ReGeo ISSN: 2177-3246

METHODOLOGICAL TRANSPARENCY AUDIT IN HUMAN-AI COLLABORATION RESEARCH: AI MODEL AND VERSION REPORTING ACROSS 439 STUDIES

AUDITORIA DE TRANSPARÊNCIA METODOLÓGICA EM PESQUISA DE COLABORAÇÃO HUMANO-IA: REPORTE DE MODELOS E VERSÕES DE SISTEMAS DE INTELIGÊNCIA ARTIFICIAL EM 439 ESTUDOS

AUDITORÍA DE TRANSPARENCIA METODOLÓGICA EN LA INVESTIGACIÓN SOBRE COLABORACIÓN HUMANO-IA: REPORTE DE MODELOS Y VERSIONES DE SISTEMAS DE INTELIGENCIA ARTIFICIAL EN 439 ESTUDIOS

https://doi.org/10.56238/revgeov17n5-154

Data de submissão: 16/04/2026 **Data de publicação:** 16/05/2026

Marcelo Caetano de Ribeiro e Melo[1], Marcos Alves da Rocha[2], Silvania Soares Brandão[3], Jordana Layara Alves Dias[4]

ABSTRACT

Reproducibility in empirical research requires that the tools of inquiry be fully identified. In studies involving AI systems, this requirement extends to naming the AI model and specifying the version or release identifier, since different versions of nominally the same model can exhibit materially distinct behavioral profiles. This paper presents a systematic audit of AI model and version reporting practices across a corpus of 439 primary studies on human-AI team collaboration. The corpus was inherited from two prior PRISMA-compliant systematic reviews and spans publications from 2018 to 2026. Studies were classified into three evidence tiers based on available documentation depth: full-text (n=67), abstract (n=272), and metadata (n=99). A five-category coding scheme was applied to each record: S2 (specific model with explicit version identifier), S1 (specific named model without version), G (generic or vague description), NA (not applicable, including theoretical and Wizard of Oz designs), and NR (not reported despite apparent empirical AI use). Primary analysis was restricted to the full-text tier. Among the 42 empirically active studies in that tier, 76.2 percent identified the AI system by name, 21.4 percent provided a version-specific identifier, and 9.5 percent disclosed no model information at all. A pronounced temporal pattern emerged: no study published before 2025 provided version-specific information, while 34.6 percent of empirically active full-text studies from 2025 onward did so. These findings indicate partial but incomplete adoption of transparent reporting norms, with version-specific disclosure representing the primary remaining gap. Three minimum reporting elements are proposed as a field-level standard: model name, version or release identifier, and access method.

[1] Ph.D. in Biofuel Science and Technology. Institution: Universidade Estadual de Montes Claros (UNIMONTES). E-mail: marcelo.melo@unimontes.br

[2] Master's degree in Economic Development and Business Strategy. Institution: Universidade Estadual de Montes Claros (UNIMONTES). E-mail: marcos.alves@unimontes.br

[3] Master's degree in Business Administration. Institution: Instituto Federal do Norte de Minas Gerais (IFNMG). E-mail: silvania.brandao@ifnmg.edu.br

[4] Specialist in Accounting Management and Corporate Controllership. Institution: Faculdades Integradas do Norte de Minas (FUNORTE). E-mail: jordana.dias@soufunorte.com.br

Keywords: Human-AI Collaboration. AI Model Reporting. Methodological Transparency. Reproducibility. Version Reporting. Reporting Standards.

RESUMO
A reprodutibilidade na pesquisa empírica exige que as ferramentas de investigação sejam completamente identificadas. Em estudos envolvendo sistemas de IA, esse requisito se estende à nomenclatura do modelo de IA e à especificação do identificador de versão ou lançamento, uma vez que diferentes versões do mesmo modelo nominalmente podem apresentar perfis comportamentais materialmente distintos. Este artigo apresenta uma auditoria sistemática das práticas de reporte de modelo e versão de IA em um corpus de 439 estudos primários sobre colaboração entre humanos e IA. O corpus foi herdado de duas revisões sistemáticas anteriores em conformidade com PRISMA e abrange publicações de 2018 a 2026. Os estudos foram classificados em três níveis de evidência com base na profundidade da documentação disponível: texto completo (n=67), resumo (n=272) e metadados (n=99). Um esquema de codificação de cinco categorias foi aplicado a cada registro: S2 (modelo específico com identificador de versão explícito), S1 (modelo nomeado específico sem versão), G (descrição genérica ou vaga), NA (não aplicável, incluindo designs teóricos e Wizard of Oz), e NR (não informado apesar do uso empírico aparente de IA). A análise primária foi restrita ao nível de texto completo. Entre os 42 estudos empiricamente ativos nesse nível, 76,2 por cento identificaram o sistema de IA pelo nome, 21,4 por cento forneceram um identificador específico de versão, e 9,5 por cento não divulgaram nenhuma informação de modelo. Um padrão temporal pronunciado emergiu: nenhum estudo publicado antes de 2025 forneceu informações específicas de versão, enquanto 34,6 por cento dos estudos empiricamente ativos de texto completo de 2025 em diante o fizeram. Esses achados indicam adoção parcial mas incompleta de normas de reporte transparente, com a divulgação específica de versão representando a lacuna principal restante. Três elementos de reporte mínimos são propostos como um padrão em nível de campo: nome do modelo, identificador de versão ou lançamento e método de acesso.

Palavras-chave: Colaboração Humano-IA. Reporte de Modelo de IA. Transparência Metodológica. Reprodutibilidade. Reporte de Versão. Padrões de Reporte.

RESUMEN
La reproducibilidad en la investigación empírica requiere que las herramientas de investigación sean completamente identificadas. En estudios que involucran sistemas de IA, este requisito se extiende a la denominación del modelo de IA y la especificación del identificador de versión o lanzamiento, ya que diferentes versiones del mismo modelo nominalmente pueden presentar perfiles comportamentales materialmente distintos. Este artículo presenta una auditoría sistemática de las prácticas de reporte de modelo y versión de IA en un corpus de 439 estudios primarios sobre colaboración humano-IA. El corpus fue heredado de dos revisiones sistemáticas anteriores conformes a PRISMA y abarca publicaciones de 2018 a 2026. Los estudios fueron clasificados en tres niveles de evidencia basados en la profundidad de la documentación disponible: texto completo (n=67), resumen (n=272) y metadatos (n=99). Se aplicó un esquema de codificación de cinco categorías a cada registro: S2 (modelo específico con identificador de versión explícito), S1 (modelo específico nombrado sin versión), G (descripción genérica o vaga), NA (no aplicable, incluyendo diseños teóricos y Wizard of Oz), y NR (no reportado a pesar del uso empírico aparente de IA). El análisis primario se restringió al nivel de texto completo. Entre los 42 estudios empíricamente activos en ese nivel, el 76,2 por ciento identificó el sistema de IA por

nombre, el 21,4 por ciento proporcionó un identificador específico de versión, y el 9,5 por ciento no divulgó información de modelo alguna. Un patrón temporal pronunciado emergió: ningún estudio publicado antes de 2025 proporcionó información específica de versión, mientras que el 34,6 por ciento de los estudios empíricamente activos de texto completo a partir de 2025 en adelante lo hicieron. Estos hallazgos indican una adopción parcial pero incompleta de normas de reporte transparente, siendo la divulgación específica de versión la brecha principal restante. Se proponen tres elementos de reporte mínimos como un estándar a nivel de campo: nombre del modelo, identificador de versión o lanzamiento y método de acceso.

Palabras clave: Colaboración entre Personas e IA. Informe Sobre el Modelo de IA. Transparencia Metodológica. Reproducibilidad. Informe de Versión. Normas de Información.

1 INTRODUCTION

The integration of AI systems into collaborative work has accelerated substantially since 2018, driven by successive advances in large language models, conversational agents, and decision-support algorithms. Teams in healthcare, engineering, education, creative industries, and organizational management now routinely include AI components that contribute to planning, analysis, communication, and decision-making (SEEBER *et al.*, 2020; DELLERMANN *et al.*, 2019). Research on these human-AI configurations has grown correspondingly, producing a body of literature that spans multiple disciplines and employs diverse theoretical frameworks and empirical designs. The volume and disciplinary breadth of this literature make it an important site for examining whether the methodological practices that underpin scientific credibility are being adequately observed.

This growth brings with it a methodological challenge that the field has not yet fully confronted. When an empirical study involves AI participation in team processes, the identity of the AI system is a methodological variable of first-order importance. AI systems are not static instruments in the manner of standardized psychological scales or fixed physical apparatus. A model designated “GPT-4” in early 2023, for instance, is a different computational artifact from what bore the same designation twelve months later, and both differ substantially from predecessor and successor versions in the same family. The behavioral consequences of such differences, including changes in reasoning patterns, factual accuracy, instruction-following reliability, and response length, are not trivial. They bear directly on what a study can be said to have measured and on whether its findings can be reproduced by an independent researcher working at a different point in time (COLLINS; MOONS, 2019; LIU *et al.*, 2020).

An additional complication, which distinguishes AI systems from virtually all prior scientific instruments, is that version changes in commercial AI products are not always formally announced or publicly documented. A study conducted with a consumer-facing AI product in a given month may have employed a substantially different model from one conducted in the same product interface three months earlier, with no indication visible to the user that any change occurred. This “silent update” problem means that even a researcher who records the product name and access date may be unable to reconstruct the precise computational entity that participated in the study, unless the specific API version or model checkpoint was also recorded. Against this backdrop, the question of whether the field routinely captures even the minimum necessary instrument information is far from trivial.

The scientific literature has long grappled with a broader reproducibility crisis (IOANNIDIS, 2005; OPEN SCIENCE COLLABORATION, 2015; BAKER, 2016). These concerns, originating primarily in biomedical and psychological research, apply with equal or greater force to human-AI collaboration research, where the instrument under study, the AI system itself, changes continuously and unpredictably across releases, updates, and fine-tuning cycles. The convergence of the general reproducibility problem with the specific versioning challenge of AI systems creates a compound methodological risk that warrants systematic investigation.

Despite the availability of reporting guidelines for AI-based clinical tools (LIU *et al.*, 2020; COLLINS; MOONS, 2019), no systematic audit has examined compliance with model-and-version reporting norms specifically within the human-AI collaboration research literature. This paper fills that gap. It presents a systematic content analysis of AI model and version reporting practices across 439 primary studies drawn from two prior PRISMA-compliant systematic reviews of human-AI team collaboration (MOHER *et al.*, 2009). The corpus spans publications from 2018 to 2026 and covers a range of domains, study designs, and AI system types.

Two primary research questions are addressed. First: how consistently and specifically do published studies on human-AI team collaboration report the AI model and version used? Second: what temporal pattern characterizes version-specific reporting, and what does it imply about methodological maturity in the field? The answers to these questions establish an empirical baseline for a field-level discussion of reporting standards and inform a set of minimum reporting recommendations addressed to authors, reviewers, and journal editors working in this domain.

The paper is organized as follows. Section 2 reviews the relevant theoretical and methodological background, covering the human-AI teaming literature, the reproducibility crisis, AI versioning as a methodological variable, and existing reporting standards. Section 3 describes the corpus, coding scheme, and analytical approach. Section 4 presents and discusses the results. Section 5 offers conclusions, limitations, and directions for future work.

2 THEORETICAL BACKGROUND

2.1 HUMAN-AI TEAMING AS A RESEARCH DOMAIN

The systematic study of teams that include AI members has emerged as a distinct area of inquiry at the intersection of organizational behavior, human factors, information systems,

and computer science. Seeber *et al.* (2020) proposed a foundational agenda for research on AI in team collaboration, organizing the domain around three categories of AI roles: information processing, decision support, and active task participation. Their framing established that AI participation in teams is not a peripheral technical matter but a substantive organizational phenomenon warranting dedicated scholarly attention, and it identified a set of research priorities concerning team dynamics, performance outcomes, and trust relationships that has substantially shaped subsequent empirical work in the field. A key implication of their agenda is that the AI team member is a theoretically consequential actor whose properties must be specified with the same care applied to human participants.

Parallel contributions developed the concept of human-autonomy teaming, in which human operators and autonomous AI systems share task space in real time. McNeese *et al.* (2018) reported an experimental investigation of human performance when teamed with a synthetic agent in a simulated mission environment, demonstrating that the nature of the AI teammate's capabilities and behavioral profile materially influenced team coordination patterns and overall task outcomes. Their work highlights the degree to which conclusions about human-AI teamwork are sensitive to the specific AI partner employed. Levine and Williams (2018) examined a complementary challenge in the form of concurrent plan recognition in human-robot teams, and reported results obtained using a named symbolic planning system (PIKE) operating under specific algorithmic parameters. Both studies exemplify the design-level specificity that rigorous human-AI collaboration research requires, and both stand as methodological reference points for how AI instruments should be documented.

Dellermann *et al.* (2019) offered a broader conceptual integration under the hybrid intelligence framework, arguing that the combination of human and AI capabilities produces a system whose collective performance exceeds what either party could achieve independently. This framing has been influential in positioning human-AI collaboration as a long-term sociotechnical configuration rather than a short-term tool adoption. From a methodological standpoint, it reinforces the importance of specifying the AI component with precision, since the performance characteristics of the hybrid system depend directly on the capabilities of the AI contribution. A change in the AI component, such as a version update that alters its reasoning or communication style, is therefore equivalent to replacing a team member, not merely updating a software utility.

Empirical studies in the domain have examined an increasingly wide range of AI system types. Earlier work centered on robotic systems and rule-based agents with relatively stable behavioral profiles (LEVINE; WILLIAMS, 2018). More recent research has engaged extensively with generative AI systems, including large language models used in creative, analytical, educational, and organizational contexts. Oppenheimer *et al.* (2025) investigated large language models as collaborative learning partners across multiple experimental conditions, while Bossen and Pine (2023) examined human-AI collaboration in healthcare knowledge work, specifically the use of AI-assisted coding tools by clinical documentation integrity specialists. Georganta *et al.* (2023) synthesized findings across disciplines in a systematic review of interactive technologies in team effectiveness contexts, identifying persistent methodological heterogeneity as a key challenge for cumulative knowledge production in the field.

At the conceptual frontier, Sowa and Przegalinska (2025) proposed the construct of Generative Artificial Experts, a theoretical architecture for human-AI collaboration in knowledge work built on generative AI capabilities. While not an empirical trial of a deployed system, this work illustrates the degree to which the field continues to elaborate new conceptual models for understanding AI participation in teams and the diversity of levels at which AI involvement can be analyzed. The coexistence of empirical studies, systematic reviews, and theoretical frameworks within a single research corpus is a feature that any methodological audit of the domain must accommodate, and it is addressed in the evidence tier design described in Section 3.

2.2 REPRODUCIBILITY AND OPEN SCIENCE

The epistemological foundations of scientific knowledge rest on the assumption that reported findings can, in principle, be reproduced by independent investigators working with comparable methods. When this assumption fails at scale, the credibility of an entire field is compromised. Ioannidis (2005) presented a formal argument, grounded in statistical reasoning about positive predictive values under realistic conditions of effect prevalence, statistical power, and researcher degrees of freedom, that the majority of published research findings are expected to be false positives. This analysis did not indict individual researchers but pointed to systemic conditions that make spurious findings probable, including small sample sizes, flexible analytical choices, selective reporting of outcomes, and inadequate methodological documentation. The argument applies to any domain in which these

conditions obtain, including the nascent and rapidly evolving field of human-AI collaboration research.

The Open Science Collaboration (2015) provided direct empirical corroboration of the reproducibility concern by attempting to replicate 100 original psychology studies using the original authors' materials and procedures as closely as possible. Only 36 of the replication attempts recovered a statistically significant effect in the same direction as the original, and effect sizes in replications were on average approximately half the magnitude of those originally reported. These results prompted a broad reassessment of how scientific findings should be evaluated and how methodological documentation should be standardized. Baker (2016) documented the extent to which this concern had penetrated the practicing scientific community, finding that more than 70 percent of surveyed scientists had tried and failed to reproduce another researcher's findings, and that 52 percent agreed science faces a significant reproducibility crisis.

Institutional responses to the reproducibility concern have included pre-registration of hypotheses and analysis plans, open data mandates, registered reports that decouple publication decisions from outcomes, and the development of more detailed reporting standards. The key insight motivating these reforms is that a result whose methodology is incompletely described cannot be meaningfully evaluated, extended, or challenged. Incompleteness in method reporting is not merely a presentation problem; it is a validity problem, because a reader cannot assess the plausibility of a finding without understanding the precise conditions under which it was obtained.

For human-AI collaboration research, the methodological documentation requirement takes on a specific character that has no close analog in traditional behavioral science. In a psychology experiment, the instruments are typically standardized tests or apparatus that remain stable across administrations. In a human-AI collaboration study, the AI instrument may change between administrations even when the same product name is used, and it will almost certainly differ from any nominally equivalent instrument used by a subsequent replication team unless the version is precisely specified. This makes the documentation obligation for AI instruments more demanding than the analogous obligation for traditional instruments, and it makes the consequences of non-compliance more severe.

2.3 AI VERSIONING AS A METHODOLOGICAL VARIABLE

The methodological significance of AI versioning is underappreciated in the human-AI collaboration literature. Unlike a physical measuring instrument that remains constant unless physically altered, AI model versions represent distinct computational entities with potentially different behavioral profiles. Major generative AI providers have released successive model versions at a pace that has outstripped the ability of many research communities to develop corresponding documentation norms.

The practical consequences for research design are substantial. A study that used a particular large language model in early 2023 may have employed a model checkpoint that was subsequently deprecated, rendering any future replication with the “current version” of the same product a comparison between materially different instruments. Version differences can manifest as changes in task performance on standard benchmarks, shifts in instruction-following fidelity, alterations in the length and structure of generated outputs, and changes in calibration and uncertainty expression. Any of these differences could plausibly affect the outcomes of a human-AI collaboration study, particularly one that examines communication quality, decision accuracy, or task completion time in a joint human-AI workflow.

Oppenheimer *et al.* (2025) illustrate the challenge concretely in an educational context. Their study of large language models as collaborative learning partners named multiple AI systems (ChatGPT, Gemini, and Claude) across experimental conditions but provided no specification of which version of any system was used. This reporting pattern is common in recent literature and is entirely understandable given the absence of established norms. However, it limits the interpretive reach of the findings: a secondary analyst who seeks to understand whether observed differences between conditions reflect genuine contrasts between the named AI systems or incidental differences between the specific version instances deployed cannot resolve this question from the available documentation.

Kirkby *et al.* (2025), examining human-AI co-creation for creative brand voice, provided a specific version indicator (“ChatGPT-3”), thereby anchoring the study to a particular model generation with a known capability profile. The contrast between these two approaches, one study naming systems without version and the other providing a version identifier, illustrates a methodologically meaningful spectrum of reporting quality. Naming a model family is a necessary but not sufficient reporting step. A version or release identifier converts a vague instrument reference into a citable, auditable methodological element, and its presence or absence distinguishes reproducible from non-reproducible methods descriptions.

Beyond the individual study level, the versioning problem has cumulative consequences for the field as a whole. A meta-analyst who attempts to synthesize findings about the effect of AI partnership quality on human team performance across a literature in which AI instruments are only partially identified must either assume that all studies using "ChatGPT" deployed equivalent instruments, an assumption that is almost certainly false, or treat AI instrument identity as a moderating variable that cannot be coded from the available documentation. In the absence of version information, the former assumption prevails by default, introducing a source of heterogeneity into pooled estimates that cannot be evaluated, controlled, or communicated to readers.

2.4 EXISTING REPORTING STANDARDS AND THEIR RELEVANCE

Formal reporting guidelines for AI-based research have developed most extensively in clinical and biomedical contexts. Collins and Moons (2019) argued in a landmark commentary that reporting of AI prediction models was systematically inadequate, with most published reports omitting key details about model architecture, training data provenance, and performance benchmarking procedures. They called for extension of existing prediction model reporting standards (TRIPOD) to cover AI-specific requirements including version identification and deployment context.

Liu *et al.* (2020) subsequently developed CONSORT-AI, an extension of the CONSORT checklist for randomized trials, specifying that trial reports must identify the AI intervention including its version, the conditions under which it was accessed, any modifications to standard configurations, and the training and validation procedures applied. These requirements were motivated by precisely the kind of methodological ambiguity that characterizes the human-AI collaboration literature: a reader who cannot determine whether an AI system was used in its standard commercial configuration or in a fine-tuned, modified, or version-pinned form cannot evaluate the external validity or reproducibility of the reported results.

These standards have achieved variable uptake even within their target domains. Their application to human-AI collaboration research, where the AI system functions not as a clinical diagnostic tool but as a team member, communicative partner, or collaborative agent, has not been systematically promoted or examined. The structural differences between clinical AI deployment and experimental human-AI teaming studies, particularly the diversity of AI roles and the absence of a unifying regulatory or publication infrastructure, have slowed

the translation of clinical reporting norms to the social and organizational science literatures where much human-AI collaboration research appears.

The absence of established norms does not imply the absence of obligation. Method sections of empirical papers are expected to provide sufficient information that a competent researcher could replicate the study. For human-AI collaboration studies, this standard requires at minimum the identification of the AI system by name, by version, and by access method. Whether the field has met this standard in practice is the empirical question this paper addresses.

3 METHODOLOGY

3.1 CORPUS PROVENANCE

The corpus for this analysis comprises 439 primary studies on human-AI team collaboration assembled through two prior PRISMA-compliant systematic reviews (MOHER *et al.*, 2009). The originating reviews searched multiple databases including Web of Science, Scopus, IEEE Xplore, and PubMed, applying a common set of inclusion and exclusion criteria centered on the requirement that studies examine collaborative arrangements between human individuals or groups and AI systems in task-relevant settings. The corpus spans the period from 2018, when systematic empirical investigation of modern AI collaboration systems became feasible at scale, through early 2026, the closing date of the search update that produced the current corpus version.

The present paper inherits the corpus and does not conduct an independent literature search. This design choice reflects the purpose of the analysis: to audit the reporting practices of an existing body of literature rather than to identify new literature. One record was excluded from analysis prior to coding on the grounds that the study used human confederates posing as AI in a deception-based experimental design and therefore did not involve the deployment of any real AI system. The working corpus for the present analysis thus comprises 438 unique records after deduplication by DOI. Following the convention established in the originating systematic reviews, the corpus is referred to throughout as comprising 439 studies.

3.2 EVIDENCE TIERS

Corpus records were classified into three evidence tiers based on the depth of documentation available to the coder at the time of analysis. The full-text tier (FT, n=67) consists of studies for which complete article text was accessible and reviewed. This tier

supports the most reliable and granular coding decisions and is the basis for all primary findings reported in this paper. The abstract tier (AB, n=272) consists of studies for which only the abstract, author information, and bibliographic metadata were reviewed, without access to the full method section. Coding from this tier is substantially less reliable because method sections, the primary location for AI instrument documentation, are excluded. The metadata tier (MD, n=99) consists of studies for which only bibliographic title and year data were available. Coding from this tier provides no reliable basis for assessing reporting practices.

Results from the AB and MD tiers are reported only as secondary observations and are accompanied by explicit reliability caveats throughout. They are included in the analysis to provide a comprehensive picture of the corpus and to establish a context for the primary FT findings, not to support independent inferential claims.

The differential distribution of studies across evidence tiers reflects both the practical constraints of large-scale corpus analysis and the access limitations associated with subscription-gated academic publishing. The FT tier does not constitute a random sample of the corpus. Differential access to full text may introduce a selection effect if studies in certain disciplinary journals, publication types, or time periods are systematically over- or underrepresented. This limitation is acknowledged and addressed further in Section 5.

3.3 CODING SCHEME

A five-category coding scheme designated FR1 was developed to capture the full spectrum of AI model reporting practices encountered in the corpus. The scheme was designed to be mutually exclusive and collectively exhaustive: every record receives exactly one code, and the definitions are specified at a level of detail sufficient to resolve most borderline cases through explicit decision rules. The categories are as follows.

S2 (Specific model with version identifier). The study names a specific AI model or system and provides an explicit version or release identifier. The version indicator may take the form of a numeric release designation, a named version within a product family, or a temporal release tag that uniquely identifies the model checkpoint. This category represents the highest level of methodological transparency with respect to AI instrument specification and corresponds to the minimum standard that full reproducibility requires.

S1 (Specific named model, no version). The study names a specific AI model, product, or system but provides no version or release identifier. The reader can identify the AI system family but cannot determine which specific instance was active during data

collection. This category represents a necessary but incomplete level of reporting. It supports cross-study comparison at the level of AI system identity but not at the level required for precise replication.

G (Generic description). The study describes the AI component in vague or generic terms, such as “AI assistant,” “a machine learning algorithm,” or “an intelligent tutoring system,” without identifying any specific product or model. This category provides no basis for replication or cross-study comparison of the AI instrument and represents a qualitatively lower level of transparency than either S1 or S2.

NA (Not applicable). The study does not include an AI system that requires model identification. This category encompasses purely theoretical or conceptual papers, studies using a Wizard of Oz protocol in which a human confederate simulates AI behavior without any actual AI system being deployed, and secondary analysis or review articles that do not themselves deploy an AI system as a research instrument. NA records are excluded from the empirical subset used for proportional calculations about reporting quality.

NR (Not reported). The study appears to involve empirical use of an AI system based on its described design, stated methods, or reported outcomes, but provides no information sufficient to identify the model or system used. Unlike the G category, where some description is provided albeit vague, NR implies a gap between the apparent empirical activity and any methodological documentation of the AI component. This is the most acute category of transparency failure audited in this study.

3.4 APPLICATION PROCEDURE

The FR1 coding scheme was applied to each FT tier record by reviewing the complete article text with specific attention to the method section, instruments or materials section, stimuli or task description, procedure section, and any supplementary or appendix materials. For each record, the coder first determined whether the study involved an AI system whose identity is relevant to reproducibility, and if so, assigned the most precise code for which evidence was present in the text.

Borderline cases were resolved according to explicit decision rules. Studies that named a commercial or proprietary AI system within a specific product generation (for example, “GPT-3.5 Turbo”) were coded S2 even if the exact API version was unstated, on the grounds that a named generation constitutes a version indicator in the relevant sense. Studies that named only a brand family (for example, “ChatGPT” without generation

specification) were coded S1. Studies that described an AI tool by function without naming any product were coded G. Studies employing only rule-based or deterministic AI systems with stable and publicly documented behavioral specifications were coded S1 if the system name was stated, on the rationale that naming a stable system is equivalent in reproducibility terms to naming a versioned generative model.

For the AB tier, coding was applied to the information available in the abstract and bibliographic metadata. This limited information was sufficient in many cases to distinguish NA from non-NA records and to identify named systems when they were mentioned in the abstract, but was generally insufficient to distinguish reliably between S1, G, and NR. AB tier codes therefore carry substantially reduced inferential weight relative to FT tier codes and are treated as illustrative rather than definitive.

Intercoder reliability assessment was not conducted in this study, which used a single systematic coder throughout. This limitation is acknowledged and its implications for the findings are discussed in Section 5. The consistency of the coding scheme was supported by the application of explicit decision rules for all anticipated ambiguity types, and borderline decisions are documented and disclosed in the results where relevant.

3.5 ANALYTICAL APPROACH

The analysis is descriptive. For the primary FT tier, the distribution of codes is reported as counts and percentages of the total FT sample (n=67) and, separately, as percentages of the empirical subset obtained by excluding NA records (n=42). The empirical subset percentage is the more informative figure for assessing the quality of reporting among studies that actually deployed AI systems, because it removes from the denominator records for which model reporting is not applicable.

A temporal analysis examines the year-by-year distribution of codes within the FT empirical subset, tracking changes in the proportions of each code category across publication years from 2018 to 2026. This analysis addresses the second research question regarding temporal patterns and methodological maturity. No FT records were available for 2020, so this year is absent from the temporal analysis.

No inferential statistics are computed. Sample sizes within individual years are too small to support hypothesis testing with adequate power, and the analytical purpose of this paper is characterization and audit rather than causal inference. Proportions are reported to one decimal place throughout.

4 RESULTS AND DISCUSSION

4.1 OVERALL CODING DISTRIBUTION IN THE FULL-TEXT TIER

Table 1 presents the distribution of FR1 codes across the 67 full-text tier studies. The table reports counts and percentages with respect to both the full FT tier (n=67) and the empirical subset (n=42).

Table 1

FR1 coding distribution: full-text tier (n=67)

Code	Category	n (FT total)	% of FT total	% of empirical subset
S2	Specific model + version identifier	9	13.4	21.4
S1	Specific named model, no version	23	34.3	54.8
G	Generic description	6	9.0	14.3
NR	Not reported	4	6.0	9.5
NA	Not applicable	25	37.3	—
Total		**67**	**100.0**	—
Empirical subset (excl. NA)		**42**	—	**100.0**

Source: FR1 coding of full-text corpus records (2018 to 2026).

The most immediately notable feature of the distribution is that NA studies account for 37.3 percent of the full-text tier. These are studies for which AI model identification is not applicable because the research design does not involve the deployment of a real AI system. They include purely theoretical frameworks, Wizard of Oz protocols in which a human operator simulates AI behavior, and secondary reviews or meta-analyses. Their prevalence reflects the breadth of inquiry in the human-AI collaboration domain, which encompasses not only empirical trials of deployed AI systems but also substantial conceptual and synthetic work. Georganta and Ulfert (2024), for example, investigated team trust in human-AI teams using a Wizard of Oz design in which participants interacted with what they believed to be an AI teammate but which was in fact operated by a human confederate. Such studies contribute important knowledge about human cognition and attitudes in human-AI contexts without deploying a real AI system; accordingly, they are appropriately coded NA under the FR1 scheme.

Among the 42 studies in the empirical subset, 76.2 percent (n=32) identified the AI system by name, receiving codes of S2 or S1. This finding indicates that the majority of empirical researchers in this domain do attend to the identity of the AI system at some level of specificity. However, the distribution between S2 and S1 is highly asymmetric. Of the 32 studies that named the AI system, only 9 (28.1 percent) also provided a version or release identifier (S2). The remaining 23 (71.9 percent) named the system without version information (S1). This asymmetry suggests that awareness of the AI system as a methodological instrument is widespread, but that awareness of version-level specification as a documentation obligation has not yet been broadly internalized.

The G and NR categories together account for 23.8 percent of the empirical subset (n=10). These studies represent distinct but related forms of methodological opacity. G studies (14.3 percent) name no specific AI system; NR studies (9.5 percent) disclose no AI model information at all despite evidence of AI use. Both categories constitute reproducibility failures at the most fundamental level: a secondary analyst cannot identify, locate, or approximate the AI instrument from the published record.

4.2 THE EMPIRICAL SUBSET IN DETAIL

The empirical subset of 42 full-text studies constitutes the primary evidence base for this audit. These are studies that, by the FR1 coding determination, involved the deployment of an AI system as a component of the research design and for which full-text documentation was available.

The nine S2 studies are those that provide version-level specificity. Kirkby *et al.* (2025) specified “ChatGPT-3” as the AI system employed in their investigation of human-AI co-creation for creative brand voice, thereby anchoring the study to a particular model generation with a known capability profile. This approach exemplifies the S2 standard: the version indicator, however informal its expression, uniquely identifies the model generation and enables a secondary analyst to reconstruct the approximate capability profile of the instrument used. The S2 coding criterion required an explicit version or release indicator; studies that named “ChatGPT” or “GPT-4” without further qualification were assigned S1 regardless of publication recency or methodological sophistication in other respects.

The 23 S1 studies exemplify the most common reporting pattern in the empirical subset. They name a specific AI system but do not specify a version. In some cases, this reflects the use of AI systems with limited publicly documented version histories. Levine and

Williams (2018), for instance, described their study in terms of a named AI planning system (PIKE) without specifying a software version, a pattern common in robotic and autonomous systems research where implementations are often custom-built rather than commercially versioned. For such systems, the lack of a version identifier reflects an architectural reality of the research context rather than an oversight in documentation practice.

In more recent studies, however, S1 reporting typically reflects the use of commercially versioned large language model products for which version information is readily available but apparently not treated as a documentation requirement. Oppenheimer *et al.* (2025), for instance, named multiple AI systems across their study conditions, including ChatGPT, Gemini, and Claude, but did not specify which version of any system was used. This omission illustrates how a methodologically sophisticated study can still fall short of the S2 standard on what might appear to be a relatively minor procedural detail but which has significant consequences for the interpretability and reproducibility of the findings. The behavioral differences between, for example, Claude 3 Haiku and Claude 3.5 Sonnet, or between Gemini 1.0 Pro and Gemini 1.5 Pro, are substantial enough to constitute distinct AI instruments for research purposes.

The six G studies present a qualitatively different challenge. These studies describe AI involvement without naming any specific system, using language such as “AI assistant,” “machine learning classifier,” or “automated decision support.” Bossen and Pine (2023) examined human-AI collaboration among clinical documentation integrity specialists, using “Computer Assisted Coding” as the descriptor for the AI component. This phrasing names a function category rather than a specific system and was assessed as S1 on the basis that it implicitly identifies a well-defined class of clinical AI product deployed in the documented workplace setting, even though no specific product name or version appears in the text. This borderline decision illustrates the interpretive complexity inherent in auditing AI reporting practices and points to the value of explicit, named instrument documentation that would eliminate such ambiguities.

The four NR studies are discussed separately in Section 4.4.

4.3 TEMPORAL TRENDS IN VERSION-SPECIFIC REPORTING

Table 2 presents the year-by-year distribution of FR1 codes within the FT empirical subset across publication years from 2018 to 2026.

Table 2

Year-by-year FR1 code distribution: full-text empirical subset (n=42)

Year	FT total	Empirical	S2	S1	G	NR	Named (%)	S2 (%)
2018	1	1	0	1	0	0	100.0	0.0
2019	3	3	0	3	0	0	100.0	0.0
2021	5	2	0	1	1	0	50.0	0.0
2022	5	3	0	2	1	0	66.7	0.0
2023	7	2	0	1	0	1	50.0	0.0
2024	8	5	0	4	1	0	80.0	0.0
2025	16	11	4	3	3	1	63.6	36.4
2026	22	15	5	8	0	2	86.7	33.3
Total	**67**	**42**	**9**	**23**	**6**	**4**	**76.2**	**21.4**

Named (%) = (S2 + S1) / Empirical × 100. S2 (%) = S2 / Empirical × 100.
Source: FR1 coding of full-text corpus records (2018 to 2026). No FT records were available for 2020.

The temporal pattern is the most consequential single finding of this analysis. Across all empirical studies in the full-text tier published from 2018 through 2024, the S2 count is zero. Not a single study in this cohort, spanning six publication years and 16 empirically active records, provided a version-level identifier for its AI system. The version-specific reporting rate for this entire pre-2025 cohort is 0.0 percent.

This absence cannot be attributed to a low prevalence of AI systems with publicly documented version histories during this period. Commercial AI products with formal version documentation existed throughout the corpus timeframe, including robotics platforms, natural language processing APIs, and decision-support systems deployed in organizational settings. The specific case of large language models with publicly tracked version histories became highly salient from late 2022 with the public release of ChatGPT and from early 2023 with the release of GPT-4. Yet the 2023 and 2024 cohorts, comprising 7 empirically active studies between them, still record zero S2 studies. The absence of version reporting during this period therefore reflects not a lack of available version information but rather the absence of a community norm requiring its documentation.

The shift occurs at the 2025 boundary. Among the 11 empirically active studies in the 2025 cohort, 4 receive S2 codes, yielding a version-specific reporting rate of 36.4 percent. Among the 15 empirically active studies in the 2026 cohort, 5 receive S2 codes, yielding a rate of 33.3 percent. Pooled across 2025 and 2026, the S2 rate is 9 of 26 empirical studies, or 34.6 percent. In absolute terms this is a modest rate, but in relative terms the shift from 0.0

percent to 34.6 percent between adjacent cohort periods represents a structural change in community reporting practice.

The 2025 to 2026 cohort also shows improvement in the Named rate (S2+S1). The 2026 cohort achieves a Named rate of 86.7 percent, substantially higher than the 2023 cohort (50.0 percent), with the 2024 cohort (80.0 percent) representing an intermediate position. This improvement in basic model identification runs broadly parallel to the emergence of version-specific reporting, suggesting that the same normative shift may be driving both developments. However, the G rate in 2025 is 27.3 percent (3 of 11), the highest of any year in the corpus, indicating that even within the more recent cohort, a substantial minority of studies remains below the minimum naming threshold. The coexistence of high S2 rates and high G rates in 2025 suggests that the transition to more rigorous AI reporting is uneven across disciplinary communities and methodological traditions rather than uniformly distributed.

Two interpretations of the temporal shift from 0.0 percent to 34.6 percent at the 2025 boundary are plausible. The first is that the rapid proliferation of large language models from 2022 onward introduced a new category of AI instruments for which version documentation practices were initially absent and have only recently begun to crystallize. Early research using ChatGPT and related products may have treated these tools as relatively uniform consumer platforms rather than as versioned research instruments requiring formal specification, in part because the AI product landscape at that time was dominated by a small number of rapidly evolving systems whose version identity was not yet treated as scientifically consequential. As the behavioral differences between versions became more visible and as the research community's attention to reproducibility in AI research increased, version reporting may have begun to emerge as an expected practice.

The second interpretation is that the shift reflects broader maturation of the human-AI collaboration research community. Researchers who entered the field after methodological commentaries on AI reporting (COLLINS; MOONS, 2019; LIU *et al.*, 2020) were published and discussed may have internalized version documentation as a standard expectation in ways that earlier entrants did not. This generational or cohort effect in normative learning about AI documentation practices is consistent with the abruptness of the shift at the 2025 boundary and with the concentration of S2 records in the most recent cohort years. These two interpretations are not mutually exclusive and may both be contributing to the observed pattern.

4.4 NON-REPORTING CASES AND THEIR METHODOLOGICAL SIGNIFICANCE

Four studies in the full-text empirical subset were assigned the NR code, indicating that despite apparent empirical use of an AI system, no model identification was provided in the available documentation. Non-reporting represents the most acute transparency failure audited in this study. While a G code indicates that some description of AI involvement is present, NR indicates that even the existence of a specific AI instrument may have gone undisclosed to readers.

Two of the four NR studies are identifiable from the available documentation. Georganta *et al.* (2023), in a study of interactive technologies in team effectiveness contexts, employed AI-assisted analytical methods as part of the research process but did not disclose the identity or version of the AI tools used. Puerta-Beldarrain *et al.* (2025) conducted a comprehensive review of human-AI collaboration frameworks and similarly failed to identify the computational tools used in the synthesis and analysis phase. Two additional NR records from 2026 were identified in the coding process but cannot be definitively attributed from the available documentation at this stage.

The NR pattern in the two identified cases is methodologically notable because both are review-type studies rather than primary experiments in the conventional sense. This observation raises a question that the FR1 scheme addresses but that merits explicit discussion: is AI model identity a documentation obligation in review articles that use AI as an analytical aid, in the same way that it is an obligation in experimental studies where AI is the primary treatment or instrument? The position taken in this paper is that it is, for the following reason. Any AI-assisted analytical step, including AI-aided literature screening, text coding, synthesis, or summarization, introduces a computational process with a specific behavioral profile that influences the output. The identity and version of that AI tool is therefore a methodological variable that bears on the validity and reproducibility of the analytical conclusions. A review whose screening or synthesis was performed with the assistance of GPT-4 will not necessarily produce the same outcomes as one assisted by Claude 3 or Gemini 1.5 Pro; readers are entitled to this information to evaluate the review's conclusions and to reproduce its methodology.

The coding of these studies as NR is therefore consistent with the broader framework of methodological transparency advocated in this paper. It does not constitute a judgment about the quality of the studies' substantive contributions, which are independent of the AI documentation question, but it does identify a specific methodological gap that future

iterations of these scholarly lines of inquiry should address. As the use of AI assistance in literature reviews and research synthesis becomes more common, the documentation obligation for AI tools used in such processes will become increasingly central to research transparency standards.

4.5 OBSERVATIONS FROM SECONDARY EVIDENCE TIERS

The AB tier (n=272) and MD tier (n=99) are examined here as secondary observations with explicitly reduced reliability relative to the FT primary analysis. For abstract-only records, coding yields information primarily about whether AI involvement is apparent from the abstract and whether any named AI system is mentioned. Because method section details are unavailable, the S2 rate in the AB tier is expected to be negligible regardless of actual reporting quality in the full text, and the G and NR rates may substantially overestimate full-text non-compliance.

The AB tier data suggest that AI involvement is apparent from the abstract in a substantial fraction of the 272 records, and that named AI systems appear in a proportion of abstracts, though the base rate is lower than in the FT tier as expected. The AB distribution cannot be directly compared to the FT tier distribution because the two tiers represent different levels of documentation depth, not equivalent samples of the underlying literature. The primary value of the AB tier data is to confirm that the qualitative patterns observed in the FT tier (predominance of named over unnamed, low version-specific rate) are broadly consistent with the available abstract-level evidence.

The MD tier provides insufficient information for any meaningful observation about reporting practices and is not analyzed further in this paper. The 99 MD records represent a methodological blind spot in the corpus: their reporting practices cannot be assessed without access to at least abstract-level documentation. Taken together, the secondary tiers encompass 371 studies for which full-text primary analysis was not conducted. These studies may include a substantially different distribution of AI system types, disciplinary contexts, and reporting practices relative to the 67 FT studies. Differential access to full text, which tends to favor open-access publications and studies from certain disciplinary journal environments, may introduce a systematic divergence between the FT tier and the broader corpus. Future research with expanded full-text access could substantially enrich the current findings and test the generalizability of the FT results to the full corpus.

4.6 IMPLICATIONS FOR METHODOLOGICAL MATURITY

The overall picture that emerges from this audit is one of partial but incomplete methodological maturation. By 2026, the human-AI collaboration research literature has reached a state in which the majority of empirical full-text studies name the AI system used (76.2 percent in the empirical subset). This naming rate represents a substantive baseline of instrument awareness that compares favorably with what would be expected in a domain operating entirely without AI-specific reporting norms. However, only 21.4 percent of empirical studies provide a version-specific identifier, and 9.5 percent provide no model information at all.

By comparison, a methodologically mature field, defined operationally as one in which reproducibility with respect to AI instrument specification is reliably achievable, would exhibit S2 rates approaching 100 percent of empirical studies. The gap between the current 21.4 percent and this standard represents the primary methodological reproducibility risk in the human-AI collaboration literature at the time of writing. Given that virtually all contemporary human-AI collaboration research involves AI systems that are versioned, updated, and subject to behavioral change between releases, the current S2 rate implies that more than three-quarters of empirical full-text studies cannot be precisely reproduced from their methodological descriptions alone.

The temporal trend provides reason for cautious optimism. The emergence of a 34.6 percent S2 rate in the 2025 to 2026 cohort, from a baseline of 0.0 percent before 2025, indicates that normative pressure and community awareness are having a measurable effect. If this trajectory continues, the field could reach S2 rates substantially higher than 50 percent within the coming publication cycles. However, the near-equivalence of S2 rates in 2025 (36.4 percent) and 2026 (33.3 percent) rather than continued growth, and the persistence of high G rates in 2025, suggest that the improvement may be approaching an early plateau below the level that full reproducibility requires. Continued progress likely depends on the adoption of explicit author guidelines, as discussed in Section 4.7.

Comparison with adjacent fields reinforces this interpretation. In biomedical AI research, where CONSORT-AI and related standards have been in effect since 2020, compliance rates with AI model reporting requirements remain inconsistent even with formal journal-level enforcement mechanisms, as Liu *et al.* (2020) anticipated. The human-AI collaboration literature, which operates without such mechanisms, achieving a named rate of 76.2 percent suggests that the community takes instrument identification seriously at a

minimal level even absent formal mandates. The primary challenge is therefore not the absence of instrument awareness but the absence of norms that translate awareness into the specific practice of version documentation. The current findings also bear on the interpretation of any meta-analytic or comparative work conducted on this literature. A meta-analyst who attempts to pool estimates of AI impact on team performance across studies must treat AI instrument identity as a moderating variable. When version information is unavailable for more than 78 percent of the included empirical studies, this moderator variable cannot be adequately coded or controlled. Under such conditions, pooled estimates implicitly assume instrument equivalence among all studies using nominally similar AI systems, an assumption that the available documentation neither supports nor refutes but that introduces unquantifiable heterogeneity into any pooled estimate.

4.7 MINIMUM REPORTING RECOMMENDATIONS

Based on the findings of this audit, three elements are proposed as a minimum standard for AI model reporting in human-AI collaboration research. These elements are designed to be achievable for any empirical study regardless of the type of AI system involved and to provide the minimum information necessary to support replication, cross-study comparison, and secondary analysis.

The first element is the **model name**. Authors should identify the AI system by its specific product name or system designation, not by a functional category or generic description. “ChatGPT-4o” is an acceptable model name; “AI assistant” is not. “IBM Watson Natural Language Understanding v1.0” is acceptable; “a machine learning API” is not. Where a proprietary enterprise system is used under terms that preclude disclosure of the exact product name, authors should note this constraint explicitly and provide as much identifying information as the disclosure terms permit, rather than omitting the identification entirely. The presence of a non-disclosure constraint is itself methodologically relevant information.

The second element is the **version or release identifier**. Authors should specify the version of the AI model used, including version number, release date, or access date where no stable version number is available. For commercial API products, the API version endpoint and the access date should both be recorded, as API endpoints may map to different underlying model checkpoints at different points in time. For locally deployed open-source models, the model checkpoint identifier and configuration should be stated. For consumer-facing products where a version setting is available (for example, the choice between GPT-

4o and GPT-4 within the ChatGPT interface), that setting should be documented. For products where the version is not controllable or visible to the user, the access date serves as a partial proxy and should be recorded even if it cannot uniquely identify the model checkpoint. The rationale for this requirement is not merely formal compliance but substantive reproducibility: without version information, a reader cannot evaluate whether the AI instrument in the reported study is the same computational entity as any subsequent comparison or replication.

The third element is the **access method**. Authors should specify whether the AI system was accessed through a consumer interface, through an API call, through a locally deployed instance, or through an institutional deployment. This information bears on whether the model was in its default configuration, whether any system prompts or configuration parameters were active, and whether the access conditions can be reproduced by a subsequent researcher. A study that used “GPT-4o via the OpenAI API with a custom system prompt” employs a different methodological configuration from one that used “GPT-4o via the ChatGPT consumer interface at default settings,” even if the underlying model is nominally the same. For studies involving prompting protocols, the prompt structure or system message should additionally be included in the method section or as a supplementary appendix.

These three elements constitute a reporting floor rather than a ceiling. Researchers who can provide additional detail, including temperature settings, context window parameters, fine-tuning information, and evaluation benchmarks, should do so, as this information supports more complete methodological evaluation. But the three elements proposed here represent the minimum necessary to establish AI instrument identity in a manner that supports replication and secondary analysis. Implementation of these standards does not require institutional infrastructure or peer-reviewed consensus processes before authors can begin applying them. They can be adopted unilaterally in any empirical study at essentially no cost, since the relevant information is available to the researcher at the time of data collection and need only be recorded and reported. The primary barrier is normative rather than practical: the absence of a community expectation that such documentation is required. This paper aims to contribute to the establishment of that expectation.

5 CONCLUSION

This paper has presented a systematic audit of AI model and version reporting practices across 439 studies on human-AI team collaboration, with primary analysis focused on the 67 full-text studies available for complete review. Four principal findings emerge.

First, model naming is the majority practice among empirically active studies. Of the 42 studies that deployed AI systems in the full-text tier, 76.2 percent identified the AI system by name. This finding indicates that the field has developed a general awareness of AI identity as a methodological consideration, at least at the level of instrument naming. It is a more encouraging baseline than might have been expected given the absence of formal reporting mandates in this literature.

Second, version-specific reporting is rare. Only 21.4 percent of empirically active full-text studies provided a version or release identifier. The gap between naming (76.2 percent) and versioning (21.4 percent) is large and represents the primary methodological transparency deficit identified by this audit. Given the well-documented behavioral variability between versions of nominally identical AI models, this gap constitutes a substantial and underappreciated reproducibility risk. Studies that report using “ChatGPT” without version specification cannot be precisely replicated by a subsequent researcher, because the ChatGPT product family encompasses a range of model generations with materially distinct capabilities.

Third, version-specific reporting is a recent and not yet universal phenomenon. All nine S2 studies in the full-text tier are from 2025 or 2026. No study published before 2025 in the full-text empirical subset provided a version identifier, yielding a version-specific reporting rate of 0.0 percent for the 2018 to 2024 period and 34.6 percent for the 2025 to 2026 period. This abrupt shift suggests that community norms regarding AI version reporting began to crystallize around 2025, a development that may be attributable to growing awareness of AI capability differentiation across model versions and to the broader visibility of reproducibility concerns in AI research more generally.

Fourth, a small but meaningful minority of empirical studies disclosed no AI model information whatsoever. The NR rate of 9.5 percent in the empirical subset indicates that even the minimum threshold of naming has not been universally reached. The identification of NR cases in review-type studies, where AI-assisted analytical steps are not documented, suggests that the AI documentation obligation extends beyond experimental designs to any research process in which AI tools make methodologically consequential contributions.

Several limitations of this study should be acknowledged. The primary analysis is restricted to 67 full-text studies, a subset that may not be fully representative of the 439-study corpus. Differential access to full text may introduce a selection effect that limits the generalizability of the FT findings to the broader corpus. The coding scheme was applied by a single coder, and the absence of an intercoder reliability assessment means that the stability of borderline decisions cannot be formally verified. Individual year cohorts in the temporal analysis are small, limiting the precision of year-by-year estimates. Future work should expand full-text coverage to encompass a larger fraction of the corpus, apply intercoder reliability protocols to key coding categories, and conduct longitudinal tracking of S2 rates in subsequent publication cohorts to determine whether the improvement observed from 2025 onward is sustained.

The central practical implication of this work is straightforward. Authors, reviewers, and journal editors working in the human-AI collaboration domain should treat the three-element minimum reporting standard proposed in Section 4.7 (model name, version or release identifier, access method) as a baseline expectation for all empirical work. This standard is achievable at no cost to researchers who record the relevant information at the time of data collection, and its widespread adoption would substantially reduce the reproducibility risk identified in this audit. The human-AI collaboration research field is at a point of genuine methodological transition: version-specific reporting has emerged as a practice but has not yet become a norm. The findings reported here provide an empirical basis for accelerating that transition.

REFERENCES

BAKER, M. 1,500 scientists lift the lid on reproducibility. Nature, v. 533, n. 7604, p. 452 to 454, 2016.

BOSSEN, C.; PINE, K. H. Batman and Robin in Healthcare Knowledge Work: Human-AI Collaboration by Clinical Documentation Integrity Specialists. ACM Transactions on Computer-Human Interaction, v. 30, n. 2, p. 1 to 29, 2023. DOI: 10.1145/3569892.

COLLINS, G. S.; MOONS, K. G. M. Reporting of artificial intelligence prediction models. The Lancet, v. 393, n. 10181, p. 1577 to 1579, 2019.

DELLERMANN, D. *et al.* Hybrid intelligence. Business & Information Systems Engineering, v. 61, n. 5, p. 637 to 643, 2019. DOI: 10.1007/s12599-019-00595-2.

GEORGANTA, E.; PEUS, C.; NIESS, J. Interactive technologies through the lens of team effectiveness: an interdisciplinary systematic literature review. European Journal of Work and Organizational Psychology, v. 33, n. 2, p. 172 to 187, 2023. DOI: 10.1080/1359432X.2023.2178904.

GEORGANTA, E.; ULFERT, A. Would you trust an AI team member? Team trust in human-AI teams. Journal of Occupational and Organizational Psychology, v. 97, n. 3, p. 1212 to 1241, 2024. DOI: 10.1111/joop.12504.

IOANNIDIS, J. P. A. Why most published research findings are false. PLOS Medicine, v. 2, n. 8, p. e124, 2005.

KIRKBY, A.; BAUMGARTH, C.; HENSELER, J. Welcome, new brand colleague! A conceptual framework for efficient and effective human-AI co-creation for creative brand voice. Journal of Brand Management, v. 32, n. 5, p. 385 to 399, 2025. DOI: 10.1057/s41262-025-00387-y.

LEVINE, S. J.; WILLIAMS, B. C. Watching and Acting Together: Concurrent Plan Recognition and Adaptation for Human-Robot Teams. Journal of Artificial Intelligence Research, v. 63, p. 281 to 359, 2018. DOI: 10.1613/jair.1.11243.

LIU, X. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. The Lancet Digital Health, v. 2, n. 10, p. e537 to e548, 2020.

McNEESE, N. J. *et al.* Teaming with a synthetic teammate: Insights into human-autonomy teaming. Human Factors, v. 60, n. 2, p. 262 to 273, 2018.

MOHER, D. *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Medicine, v. 6, n. 7, p. e1000097, 2009.

OPEN SCIENCE COLLABORATION. Estimating the replicability of psychological science. Science, v. 349, n. 6251, p. aac4716, 2015.

OPPENHEIMER, D. M.; CASH, T. N.; CONNELL PENSKY, A. E. You've Got AI Friend in Me: LLMs as Collaborative Learning Partners. International Journal of Artificial Intelligence in Education, v. 35, n. 6, p. 3896 to 3921, 2025. DOI: 10.1007/s40593-025-00521-6.

PUERTA-BELDARRAIN, M. *et al.* A Multifaceted Vision of the Human-AI Collaboration: A Comprehensive Review. IEEE Access, v. 13, p. 29375 to 29405, 2025. DOI: 10.1109/ACCESS.2025.3536095.

SEEBER, I. *et al.* Machines as teammates: A research agenda on AI in team collaboration. Information & Management, v. 57, n. 2, p. 103174, 2020. DOI: 10.1016/j.im.2019.103174.

SOWA, K.; PRZEGALINSKA, A. From Expert Systems to Generative Artificial Experts: A New Concept for Human-AI Collaboration in Knowledge Work. Journal of Artificial Intelligence Research, v. 82, p. 2101 to 2124, 2025. DOI: 10.1613/jair.1.17175.

APPENDIX

Table A1

FR1 Coding Scheme: Category Definitions and Decision Criteria

Code	Label	Definition	Inclusion criteria	Example
S2	Specific model + version	Study names a specific AI model and provides an explicit version or release identifier	Version number, named release within a family, API version, temporal access identifier	"ChatGPT-3"; "GPT-4o (version 2024-05-13)"; "Claude 3.5 Sonnet"
S1	Specific named, no version	Study names a specific AI model or product without any version information	Product or system name sufficient to identify the instrument family	"ChatGPT"; "GPT-4"; "IBM Watson"; named custom system (e.g. PIKE)
G	Generic description	Study describes AI involvement in vague or functional terms without naming any specific model or product	AI function category, generic capability descriptor, unnamed AI type	"AI assistant"; "machine learning algorithm"; "intelligent agent"; "automated decision aid"
NA	Not applicable	Study does not deploy a real AI system requiring model identification	Purely theoretical or conceptual studies; Wizard of Oz designs; secondary reviews or meta-analyses	Conceptual frameworks; WoZ experimental designs; literature syntheses without AI tool deployment
NR	Not reported	Study shows evidence of empirical AI deployment but provides no information sufficient to identify the model	AI use is apparent from design or outcomes description but no system is named or described	Studies describing AI-mediated tasks or AI-assisted analyses with no model identification

Source: FR1 coding scheme developed for this analysis.