Research Square

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

An Error Taxonomy for AI-Copilot Assisted Spreadsheet and Analytics Work

Md. Anamul Hasan

developer.anamulhasan@gmail.com

national university bangladesh https://orcid.org/0009-0005-9925-0855

Md.Abu sayeed

East West University https://orcid.org/0009-0003-0698-2607

Research Article

Keywords: spreadsheet errors; error taxonomy; AI copilots; large language models; automation bias; human review; analytics quality assurance; controlled experiment

Posted Date: October 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11275396/v1

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

Additional Declarations:
The authors declare no competing interests.

Ethics Approval: "This study did not involve human participants, animal subjects, or sensitive personal data. Therefore, ethics committee / Institutional Review Board (IRB) approval was not required for this work." Karina confirmed: "This is a protocol, so no IRB needed.

RESEARCH PAPER · FULL MANUSCRIPT

An Error Taxonomy for AI-Copilot Assisted Spreadsheet and Analytics Work

A Controlled Comparative Study of Human and AI Introduced Errors

Author: Md. Anamul Hasan

Designation : Data Analyst | Statistics & Data Science Background

Orchid ID : 0009-0005-9925-0855

Co-author : Md.Abu sayeed | Founder of Cuteadmoa | AI model and system developer

Orchid ID : 0009-0003-0698-2607

Document type : full research paper · 5 figures · 4 tables

ABSTRACT

BACKGROUND : Decades of research beginning with Panko established that most operational spreadsheets contain material errors, classified under a stable taxonomy. That literature assumes a human authored every formula. Since roughly 2023, AI copilots have taken over a meaningful share of that work, and benchmarks show they fail on a large fraction of realistic tasks frontier models reach only 30–45% execution accuracy on a 912 task real world benchmark.

THE GAP : Benchmarks score the correctness of generated artefacts but never measure whether a human reviewer would notice an error. The human-factors literature documents under-detection of errors in polished automated output in other domains. No study connects the two: a controlled comparison using one taxonomy to classify AI-introduced and human-introduced spreadsheet errors and to measure how often each is caught.

METHOD : A controlled, matched task comparison with an embedded detection experiment. Two matched groups of 20–30 working analysts complete identical tasks, one building manually and one using a copilot with prompts logged. All artefacts are audited blind under a taxonomy adapted from the Panko Halverson classification and extended for AI era failure modes. An independent blind panel then reviews a mixed set under normal time pressured conditions, without being told errors are present, and rates confidence in each artifact.

CONTRIBUTION : A taxonomy built for the post-2023 authoring environment while remaining comparable with the classical literature, and the first measurement of the detectability asymmetry between AI introduced and severity matched human-introduced spreadsheet errors the quantity that determines whether existing review practice is adequate.

Keywords : spreadsheet errors; error taxonomy; AI copilots; large language models; automation bias; human review; analytics quality assurance; controlled experiment

1. Introduction

Spreadsheets remain the most widely used analytical tool in business, and their error rates have been studied for longer than most people realise. Audits going back to the 1990s [22], and larger studies since, keep landing

on roughly the same conclusion: the great majority of operational spreadsheets contain at least one material error, with a cell level formula error rate around 5% treated as a long-standing benchmark [2]. A 2024 systematic review put the figure for decision making spreadsheets far higher, around 94% containing critical errors, and observed that quality assurance still focuses overwhelmingly on catching mistakes after the fact rather than preventing them at the point of construction [4].

That entire literature describes a world that no longer quite exists. Since roughly 2023, a growing share of formulas, pivot logic, and whole analytical workflows are generated from natural language prompts by AI copilots embedded directly in spreadsheet and business intelligence tools [5], [18], [19]. The early evidence on how well this works is not comforting. On SpreadsheetBench, a 912 task benchmark drawn from real business and financial modelling scenarios, state-of-the-art models reach only 30–45% execution accuracy [8]. The FLARE evaluation found 35–42% of LLM-generated formulas carrying range-mismatch or boundary errors that pass syntax validation while silently miscalculating the metric [10]. SpreadsheetLLM's own authors report baseline models at 40–60% accuracy on complex spreadsheet tasks [7]. Meanwhile survey evidence shows 96% of developers do not fully trust AI generated logic, yet only 48% perform structured verification of it [16].

Those two facts, high error rates and low verification combine into a specific worry that practitioners have begun to voice: that AI-generated spreadsheet content may be harder to catch than an old-fashioned manual mistake, because it looks right to a reviewer who did not derive the logic themselves [19]. Practitioner guidance now recommends mandatory three-way checks on every AI assisted model [19], and industry analysis describes a measurable shift toward reviewing rather than writing, with up to 52% of AI-generated models approved without line-by-line verification [17]. The claim is plausible, consequential, and as far as the literature goes untested.

This paper specifies the study that would test it. It is organised as follows. Section 2 reviews the classical spreadsheet error literature, the arrival of copilots, the early evidence on AI generated spreadsheet faults, and the human-factors literature on reviewing automation. Section 3 states the research objectives. Section 4 details the methodology: the adapted error taxonomy, the matched-task design, the audit procedure, the detection sub-study, the calibration component, and the analysis plan. Sections 5 to 8 set out expected contributions, significance, limitations, and conclusion.

2. Background and Literature Review

2.1 Classical spreadsheet-error research and its assumptions

The spreadsheet-error field is unusually mature for an applied computing topic. Panko's foundational work established both the prevalence of errors and, more durably, a taxonomy for classifying them: formula errors, data errors, structural errors, and comprehension or interpretation errors, later refined in the Panko Halverson taxonomy [1]. Powell, Baker and Lawson's audit of operational spreadsheets supplied the large-sample evidence, finding material errors in the great majority of real business workbooks and establishing the ~5% cell-level formula error rate as a reference point [2]. Irons' controlled task study demonstrated that domain-referent errors where the formula computes something other than the intended business quantity could be generated and measured under laboratory conditions [3].

Every one of these studies shares a single structural assumption: a human author builds each formula deliberately, and the error is therefore a human cognitive artefact. The taxonomy is built to classify human slips, a mistyped range, a copy-paste that dragged the wrong reference, a SUM where an AVERAGE was meant. That assumption is what the last two years have invalidated, and the taxonomy has not been revised to reflect it.

2.2 The arrival of AI copilots in analytical work

The scale of the change is easy to underestimate because it happened inside tools people already used. Formula generation, pivot construction, natural language querying and even whole workbook reasoning are now mediated by language models in mainstream spreadsheet and BI environments [5], [18]. The practitioner literature reflects the shift: FM Magazine's guidance on AI-assisted financial modelling notes that generative systems lack spreadsheet spatial awareness and recommends structural checks on every AI assisted model [19]; CFO.com reports 79% of FP&A teams adopting AI in modelling workflows while professionals spend roughly 13 hours a week auditing AI output [18]. The tooling is in place, the adoption has happened, and the verification burden has quietly moved to the reviewer.

2.3 Early evidence on AI-generated spreadsheet errors

A small but consistent body of evidence has begun to characterise how these systems fail. The earliest controlled test found ChatGPT produced consistently correct output only when the prompt left no room for inference, with accuracy degrading sharply as interpretation was required [6]. Recent benchmarks sharpen the picture: SpreadsheetBench shows frontier models at 30–45% execution accuracy on realistic workbook manipulation [8]; a formula-repair benchmark finds models fixing explicit syntax errors 75–80% of the time while missing silent semantic or logic errors in 40–50% of cases [9]; FLARE identifies range mismatch and off by one faults that survive syntax checks [10]; SODBench finds baseline enterprise workbooks exhibiting 88–94% error rates in multi-sheet workbooks, with 30%+ carrying unmapped formula chaining errors [11]. Formula R1 reports baseline models producing structure, range or calculation errors on 48.3% of complex multi-step queries [12]. Spreadsheet RL shows interactive execution feedback lifting agent success from 41.2% to 76.8%, which is instructive precisely because it implies single-turn copilot generation fails on the majority of complex tasks [13]. FRTR reaches 87% accuracy with retrieval and vision augmentation against 52% unaugmented, leaving a residual 13% failure rate concentrated on implicit table relationships [14].

Two things stand out across this evidence. First, the errors that persist are disproportionately *silent* range mismatches, boundary conditions, misinterpreted implicit relationships the kind that produce a plausible number rather than an obvious breakage. Second, every one of these benchmarks scores raw correctness of the generated artefact. None measures whether a human reviewer, working under normal conditions, would notice that the output was wrong. That distinction is the entire question for practice, because an error nobody catches is not equivalent to an error that gets caught, even when both are counted identically in an accuracy metric.

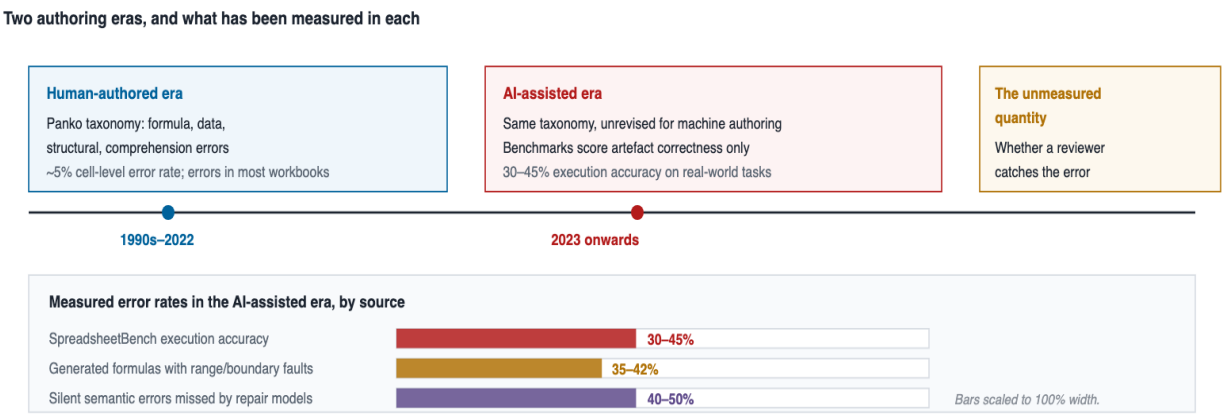


Figure 1. The two authoring eras and what has been measured in each. Classical error rates are established; AI-era error rates are quantified by benchmarks; the detection question in the third column remains unmeasured.

2.4 Human review of automation: the human factors literature

The human-factors literature has studied exactly this asymmetry in other domains for decades. Complacency and bias in the use of automation are well documented, with automation bias producing both omission errors [21] (failing to notice a problem the automation did not flag) and commission errors (acting on an incorrect automated recommendation) [20]. The pattern generalises across aviation, process control, and clinical decision support.

Recent work has begun to extend it to AI assisted knowledge work. A participatory design study of trust-calibrated code review found human error-detection rates dropping 28% when reviewing clean-looking AI output, a direct measurement of the automation-bias effect in a software context [15]. Industry analysis of the same shift describes 'LGTM rubber-stamping' approval without line-by-line verification affecting up to 52% of AI generated models, and explicitly draws the comparison to vigilance failures in aviation and radiology [17]. Sonar's State of Code survey finds the trust-verification gap widening: 96% of developers distrust AI generated logic while fewer than half verify it [16]. If the same dynamic operates in spreadsheets and the FLARE and SODBench findings suggest the failure modes are precisely the silent kind that automation bias would suppress then AI assisted analytical work may be systematically less reliable than its error rate alone implies.

2.5 The gap

The literature establishes three things firmly. Traditional spreadsheet errors are pervasive, well classified, and measured with mature methods [1], [2], [3], [4]. AI copilots generate spreadsheet content at error rates that benchmarks now quantify, with residual failures concentrated in silent semantic categories [6]–[14]. Human reviewers demonstrably under-detect errors in polished automated output in adjacent domains [15], [17], [20]. What does not exist is the study connecting them: a controlled, practitioner-based comparison using a single taxonomy to classify both AI-introduced and human-introduced errors *and* to measure how often each is caught by an independent reviewer under normal review conditions. The benchmarks score correctness; the human-factors work studies review behaviour without a controlled error taxonomy; the classical literature predates AI authoring entirely. That three-way gap is what this study fills.

3. Research Objectives and Questions

The objective is to determine whether AI copilot assisted spreadsheet and analytics work differs from manual work in the rate, the category, and most importantly the detectability of its errors, using a taxonomy that covers both authoring modes. Table 1 maps the four research questions to their instruments and analyses.

RQ1 — Error rate. Do analysts using an AI copilot produce a different overall error rate than analysts working manually on matched tasks? This establishes whether the copilot changes reliability at all, before asking how.

RQ2 — Error category. Do AI-assisted and manually built errors fall into different categories under an adapted version of the established taxonomy (formula/logic, data, structural, comprehension/interpretation)? A shift in category mix would explain any change in detectability and would itself be a contribution to the taxonomy literature.

RQ3 — Detectability (primary contribution). Are AI introduced errors caught less often by an independent human reviewer performing a normal review pass, compared with matched manually introduced errors of similar severity? This is the question practice actually cares about and the one no benchmark answers.

RQ4 — Calibration. Does reviewer confidence in an artifact track its actual accuracy differently depending on whether it was AI assisted or built manually in other words, is there an AI specific trust miscalibration effect? This tests whether the mechanism behind any detectability gap is misjudged confidence rather than inattention.

Table 1. Research questions mapped to instruments and analysis.

RQ	Question in brief	Instrument	Analysis
RQ1	Does copilot use change the overall error rate?	Blind audit of all artefacts produced under both conditions	Negative binomial regression on error counts
RQ2	Do the two conditions produce different error categories?	Same audit, classified under the adapted taxonomy	Chi square on category composition; severity mix
RQ3	Are AI introduced errors less detectable?	Independent blind reviewer panel under time pressure	Logistic regression on detection, error level random effects
RQ4	Does reviewer confidence track accuracy differently?	Post-review confidence ratings	Confidence accuracy association compared across conditions

Table 2. Evidence base: benchmark and industry sources informing the study.

Source	Type	Key figure	Status
Powell, Baker & Lawson [2]	Peer-reviewed audit	~5% cell level formula error rate; errors in most operational spreadsheets	Established reference point
2024 systematic review [4]	Peer-reviewed review	~94% of decision-making spreadsheets contain critical errors	Reported via Frontiers of Computer Science
SpreadsheetBench [8]	NeurIPS 2024 benchmark	912 real-world tasks; SOTA models 30–45% execution accuracy	Published benchmark
FLARE (EuSpRIG 2025) [10]	Peer-reviewed framework	35–42% of generated formulas contain range/boundary errors passing syntax checks	Published
Formula repair benchmark [9]	Benchmark	Syntax fixes 75–80%; silent semantic errors missed in 40–50% of cases	Published
SODBench [11]	Benchmark	88–94% error rates in multi-sheet workbooks; 30%+ unmapped chaining errors	Published
Spreadsheet-RL [13]	Benchmark	Agent success 41.2% → 76.8% with execution feedback	Published
FRTR [14]	Preprint	87% with RAG+vision vs. 52% unaugmented; residual 13% on implicit relations	Preprint
Trust-calibrated code review [15]	Peer-reviewed study	Detection rate falls 28% on clean-looking AI output	Published
Sonar State of Code [16]	Industry survey	96% distrust AI logic; only 48% verify structurally	Industry report
Atomic Robot analysis [17]	Industry report	Up to 52% of AI models approved without line-by-line review	Industry report
CFO.com / FMI [18]	Industry reporting	79% of FP&A teams use AI in modelling; ~13 hrs/week auditing AI output	Industry report

3.1 Hypotheses

Each research question is expressed as a directional hypothesis below, fixed before analysis. The primary hypothesis is H3; H1, H2 and H4 are supporting results that make the primary finding interpretable.

H1 (error rate). The copilot condition produces a different overall error rate than the manual condition on matched tasks. The direction is not pre-specified, because the benchmark evidence permits both readings:

copilots may reduce mechanical slips while introducing semantic ones, and the net effect is an empirical question.

H2 (error category). Errors in the copilot condition concentrate in the formula/logic and comprehension categories, and disproportionately in the silent severity class errors that compute a plausible number rather than breaking visibly. This follows directly from the benchmark evidence, which consistently finds residual failures in range, boundary and implicit relationship reasoning rather than in syntax.

H3 (detectability — primary). AI introduced errors are detected at a lower rate by an independent reviewer performing a normal review pass than manually introduced errors matched on severity and category. This is the study's central claim. It predicts a specific, measurable asymmetry rather than a general claim that AI output is unreliable, and it is falsifiable: comparable detection rates across conditions would refute it.

H4 (calibration). The association between reviewer confidence and actual artifact accuracy is weaker in the copilot condition reviewers are approximately as confident but less accurate. If supported alongside H3, this identifies misjudged confidence rather than simple inattention as the mechanism, which is the difference between an intervention targeting attention and one targeting trust calibration.

Interaction expectations are stated as secondary. H2 predicts the category shift and H3 predicts the detection gap; if the detection gap is explained entirely by the category shift, that is a coherent and reportable result. The analysis therefore tests the detection gap both unconditionally and conditional on category and severity, so the two explanations can be distinguished rather than conflated.

4. Proposed Methodology

4.1 Design

The design is a controlled, matched-task comparison with an embedded error-detection experiment layered over it. Its defining feature is that it treats the *reviewer* as a second experimental subject: the study measures not only what errors are produced under each authoring condition, but what happens to those errors downstream when someone reviews the work. Figure 1 shows the overall design.

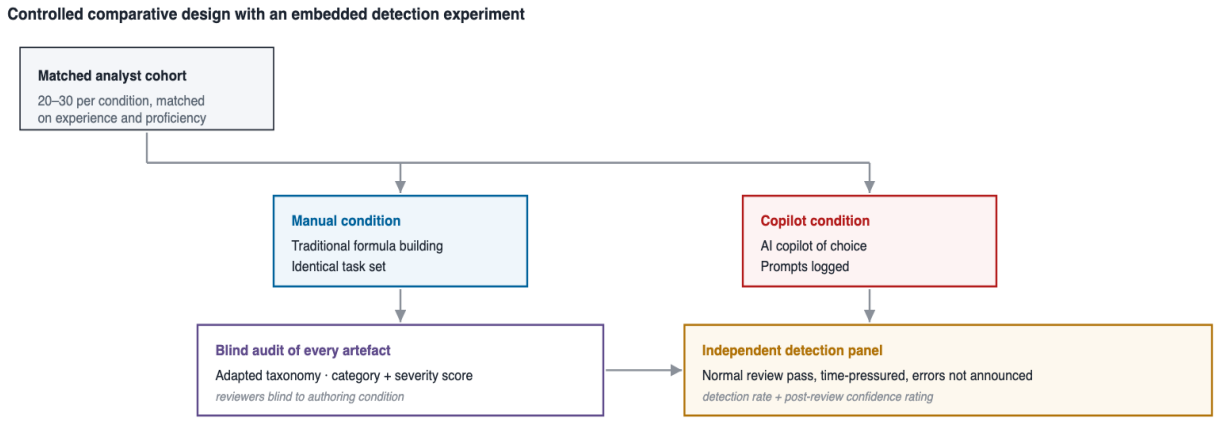


Figure 3. Controlled comparative design. The reviewer is treated as a second experimental subject: the study measures both what errors are produced under each condition and what happens to them downstream.

4.2 The adapted error taxonomy

Audit classification uses the established four-category structure formula/logic, data, structural, and comprehension/interpretation [1], [2] deliberately preserved so that results remain comparable with the classical literature. Each category is extended with manifestations specific to AI assisted authoring, derived from the failure modes documented in the benchmark studies [6]–[14]. The adaptation is detailed in Table 3 and summarised in Figure 2. The critical addition is a severity dimension distinguishing *loud* errors (which break visibly a #REF!, a total that obviously disagrees) from *silent* errors (which compute a plausible but wrong number). Detectability is expected to depend far more on this dimension than on the category, which is why it is scored separately.

Table 3. The adapted error taxonomy: classical categories extended for AI-assisted authoring.

Category	Classical definition	AI-era manifestation	Severity
Formula / logic	Formula computes something other than the intended business quantity	Plausible-but-wrong aggregation; off by one range boundaries; correct-looking nested logic that encodes the wrong rule	Usually silent
Data	Wrong or mis-scoped input data selected	Implicit table relationship misread; wrong column inferred from ambiguous prompt; stale range not extended	Silent to loud
Structural	Workbook organisation causes incorrect propagation	Formula chaining across sheets that does not update; duplicated logic diverging after an edit; unlinked summary	Usually loud, sometimes silent
Comprehension	Author misunderstood the task or the domain rule	Ambiguous instruction resolved wrongly without flagging; assumption invented rather than asked about	Silent

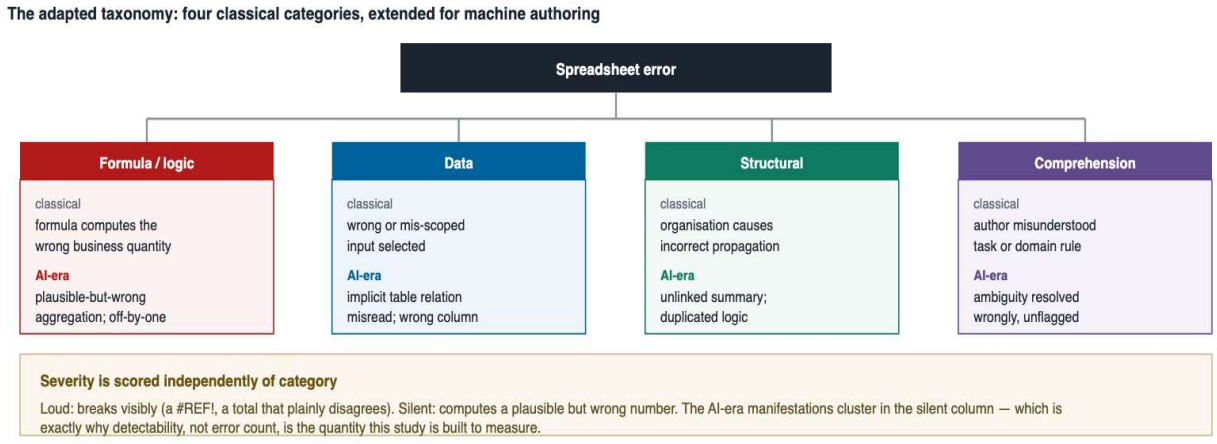


Figure 2. The adapted taxonomy. Classical categories are preserved for comparability; each is extended with AI-era manifestations. Severity is scored independently because detectability is expected to depend on visibility rather than category.

4.3 Participants and task suite

Two matched groups of working data analysts, 20–30 per group, are recruited through professional networks and matched on experience, spreadsheet proficiency, and domain familiarity. Both groups complete an identical task set: a multi criteria lookup, a pivot-based summary report, and a conditional budget calculation. Tasks were selected because each exercises a different failure surface reference construction, aggregation logic, and conditional reasoning respectively and because their outputs are objectively checkable against a reference solution. The manual group uses traditional formula building only; the copilot group works with an AI copilot of their choice, with all prompts logged for later analysis. Tool and model versions are recorded exactly, since copilot capability changes quickly and the findings are time bound to the versions tested.

4.4 Audit procedure

Every resulting workbook is audited by trained reviewers blind to authoring conditions, using the adapted taxonomy. Each artefact receives an error count, a category breakdown, and a severity classification. To make the comparison fair, errors are also matched across conditions on severity and category where possible, so that RQ3 tests detectability at comparable error quality rather than at whatever mix each condition happened to produce. Inter rater reliability is established on a common subset before the main audit proceeds.

4.5 Detection-difficulty sub-study

This is the study's novel core and the reason for the design's second layer. A separate panel of independent reviewer analysts, blind to which artefacts are AI assisted, reviews a mixed set of outputs under normal time-pressured conditions, flagging any errors they find. Crucially, reviewers are *not* told that every file contains an error, so that false-positive behaviour is observable and the review task resembles real work rather than a hunt. Detection rates for AI introduced errors are compared with detection rates for matched manually introduced errors. The protocol is shown in Figure 4.

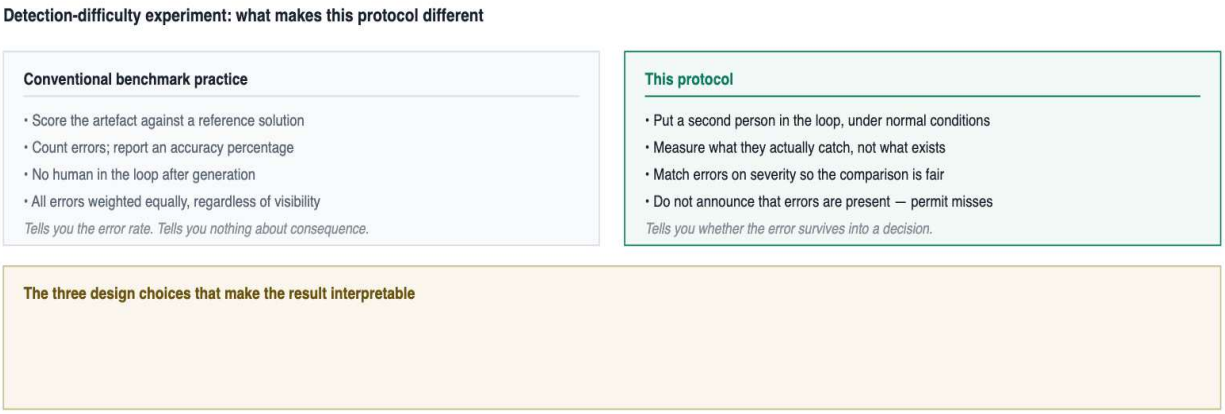


Figure 4. Detection-difficulty protocol. Three design choices make the result interpretable: blind assignment, unannounced errors, and severity-matched comparison.

4.6 Calibration component

After each review, participants rate their confidence in the artefact's correctness. The relationship between stated confidence and actual accuracy is then compared across the two authoring conditions. A positive association that is weaker in the AI condition reviewers equally confident but less accurate would support the

trust miscalibration mechanism proposed in RQ4, and would connect these findings to the broader human factors literature on automation bias [15], [20].

4.7 Analysis plan

Error counts and category distributions are compared between conditions by chi square tests for composition and negative binomial regression for rates, which is appropriate for over dispersed count data. Detection rates are compared by logistic regression with error-level random effects, controlling for participant experience and time on task. Calibration is assessed by comparing confidence accuracy association across conditions. Full specifications, including those that do not support the hypotheses, are reported. The analysis pipeline is summarised in Figure 5 and Table 4.

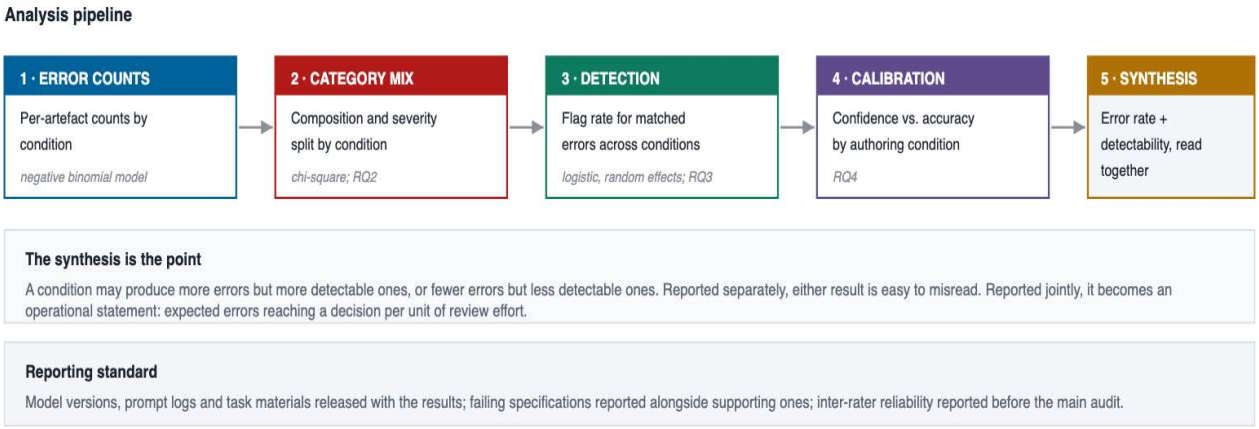


Figure 5. Analysis pipeline. Error rate and detectability are synthesised rather than reported separately, because either alone is easy to misread operationally.

Table 4. Study conditions, tasks and measures.

Element	Manual condition	Copilot condition	Compared on
Authoring	Traditional formula building only	AI copilot of choice, prompts logged	—
Tasks	Multi-criteria lookup; pivot summary; conditional budget	Identical task set	Matched difficulty
Primary measure	Error count per artefact	Error count per artefact	RQ1
Secondary measure	Category and severity mix	Category and severity mix	RQ2
Detection measure	Blind reviewer flag rate	Blind reviewer flag rate	RQ3
Calibration measure	Reviewer confidence vs. accuracy	Reviewer confidence vs. accuracy	RQ4

4.8 Validity and ethics

Three threats are addressed by design. Participant AI literacy and prompting skill confound the copilot condition, so both are measured and reported, and prompt logs allow a sensitivity analysis restricted to competent prompting. Task artificiality limits generalisation, which is why the tasks mirror real work and why findings are framed as a controlled baseline rather than a field estimate. Tool version drift dates the findings, which is why exact versions are recorded and reported in the title block of any published result. Participants

are volunteers recruited with informed consent, artefacts contain no confidential business data, and reviewer panels are debriefed after the detection sub study since the design necessarily withholds the presence of errors during it.

5. Expected Contributions

The study is expected to contribute in four ways. It produces the first controlled comparison of AI-assisted and manual spreadsheet work by working analysts rather than by researchers testing models in isolation, a distinction that matters because prompts, tool choice and workflow differ in practice. It delivers an error taxonomy adapted for the post-2023 authoring environment while remaining compatible with the classical categories, so results can be read against thirty years of prior findings. It supplies the first measurement of the detectability asymmetry between AI introduced and human-introduced spreadsheet errors, which is the quantity that determines whether existing review practice is adequate. And it provides a calibration result linking reviewer confidence to accuracy across authoring conditions, connecting spreadsheet practice to the wider automation bias literature.

6. Significance

Every finance, operations and analytics function is adopting copilots for exactly this kind of work, right now, with essentially no evidence base on whether it makes their numbers more or less reliable. The available benchmarks quantify how often models get spreadsheet tasks wrong, but they say nothing about whether the resulting errors survive contact with a human reviewer and if AI introduced errors are systematically less detectable, then organisations adopting copilots may be increasing their effective error exposure even where raw accuracy improves. Conversely, if detectability proves similar across conditions, practitioners can adopt the tools with a clearer conscience and redirect review effort to where it actually matters. Either result is decision-relevant. The study also gives the long established spreadsheet error literature its most timely extension: the same methods, applied to the authoring mode that has quietly displaced the one they were built for.

7. Limitations

Five limitations should be stated at the outset. Findings from controlled tasks may not fully generalise to the ambiguity and time pressure of production analytics work; the study is framed as a controlled baseline for future field replication. Copilot capability changes rapidly, so results are time bound to the specific tools and model versions tested and must be dated accordingly. Participant AI literacy and prompting skill vary and constitute a confound, measured and reported rather than eliminated. The detection sub-study uses a panel of analyst reviewers rather than the full range of people who actually review spreadsheets in organisations, which likely includes less analytically trained colleagues; detection rates measured here should be read as an upper bound on real world performance. Finally, error severity matching across conditions, while necessary for a fair detectability comparison, depends on the severity rubric being applied consistently which is why inter-rater reliability is established before the main audit rather than after.

8. Conclusion

Spreadsheet error research built its reputation on a simple insight: the errors that matter in practice are the ones that survive into a decision. AI copilots have changed who writes the formulas, but nobody has yet asked whether the resulting errors are more or less likely to survive review. This paper specifies the controlled study that answers matched tasks across manual and copilot conditions, a taxonomy that covers both, a blind detection experiment that measures what human reviewers actually catch, and a calibration component that tests the mechanism. The result would give organisations a testable answer to a question they are currently answering by assumption, and would extend a well respected literature into the authoring environment that has already replaced the one it studied.

Positioning statement

This study is deliberately incremental in method and novel in object. It reuses the established taxonomy and audit procedures, which makes the results comparable with decades of prior work, but applies them to an authoring mode those procedures were not designed for. The novel contribution is not a new error classification from scratch; it is the measurement of how a well understood class of errors behaves when a machine produces them and a human reviews them under ordinary conditions.

Author note. Correspondence concerning this manuscript should be addressed to the corresponding author. The study described is a proposed methodology; no empirical results are asserted. Benchmark error rates cited are those reported by the sources listed in Table 2 and are attributed to those sources, not measured in this study.

References

[1] Panko, R. R., & Aurigemma, S. (2010). Revising the Panko–Halverson taxonomy of spreadsheet errors. *Decision Support Systems*, 49(2), 235–244.

[2] Powell, S. G., Baker, K. R., & Lawson, B. (2009). Errors in operational spreadsheets. *Journal of Organizational and End User Computing*, 21(3), 24–36.

[3] Irons, R. (2003). The Wall and the Ball: A study of domain referent spreadsheet errors. *Proceedings of the Third Annual Conference of the European Spreadsheet Risks Interest Group (EuSpRIG)*, Dublin.

[4] Poon, et al. (2024). Systematic review of spreadsheet quality assurance. *Frontiers of Computer Science*. (Reported via Phys.org.)

[5] Sarkar, A. (2024). When Copilot becomes autopilot: Generative AI's critical risk to knowledge work and a critical solution. arXiv:2412.15030.

[6] Experimenting with ChatGPT for spreadsheet formula generation: Evidence of risk in AI generated spreadsheets. (2023). arXiv:2309.00095.

[7] SpreadsheetLLM: Encoding spreadsheets for large language models. (2024). arXiv:2407.09025. (Microsoft Research.)

[8] SpreadsheetBench: Towards challenging real world spreadsheet manipulation. (2024). arXiv:2406.14991. (NeurIPS 2024.)

[9] Benchmark dataset generation and evaluation for Excel formula repair with LLMs. (2025). arXiv:2508.11715.

[10] Large language models for spreadsheets: Benchmarking progress with FLARE. (2025). *EuSpRIG 2025 Proceedings*; arXiv:2506.17330.

[11] SODBench: A large language model approach to documenting spreadsheet operations. (2025). arXiv:2510.19864.

[12] Formula-R1: Incentivizing LLM reasoning over complex tables via formula-driven reinforcement learning. (2025). arXiv:2505.23667.

[13] Spreadsheet-RL: Advancing LLM agents on realistic spreadsheet workflows. (2026). arXiv:2605.22642.

[14] From rows to reasoning: A retrieval-augmented multimodal framework for spreadsheet understanding. (2026). arXiv:2601.08741.

[15] Trust-calibrated code review: A participatory design study. (2026). arXiv:2606.01969. (Lund University.)

[16] Sonar. (2026). *The AI trust gap: Why code verification matters*. State of Code survey.

[17] Atomic Robot. (2026). *AI writes better code. We're getting worse at reviewing it*.

[18] CFO.com. (2026). Who's at fault when AI messes up a financial model? *Financial Modeling Institute*.

[19] FM Magazine, AICPA & CIMA. (2025). *Financial modelling with AI: Part 1*.

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

[21] Skitka, L. J., Mosier, K. L., & Burdick, M. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, 52(4), 701–717.

[22] Coopers & Lybrand. (1997). *Spreadsheet errors: The state of the art*. (As cited in the spreadsheet-error literature.)

Author: Md. Anamul Hasan · Data Analyst | Statistics & Data Science Background

Author's Linkedin:https://www.linkedin.com/in/anamulhasan1/

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1aicopilot.png
- Table2aicopilot.png
- Table3aicopilot.png
- Table4aicopilot.png