream costs associated with verification and error correction, which remain poorly measured despite their practical significance for adoption decisions. Finally, studies explicitly testing the task-structure hypothesis proposed in this review, comparing bounded versus complex tasks within controlled experimental designs, would help establish more definitive causal evidence regarding when ChatGPT can be trusted for unsupervised use.

Conclusion

This review demonstrates that ChatGPT delivers meaningful and reproducible benefits across workplace, educational, technical, marketing, and healthcare domains, particularly for bounded, language-heavy tasks, while consistently unsettling assumptions about authorship, literacy, and interpretive authority that have long been central to humanities inquiry. These benefits are consistently accompanied by reliability limitations that transcend sector boundaries, underscoring that trust in ChatGPT-generated language cannot be assumed uniformly across contexts and must instead be earned through critical reading and verification. The evidence points toward a clear practical model: ChatGPT functions most effectively as an augmentative tool within human-supervised workflows rather than as an autonomous decision-maker or an autonomous author. As adoption continues to expand from the workplace to the classroom and beyond, sustained attention to task-appropriate verification and to the cultivation of critical AI literacy will remain essential to realizing ChatGPT's benefits while minimizing the risks associated with its documented limitations — a task to which the digital humanities, with their long engagement in questions of language, authorship, and interpretation, are well suited to contribute.

Funding

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/37091)

References

- Alsawaylim, A. M., Alshehri, M. A., Lamouchi, A., Abdellatif, M. S., Ibrahim, A. R., & Nemt-Allah, M. A. (2026). ChatGPT use and learner autonomy among EFL university students: the mediating role of English speaking anxiety. *Frontiers in Psychology*, 17, 1834791. <https://doi.org/10.3389/fpsyg.2026.1834791>
- Arslanoğlu, K., & Karaköse, M. (2025). A trustworthy analysis approach for chatbots on health data: ChatGPT-4 example. In *2025 29th International Conference on Information Technology (IT)* (pp. 1–4). IEEE. <https://doi.org/10.1109/IT64745.2025.10930304>
- Azaria, A., Azoulay, R., & Reches, S. (2023). ChatGPT is a Remarkable Tool—For Experts. *Data Intelligence*, 6(1), 240–296. https://doi.org/10.1162/dint_a_00235
- Bego, C. R. (2023). Using ChatGPT for homework: Does it feel like cheating? (WIP). In *2023 IEEE Frontiers in Education Conference (FIE)* (pp. 1–4). IEEE. <https://doi.org/10.1109/FIE58773.2023.10343397>
- Bilika, P., Stefanouli, V., Strimpakos, N., & Kapreli, E. V. (2024). Clinical reasoning using ChatGPT: Is it beyond credibility for physiotherapists use? *Physiotherapy Theory and Practice*, 40(12), 2943–2962. <https://doi.org/10.1080/09593985.2023.2291656>
- Brüns, J. D., & Meißner, M. (2024). Do you create your content yourself? Using generative artificial intelligence for social media content creation diminishes perceived brand authenticity. *Journal of Retailing and Consumer Services*, 79, 103790. <https://doi.org/10.1016/j.jretconser.2024.103790>
- Chen, X., & Xu, L. (2024). Effectiveness of ChatGPT in education: A meta-analysis. In *2024 5th International Conference on Information Science and Education (ICISE-IE)* (pp. 428–431). IEEE. <https://doi.org/10.1109/ICISE-IE64355.2024.11025431>
- Choi, W., Zhang, Y., & Stvilia, B. (2023). Exploring Applications and User Experience with Generative AI Tools: A Content Analysis of Reddit Posts on ChatGPT. *Proceedings of the Association for Information Science and Technology*, 60(1), 543–546. <https://doi.org/10.1002/pra2.823>
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. <https://doi.org/10.1080/14703297.2023.2190148>
- Crompton, H., & Burke, D. (2024). The Educational Affordances and Challenges of CHATGPT: State of the Field. *TechTrends*, 68(2), 380–392. <https://doi.org/10.1007/s11528-024-00939-0>
- Da Silva, L., Samhi, J., & Khomh, F. (2025). LLMs and Stack Overflow discussions: Reliability, impact, and challenges. *Journal of Systems and Software*, 230, 112541. <https://doi.org/10.1016/j.jss.2025.112541>
- DePalma, K., Miminoshvili, I., Henselder, C., Moss, K., & AlOmar, E. A. (2024). Exploring ChatGPT's code refactoring capabilities: An empirical study. *Expert Systems With Applications*, 249, 123602. <https://doi.org/10.1016/j.eswa.2024.123602>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Fatahi, S., Vassileva, J., & Roy, C. K. (2024). Comparing emotions in ChatGPT answers and human answers to the coding questions on Stack Overflow. *Frontiers in Artificial Intelligence*, 7, 1393903. <https://doi.org/10.3389/frai.2024.1393903>
- Garcia, R., Fojas, J. G., Formanes, J. N., Garcia, D., Habol, C., & Orapa, L. (2025). Perspective of science Educators and students on the use of ChATGPT: A study of Artificial intelligence. *International Journal of Technology in Education and Science*, 9(4), 474–491. <https://doi.org/10.46328/ijtes.649>
- Han, Z., Battaglia, F., Udaiyar, A., Fooks, A., & Terlecky, S. R. (2023). An explorative assessment of ChatGPT as an aid in medical education: Use it with caution. *Medical Teacher*, 46(5), 657–664. <https://doi.org/10.1080/0142159x.2023.2271159>
- Imran, M., Almusharraf, N., & Dalbani, H. (2024). Future of writing in higher education: A review of prospective use of artificial intelligence and ChatGPT in academia. In M. M. Asad, P. P. Churi, F. Sherwani, & R. B. Hassan (Eds.), *Innovative pedagogical practices for higher education 4.0: Solutions and demands of the*

- modern classroom* (pp. 108–125). CRC Press. <https://doi.org/10.1201/9781003400691>
- Ishida, K., Arisaka, N., & Fujii, K. (2024). Analysis of responses of GPT-4 V to the Japanese National Clinical Engineer Licensing Examination. *Journal of Medical Systems*, 48(1), 83. <https://doi.org/10.1007/s10916-024-02103-w>
- Iyengar, K. P., Yousef, M. M. A., Nune, A., Sharma, G. K., & Botchu, R. (2023). Perception of Chat Generative Pre-trained Transformer (Chat-GPT) AI tool amongst MSK clinicians. *Journal of Clinical Orthopaedics and Trauma*, 44, 102253. <https://doi.org/10.1016/j.jcot.2023.102253>
- Jain, R., Thanvi, J., & Subasinghe, A. (2025). The evolution of ChatGPT for programming: a comparative study. *Engineering Research Express*, 7(1), 015242. <https://doi.org/10.1088/2631-8695/ada51d>
- Jo, H., & Park, D. (2023). AI in the Workplace: Examining the Effects of ChatGPT on Information Support and Knowledge Acquisition. *International Journal of Human-Computer Interaction*, 40(23), 8091–8106. <https://doi.org/10.1080/10447318.2023.2278283>
- Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6, 100048. <https://doi.org/10.1016/j.nlp.2023.100048>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kim, B., & Lee, J. (2025). AI adoption, employee depression and knowledge: How corporate social responsibility buffers psychological impact. *Journal of Innovation & Knowledge*, 10(6), 100815. <https://doi.org/10.1016/j.jik.2025.100815>
- Kim, D., & Ming, H. (2025). Assessing output reliability and similarity of large language models in software development: A comparative case study approach. *Information and Software Technology*, 185, 107787. <https://doi.org/10.1016/j.infsof.2025.107787>
- Koubaa, A., Qureshi, B., Ammar, A., Khan, Z., Boulila, W., & Ghouti, L. (2023). Humans are still better than ChatGPT: Case of the IEEEExtreme competition. *Heliyon*, 9(11), e21624. <https://doi.org/10.1016/j.heliyon.2023.e21624>
- Kuhail, M. A., Mathew, S. S., Khalil, A., Berengueres, J., & Shah, S. J. H. (2024). “Will I be replaced?” Assessing ChatGPT’s effect on software development and programmer perceptions of AI tools. *Science of Computer Programming*, 235, 103111. <https://doi.org/10.1016/j.scico.2024.103111>
- Li, H., Huang, J., Liu, K., Liu, J., Liu, Q., Zhou, Z., Zong, Z., & Mao, S. (2025). ChatGPT-4o outperforms gemini advanced in assisting multidisciplinary decision-making for advanced gastric cancer. *European Journal of Surgical Oncology*, 51(8), 110096. <https://doi.org/10.1016/j.ejso.2025.110096>
- Libera, P. S., Bilgram, V., Schötteler, S., & Mammen, J. (2025). ChatGPT in the working world. *Information Resources Management Journal*, 38(1), 1–26. <https://doi.org/10.4018/irmj.386593>
- Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2), 100017. <https://doi.org/10.1016/j.metrad.2023.100017>
- Matzko, R. O., & Konur, S. (2024). BioNexusSentinel: a visual tool for bioregulatory network and cytohistological RNA-seq genetic expression profiling within the context of multicellular simulation research using ChatGPT-augmented software engineering. *Bioinformatics Advances*, 4(1), vbae046. <https://doi.org/10.1093/bioadv/vbae046>
- Navas, G., Navas-Reascos, G., Navas-Reascos, G. E., & Proaño-Orellana, J. (2024). Exploring the effectiveness of advanced chatbots in educational settings: A Mixed-Methods Study in Statistics. *Applied Sciences*, 14(19), 8984. <https://doi.org/10.3390/app14198984>
- Nemt-Allah, M., Khalifa, W., Badawy, M., Elbably, Y., & Ibrahim, A. (2024). Validating the ChatGPT Usage Scale: psychometric properties and factor structures among postgraduate students. *BMC psychology*, 12(1), 497. <https://doi.org/10.1186/s40359-024-01983-4>
- Pavone, M., Palmieri, L., Bizzarri, N., Rosati, A., Campolo, F., Innocenzi, C., Taliento, C., Restaino, S., Catena,

- U., Vizzielli, G., Akladios, C., Ianieri, M., Marescaux, J., Campo, R., Fanfani, F., & Scambia, G. (2024). Artificial Intelligence, the ChatGPT Large Language Model: Assessing the Accuracy of Responses to the Gynaecological Endoscopic Surgical Education and Assessment (GESEA) Level 1-2 knowledge tests. *Facts Views and Vision in ObGyn*, 16(4), 449–456. <https://doi.org/10.52054/fvvo.16.4.052>
- Pei, L., Jong, M. S.-Y., Huang, B., Pang, W.-C., & Shang, J. (2025). Formally integrating generative AI into secondary education: Application of ChatGPT in EFL writing instruction. *Educational Technology & Society*, 28(3), 281–297. [https://doi.org/10.30191/ETS.202507_28\(3\).TP05](https://doi.org/10.30191/ETS.202507_28(3).TP05)
- Rasul, T., Nair, S., Kalendra, D., Robin, M., De Oliveira Santini, F., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning & Teaching*, 6(1), 41-56. <https://doi.org/10.37074/jalt.2023.6.1.29>
- Ravšelj, D., Keržič, D., Tomažević, N., Umek, L., Brezovar, N., Iahad, N. A., Abdulla, A. A., Akopyan, A., Segura, M. W. A., AlHumaid, J., Allam, M. F., Alló, M., Andoh, R. P. K., Andronic, O., Arthur, Y. D., Aydın, F., Badran, A., Balbontín-Alvarado, R., Saad, H. B., . . . Aristovnik, A. (2025). Higher education students' perceptions of ChatGPT: A global study of early reactions. *PLoS ONE*, 20(2), e0315011. <https://doi.org/10.1371/journal.pone.0315011>
- Riedel, M., Kaefinger, K., Stuehrenberg, A., Ritter, V., Amann, N., Graf, A., Recker, F., Klein, E., Kiechle, M., Riedel, F., & Meyer, B. (2023). ChatGPT's performance in German OB/GYN exams – paving the way for AI-enhanced medical education and clinical practice. *Frontiers in Medicine*, 10, 1296615. <https://doi.org/10.3389/fmed.2023.1296615>
- Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the promising perspectives and valid concerns. *Healthcare*, 11(6), 887. <https://doi.org/10.3390/healthcare11060887>
- Sanatizadeh, A., Lu, Y., Zhao, K., & Hu, Y. (2025). Engagement or entanglement? The dual impact of generative artificial intelligence in online knowledge exchange platforms. *Information & Management*, 62(6), 104178. <https://doi.org/10.1016/j.im.2025.104178>
- Shiraishi, M., Tomioka, Y., Miyakuni, A., Ishii, S., Hori, A., Park, H., Ohba, J., & Okazaki, M. (2024). Performance of CHATGPT in answering clinical questions on the practical guideline of blepharoptosis. *Aesthetic Plastic Surgery*, 48(13), 2389–2398. <https://doi.org/10.1007/s00266-024-04005-1>
- Singh, H., Arora, M., & Singh, A. (2024). ChatGPT in marketing: innovative pathways, decision systems, and forward perspectives. *Journal of Decision System*, 1–28. <https://doi.org/10.1080/12460125.2024.2438615>
- Sohail, S. S., Farhat, F., Himeur, Y., Nadeem, M., Madsen, D. Ø., Singh, Y., Atalla, S., & Mansoor, W. (2023). Decoding ChatGPT: A taxonomy of existing research, current challenges, and possible future directions. *Journal of King Saud University - Computer and Information Sciences*, 35(8), 101675. <https://doi.org/10.1016/j.jksuci.2023.101675>
- Suárez, A., García, V. D., Algar, J., Sánchez, M. G., De Pedro, M. L., & Freire, Y. (2024). Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *International Endodontic Journal*, 57(1), 108–113. <https://doi.org/10.1111/iej.13985>
- Tafesse, W., & Wien, A. (2024). ChatGPT's applications in marketing: a topic modeling approach. *Marketing Intelligence & Planning*, 42(4), 666–683. <https://doi.org/10.1108/mip-10-2023-0526>
- Takerngsaksiri, W., Charakorn, R., Tantithamthavorn, C., & Li, Y. (2025). Pytester: Deep reinforcement learning for text-to-testcase generation. *Journal of Systems and Software*, 224, 112381. <https://doi.org/10.1016/j.jss.2025.112381>
- Tangsrivimol, J. A., Darzidehkalani, E., Virk, H. U. H., Wang, Z., Egger, J., Wang, M., Hacking, S., Glicksberg, B. S., Strauss, M., & Krittanawong, C. (2025). Benefits, limits, and risks of ChatGPT in medicine. *Frontiers in Artificial Intelligence*, 8, 1518049. <https://doi.org/10.3389/frai.2025.1518049>
- Thant, K. S., Mon Khaing, M., & Khaung Tin, H. H. (2024). Evaluating the efficacy of ChatGPT in different domains: Customer support vs. educational assistance. In *2024 5th International Conference on Advanced Information Technologies (ICAIT)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICAIT65209.2024.10754924>
- Tsai, C., Lin, Y., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility

and challenges. *Education and Information Technologies*, 29(17), 22427–22445. <https://doi.org/10.1007/s10639-024-12722-y>

Yu, Z., & Yu, Z. (2026). Reconstruction of knowledge worker performance evaluation system in the ChatGPT era: an exploratory study based on human-AI collaborative work model. *Future Technology*, 5(1), 47–54. <https://doi.org/10.55670/fpl.futech.5.1.5>

Zehir, M., & Yilmaz, M. K. (2025). Unveiling the transformative influence of ChatGPT on service sector. *Lecture Notes in Mechanical Engineering*, 145–166. https://doi.org/10.1007/978-3-031-83583-4_10