Reproducing with AI Across the Expertise Ladder

Ghina Abdul Baki, Juan P. Aparicio, Bruno Barbarioli, Abel Brodeur,
Lenka Fiala, Derek Mikola, David Valenta

2025-10-09

1 Abstract

We test whether providing large-language-model (LLM) assistance compresses performance gaps in computational reproduction tasks along two pre-specified dimensions. Vertically, we study expertise tiers using two complementary stratifications: (i) the five-tier ladder (undergraduate, master's, PhD, postdoc/researcher, professor) and (ii) a three-category grouping (undergraduate, graduate student, professor/researcher) deployed when event-level counts require a coarser partition. Horizontally, we study cross-discipline performance in three quantitative social sciences (Economics, Political Science, Psychology) to assess whether AI reduces the penalty from working outside one's primary discipline (out-of-discipline, OOD). Participants are randomized 1:1 to AI access (ChatGPT Plus with tools) versus human-only controls within the applicable expertise strata; tasks carry a single discipline tag, with undergraduates always receiving inside-discipline assignments and approximately 30% of the remaining participants randomly assigned to an outside-discipline paper. Primary outcomes are: (i) successful reproduction, (ii) minutes to first success, and (iii) detection of coding errors (major, minor); referee-report outcomes evaluate summary accuracy, literature positioning, weakness diagnosis, recommendation quality, communication clarity, and an overall assessment. The study is designed to reveal whether AI acts as an equalizer within expertise tiers and whether it enables researchers to operate effectively across disciplines.

2 Registration and Funding

This pre-analysis plan is registered on the Open Science Framework (OSF) at https://osf.io/dkfzt/. Funded by Open Philanthropy and the Institute for Replication (I4R). The locked PAP together with the analysis scripts used to generate the mock tables and figures are mirrored there; live study code and data will be added only after the analysis lock.

3 Background and Rationale

Prior multi-site "AI Replication Games" documented measurable AI effects on success and speed in reproduction tasks, while also revealing substantial variation across teams and events. Building on that foundation, we pre-register two complementary dimensions of distributional impacts. First, the vertical dimension (expertise ladder): whether AI narrows performance gaps by disproportionately lifting less-experienced participants (equalizer) or widens gaps by enabling experts to better leverage tools (amplifier). Second, the horizontal dimension (cross-discipline): whether AI reduces the penalty from working outside one's primary discipline (out-of-discipline, OOD) when reproducing

studies across Economics, Political Science, and Psychology. These questions matter for pedagogy, workforce development, and equity: vertical compression would support broad inclusion strategies; horizontal compression would support cross-field mobility and knowledge diffusion; in both cases, amplification would call for targeted training and governance to avoid widening disparities.

We keep tasks, instructions, and grading rubrics closely aligned with prior exercises to ensure comparability while tailoring the design to individual-level randomization within strata and a discipline-tagged task pool. The design isolates intent-to-treat effects within expertise tiers and introduces an orthogonal OOD contrast by allocating approximately 30% of non-undergraduate participants to an outside-discipline article while keeping all undergraduates inside their primary field. This enables transparent tests of heterogeneous effects across tiers (vertical) and along the OOD dimension (horizontal). To support interpretability, we pre-specify a compact outcome set (levels, timing, error detection, and referee/report assessments) and a small list of precision-enhancing controls; discipline fixed effects are nested within article fixed effects and are thus absorbed in all models.

Several design features anchor the study and deserve to be unpacked carefully. We begin with randomization: participants enter either (i) one of five expertise strata—undergraduate, master's, PhD, postdoc/researcher, or professor—or (ii) a three-bin grouping that combines master's and PhD students into a graduate tier and postdocs with professors into a professor/researcher tier when sample sizes necessitate coarser cells. Within whichever stratification applies to a given event, participants are randomized 1:1 to either AI assistance or a human-only control condition. This stratification keeps comparisons within peer groups while preserving overall balance; it also fixes the reference point for the "vertical" estimands that trace how treatment effects evolve along the expertise ladder.

Task assignment introduces the horizontal dimension. Articles arrive with a single discipline tag (Economics, Political Science, or Psychology), enabling us to contrast performance inside versus outside a participant's primary field. We explicitly model two families of estimands—vertical (AI $\times$ tier) and horizontal (AI $\times$ out-of-discipline)—and we test each dimension on its own set of outcomes. This separation keeps the interpretation of compression tests transparent and avoids conflating discipline moves with tier moves.

Outcome measurement rounds out the design. The referee-report outcome is scored by both human referees—blinded to treatment and discipline—and an AI evaluator that follows the same rubric. The rubric distinguishes an Appropriateness indicator (binary) and six 0–5 scored dimensions: summary accuracy, contribution placement within the literature, diagnosis of weaknesses, actionability of recommendations, clarity and structure of the writing, and an overall holistic assessment. Section "Grading Rubric" details the anchors that guide human judgments, while Appendix "AI Referee Prompt" reproduces the corresponding instructions for the AI grader, ensuring that communication quality is assessed consistently across arms.

Our design choices address practical concerns. First, measurement: we standardize the classification of major/minor errors and use independent human and AI judges for referee outcomes to triangulate communication quality, with blinding to treatment and discipline. We also collect participants' robustness proposals and implementations as secondary outcomes to contextualize main effects (not part of primary inference families). Second, scope and external validity: by spanning multiple events, software ecosystems (R/Stata/Python), a broad experience range (undergraduate to professor), and three disciplines, we gauge how AI assistance interacts with realistic heterogeneity in tools, backgrounds, and fields. These features, combined with preregistration, separate multi-

plicity control for the vertical and horizontal families, and a limited set of pre-specified estimands, aim to balance credibility with informativeness.

4 Research Questions and Hypotheses

This study addresses three linked questions. First, we ask whether AI assistance compresses or widens the expertise gradient by comparing treatment effects across the five pre-defined tiers and, when applicable, across the three-category grouping; this provides the "vertical" lens that motivates the stratified randomization. Second, we investigate whether AI attenuates the penalty from tackling a problem outside one's primary discipline, thereby capturing "horizontal" mobility across Economics, Political Science, and Psychology. Third, we consider the intent-to-treat contrast averaged across all participants, which anchors the benchmark effect size against which heterogeneous responses are interpreted.

The corresponding hypotheses follow naturally. H1 posits that access to AI increases the probability of a successful reproduction and shortens the time required to reach that milestone. H2 states that any gains are at least as large for undergraduates and master's students as for postdocs and professors, implying a compression of tier gaps. H2b mirrors this logic across disciplines, anticipating that AI shrinks the outside-of-discipline penalty for success rates, time, and error detection. H3 focuses on quality control, predicting that AI exposure raises the detection of both major and minor coding errors. Each hypothesis is tested within the relevant estimand family, with multiplicity handled as described above.

Throughout the document we rely on several recurring definitions. "Reproduction success" means the participant's final output matches the pre-specified focal result within the documented tolerance; adjudication follows the rubric in Section "Grading Rubric." "Time-to-success" records the minutes from the start of the session until the first successful reproduction, with non-successes right-censored at 420 minutes. "Error detection" counts correctly identified coding errors and distinguishes between major issues—those that would alter the substantive result—and minor issues that affect presentation or reproducibility without changing the estimate. "Communication quality" aggregates the referee rubric components described earlier, combining the Appropriateness indicator with six scored dimensions of the written assessment. Finally, "out-of-discipline (OOD)" labels any case in which the task's discipline tag differs from the participant's self-reported primary discipline.

5 Experimental Design

We recruit participants across the five canonical strata—undergraduates, master's students, PhD students, postdocs/researchers, and professors—and observe each in a single, timed, one-day session. When registration numbers for a given event would leave some strata sparsely populated, we collapse to the three-tier grouping (undergraduate, graduate student, professor/researcher) for randomization while continuing to track individual titles. Every participant attempts to reproduce one pre-specified result using the same software ecosystems as the original studies (R, Stata, or Python) within a seven-hour working window. The window mirrors prior AI Replication Games, where most teams completed their submissions within seven hours while still allowing careful documentation. We randomly assign access to AI assistance (ChatGPT Plus with tools) within the relevant strata and events. Participants in the control arm pledge not to use AI tools. Any deviations are documented and, when material, addressed through the pre-specified per-protocol sensitivity analyses.

Human referees are blinded to both treatment and discipline, and filenames/metadata that could reveal either are redacted.

Within each event, reproduction papers are randomly assigned from a pre-curated pool once participant rosters are finalized. After the random draw, the research team audits each assigned paper to confirm that (i) the replication package includes executable code together with a README or documentation, (ii) all requisite datasets and intermediate files are accessible without additional permissions, and (iii) the combined data footprint remains tractable for local execution (target: compressed package 500 MB and memory requirements within a standard 16 GB RAM laptop). Papers that fail any check are replaced before assignments are communicated to participants.

Task materials are version-controlled in `Papers/`, which stores a journal-level folder for each study together with the inventory workbook `Papers/papers.xlsx`. As of the lock, the workbook enumerates 15 studies (AJPS = 5, AEJ: Applied Economics = 5, Psychological Science = 5) with DOI identifiers and replication-package URLs. Each journal folder contains the published article (`paper.pdf`) and the original replication package supplied by the journal or data repository; where only code is distributed (for example, "The Willingness to Pay for a Cooler Day"), the folder retains the vendor-provided archive so teams receive the same assets that were audited. This structure lets us share identical bundles across events while keeping provenance and updates transparent.

The design studies vertical compression across expertise and horizontal compression across disciplines in a unified framework (run simultaneously). Each article carries a single discipline tag—Economics, Political Science, or Psychology—and participants self-report a primary discipline at registration. The task pool spans all three fields at every event. Undergraduates always receive inside-discipline papers; among graduate students, postdocs, and professors we randomly select enough individuals to ensure that approximately 30% of the overall roster (subject to feasibility) works outside their primary discipline (OOD), with the remainder kept inside. This prioritizes a fixed OOD exposure rate over per-cell balance and reflects our expectation that undergraduates benefit from staying within field during a timed replication task. We report realized counts by tier, the collapsed tier grouping described below, discipline, and OOD status prior to the analysis lock. Event×article fixed effects absorb site, tooling, and task heterogeneity; because each article maps to a single discipline, separate discipline indicators are redundant.

For descriptive summaries we supplement the five-tier stratification with a collapsed `tier_2` indicator: undergraduates form the first category, master's and PhD students form a graduate grouping, and postdocs plus faculty comprise the professor/researcher tier. Events that implement three-bin randomization use this same collapsed partition for treatment assignment. Unless otherwise noted, inference continues to rely on the finest stratification available for each cohort; the collapsed measure streamlines tables and figures shared with partners.

5.1 Pre-Event Training

Participants complete a 60-minute orientation, "Getting Research Done with ChatGPT Plus and Modern AI," delivered the week before each event (slides in `Pre game/` with file `ai_research_webinar_codex_cli_v2.pdf`). The session covers (i) plain-language intuition for LLMs and prompting patterns, (ii) the capabilities of ChatGPT Plus with Advanced Data Analysis, Deep Research, and Agent Mode, (iii) an end-to-end research workflow illustrating literature reviews, data documentation, cleaning, interpretation, and writing, and (iv) guardrails for reproducible and responsible use. Demonstration prompts embedded in the deck provide copy-ready examples for literature scaffolds, data dictionaries, cleaning scripts, interpretive

summaries, reviewer checklists, and Codex CLI micro-automations. The training closes with a five-point reproducibility checklist and a quick Codex CLI primer so treated participants and graders share the same vocabulary and expectations going into the sessions.

5.2 Post-Pilot Focus Groups

To contextualize quantitative findings, we will run five parallel focus-group sessions (six participants per group) immediately after the pilot wave concludes. Scheduling them post-pilot minimizes contamination risk: participants cannot brief future cohorts about task structure or AI prompts before those sessions occur. Groups are stratified by treatment status (AI-assisted vs. human-only) and, when numbers permit, by inside/outside-discipline assignments so that discussions surface arm-specific workflows and cross-discipline frictions. Each 60-minute session is moderated by Institute for Replication staff using the standardized guide in `focus_groups/focus_group_guide.md`; facilitators remind attendees of confidentiality expectations and collect recorded consent before beginning. Discussion notes and transcripts are coded with the companion qualitative codebook (`focus_groups/codebook.md`), which maps themes on motivations, preparation, AI use or workarounds, cross-discipline challenges, and suggestions for future waves. Insights from these sessions inform protocol refinements before scaling beyond the pilot.

Design cells overview (what is crossed with what):

Table 1: Design factors and levels (crossed: Arm × Tier × OOD).

Factor	Levels	Notes
Arm	Human-Only; AI-Assisted	1:1 allocation within tier
Expertise tier	UG; MA; PhD; PD; P	Stratification (randomization blocks)
OOD status	Inside; Outside	Derived from participant vs task discipline

Randomization is implemented with a reproducible script and a fixed seed recorded in the registry; the assignment file is timestamped and stored read-only. Allocation is concealed until check-in, when the onsite coordinator reveals arm and task. No-shows remain in their assigned arm for intent-to-treat analyses. Replacements are permitted only before the event begins and are re-randomized using the same stratum-specific seed. The same concealment and documentation protocol applies to inside/outside assignments. Any late swaps or deviations are logged prior to accessing outcomes.

We plan to enroll roughly 300 participants across multiple events. For power, we assume a plausible tier composition (more undergraduates than postdocs/professors) rather than equal counts by tier. Simulations suggest that, with baseline success gaps between undergraduates and professors of 15–20 percentage points and AI compressing roughly 40% of that gap, we achieve at least 80% power for the vertical interactions at = 0.05. For the horizontal dimension, we assume a baseline OOD penalty of 20–25 percentage points in the control arm and target detectable AI-induced reductions of about 8–10 percentage points, while preserving the same overall sample size. Standard errors are clustered at the event×software level; when the number of clusters is modest we report wild-cluster bootstrap p-values alongside conventional ones. Table 2 summarizes the core inputs; we will freeze any updates to these assumptions prior to registry lock.

The primary analysis set follows intent-to-treat principles and includes all randomized individuals with any outcome data. We do not impute outcomes. For success and error counts, nonresponse

results in missing outcomes that are excluded from that specific regression but retained for other outcomes. For timing, non-successes are right-censored at the session cap (420 minutes) in survival analyses; we do not impute minutes for OLS models. For covariates only, we handle missingness as follows: categorical covariates (tier, software, prior ChatGPT familiarity, and—where used descriptively—participant and task discipline) gain an explicit "Missing" category if needed; continuous covariates (years of coding) use within-stratum median imputation with a missingness indicator. We report outcome and covariate missingness by arm and verify robustness to listwise deletion.

Table 2: Prospective power and design inputs (pre-lock)

Quantity	**Value**
Participants (N)	300 (approximate)
Tier composition (assumed)	UG 30–35%, MA 20–25%, PhD 20–25%, PD 10–15%, P 8–10%
Discipline composition	≈50% Econ, 25% PolSci, 25% Psych
Allocation	1:1 within tier (AI vs Control)
Inside vs Outside discipline	Undergrads always inside; 30% of roster assigned outside (selected from non-UG)
Clusters for SE	Event × software (10–20 anticipated)
Minutes cap	420 minutes (right-censor in survival)
Vertical: control success (UG; P)	40% ; 55–60% (assumed)
Vertical: baseline gap (UG vs Prof)	15–20 pp (assumed)
Vertical: detectable compression	≈ 40% of gap @ 80% power, $\alpha = 0.05$
Vertical: time outcome variability	SD ≈ 60 minutes
Horizontal: OOD penalty (control)	20–25 pp (assumed)
Horizontal: detectable AI reduction (Success)	≈ 8–10 pp @ 80% power, $\alpha = 0.05$
Horizontal: detectable AI reduction (Minutes)	≈ 8–12 minutes @ 80% power (illustrative)
Multiplicity	Tests conducted separately within vertical and horizontal families
Small-cluster inference	Wild-cluster bootstrap if clusters < 30 (9,999 reps)

6 Outcomes and Measurement

Participants receive a discipline-tagged article, a standardized instruction sheet, and a data/code package (as available) mirroring the original study's setup. They are asked to reproduce a pre-specified focal result and to document their workflow. At the end of the seven-hour window, participants submit (i) a final result file (tables/figures or numeric outputs), (ii) executable code and a short README describing the steps taken, (iii) an error log listing any major and minor coding issues identified (with file and line references when possible), and (iv) a brief referee-style assessment of the credibility and clarity of the reproduced evidence. Human graders (blinded to treatment and discipline) use a rubric to classify success, time to first success, and error detection, and to score each referee component (summary accuracy, literature placement, weakness diagnosis, recommendation quality, clarity, and overall assessment) as well as the Appropriateness indicator. AI-assisted grading (for the AI referee outcomes) follows the same prompts and rubric; Appendix "AI Referee Prompt" reproduces the exact instructions, including required outputs and scoring anchors.

We standardize these deliverables with the empty workbook `Reports/Replication_Log_Referee_Template.xlsx`, which participants fill as they work. Sheet `00_Main` captures session metadata (participant name, article identifiers, software, event, discipline tags, and the out-of-discipline flag). Sheet `01_CodingErrors` provides the structured log for major/minor issues, including timestamp, affected element, narrative justification, evidence pointer, and a minor/major classification toggle. Sheet `02_Computation` records whether the focal result was reproduced and when. Sheet

`03_Robustness` allocates parallel columns for up to two robustness checks—each with motives, specification changes, original versus reproduced estimates, and interpretation—enforcing the "maximum two" rule in the protocol. Sheet `04_Referee_Report` houses the scored rubric, combining binary/0–5 entries with prompts about what to cover in the narrative. Aligning the reporting template with the PAP ensures fields are named consistently across sites and can be ingested without ad-hoc cleaning.

We organize outcomes into primary and secondary categories to align directly with our hypotheses and to keep inference focused. Primary outcomes capture whether participants reproduced the pre-specified result (level), how long it took to first achieve a reproduction (timing), their ability to detect coding errors (major and minor), and performance on the structured referee rubric (Appropriateness, five content components, and an overall score). These outcomes together reflect the core goals of the exercise: getting to the right result, getting there efficiently, avoiding substantive mistakes, and communicating clearly.

Secondary and exploratory outcomes provide contextual texture and mechanisms. In particular, we summarize robustness proposals and implementation, and—within the AI arm only—self-reported usage intensity via prompts/files/images/words (entered as inverse hyperbolic sine transformations). These help differentiate under- or over-use patterns and support interpretation of treatment effects. Finally, two pre-specified moderators (self-reported years of coding and prior AI usage) enter as covariates in the main models and, where noted, as separate heterogeneity analyses; they are not outcomes themselves. The table below consolidates definitions, types, and assessment sources for quick reference; the analyses and figures throughout the plan use these exact definitions.

Table 3: Outcomes and measurements (primary and secondary).

Category	Outcome	Type	Measurement / Assessment
	Success	Binary	Core result reproduced by endline (yes/no).
	Time-to-success	Time (min)	Minutes until first successful reproduction (visualized via KM curves).
	Error detection — major	Count	Number of major coding errors correctly identified (pre-defined rubric).
	Error detection — minor	Count	Number of minor coding errors correctly identified (pre-defined rubric).
	Referee — appropriateness	Binary	Appropriate vs. Not appropriate; human judges: A. Brodeur, J. Aparicio, D. Mikola; AI judge separately.
	Referee — summary accuracy	0–5	0–5 (higher is better); captures fidelity of the paper summary; human score averaged across three judges; AI judge recorded separately.
	Referee — literature placement	0–5	0–5; assesses how well the report situates the paper's contribution in the literature; human score averaged across three judges; AI judge recorded separately.
	Referee — weakness diagnosis	0–5	0–5; evaluates identification and prioritization of key weaknesses or threats to validity; human score averaged across three judges; AI judge recorded separately.

Primary			
	Referee — recommendations	0–5	0–5; measures specificity and actionability of suggested fixes or robustness checks; human score averaged across three judges; AI judge recorded separately.
	Referee — clarity	0–5	0–5; judges organization, tone, and clarity of the write-up; human score averaged across three judges; AI judge recorded separately.
	Referee — overall	0–5	0–5 composite overall assessment (not an average of components; scored directly); human score averaged across three judges; AI judge recorded separately.
Secondary	Robustness proposals — quality	Ordinal	Quality of robustness proposals (standardized rubric).
	Robustness implementations — count	Count	Number of robustness checks successfully implemented (standardized rubric).
	Prompt usage — count (AI arm)	Count	Self-reported prompt count (asinh transformed; AI arm only).
	Prompt usage — length (AI arm)	Continuous	Self-reported prompt length in characters (asinh transformed; AI arm only).
Moderator	Years of coding	Continuous	Self-reported; used as moderator (not an outcome).
	Prior AI usage	Categorical	Self-reported; used as moderator (not an outcome).

We pre-define classification of "major" versus "minor" errors and use a standardized grading rubric. As a preview of magnitudes, the table below (Table 4) reports simple means (and standard deviations) by arm, together with Welch tests for the difference in means. Figures then visualize the core outcomes and timing distributions that the analysis will focus on (Figures 1 and 2).

The figures referenced here are predefined mock-ups of the panels we will produce once data are available; they are rendered using synthetic or placeholder data solely to illustrate the visual layout and expected content. Their purpose is to communicate the intended presentation of our preregistered estimands rather than to reveal any substantive pattern at this stage.

Table 4: Comparison of Human-Only and AI-Assisted Metrics

Variable	Human-Only	AI-Assisted	Human-Only vs AI-Assisted
Reproduction	0.455 (0.501)	0.473 (0.502)	-0.019 [0.804]
Minutes to success	243.674 (194.493)	235.241 (196.369)	8.433 [0.772]
Number of minor errors	0.398 (0.598)	0.419 (0.631)	-0.022 [0.813]
Number of major errors	0.193 (0.425)	0.183 (0.389)	0.010 [0.864]
At least one good robustness check	0.443 (0.500)	0.409 (0.494)	0.035 [0.640]
At least two good robustness checks	0.102 (0.305)	0.237 (0.427)	-0.134 [0.016]
Appropriate (human)	0.500 (0.503)	0.355 (0.481)	0.145 [0.049]
Summary 0–5 (human)	3.249 (0.871)	3.271 (0.778)	-0.022 [0.858]
Literature 0–5 (human)	3.249 (0.871)	3.271 (0.778)	-0.022 [0.858]
Weakness 0–5 (human)	3.249 (0.871)	3.271 (0.778)	-0.022 [0.858]
Recommendations 0–5 (human)	3.249 (0.871)	3.271 (0.778)	-0.022 [0.858]
Clarity 0–5 (human)	3.249 (0.871)	3.271 (0.778)	-0.022 [0.858]
Overall 0–5 (human)	3.249 (0.871)	3.271 (0.778)	-0.022 [0.858]
Appropriate (AI)	0.466 (0.502)	0.462 (0.501)	0.004 [0.962]
Summary 0–5 (AI)	3.239 (0.892)	3.277 (0.786)	-0.038 [0.760]
Literature 0–5 (AI)	3.239 (0.892)	3.277 (0.786)	-0.038 [0.760]
Weakness 0–5 (AI)	3.239 (0.892)	3.277 (0.786)	-0.038 [0.760]
Recommendations 0–5 (AI)	3.239 (0.892)	3.277 (0.786)	-0.038 [0.760]
Clarity 0–5 (AI)	3.239 (0.892)	3.277 (0.786)	-0.038 [0.760]
Overall 0–5 (AI)	3.239 (0.892)	3.277 (0.786)	-0.038 [0.760]

Note: Columns 2–3 present means and standard deviations in parentheses for the two arms; column 4 presents the difference in means (Human-Only − AI-Assisted) and two-sided Welch p-values in brackets.

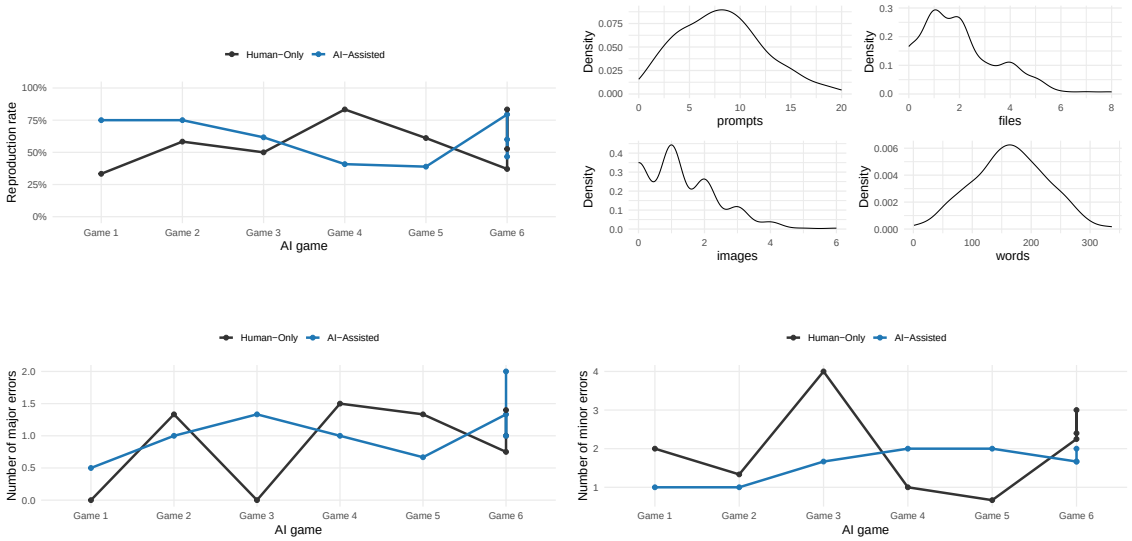

Figure 1: Primary outcomes (levels): reproduction and errors, plus usage context. Notes: Four-panel layout with (top-left) reproduction rates (raw), (top-right) prompt distribution (usage), (bottom-left) major errors (raw), (bottom-right) minor errors (raw). Difference-style plots are intentionally omitted.

6.1 Grading Rubric (Operational Definitions)

Success and timing - Success (binary): "Yes" when the participant's final output matches the pre-specified focal result within documented tolerance (numeric threshold or visual equivalence for figures), with code that runs cleanly to produce the output. Partial or alternative results do not count. - Time-to-success: Minutes from start until first successful reproduction; right-censored at 420 minutes for non-successes. Recorded from code logs and submissions.

Error detection - Major error (count): Any coding/data/model issue that, once corrected, changes the focal result's sign, statistical significance, or substantive interpretation; for example a wrong sample filter or an omitted transformation that alters the reported effect size. Issues that could matter but do not demonstrably affect the focal estimate are coded as minor. - Minor error (count): Formatting, non-substantive code, or reproducibility issues that do not change the focal result. Examples: mislabeled axes; non-deterministic seed; inefficient code without impact on estimate.

Referee/report outcomes - Appropriateness (binary): "Appropriate" when the report accurately identifies whether the reproduced evidence supports the original claim, flags substantive issues if present, and substantiates claims with code/snippets or references to outputs. - Summary accuracy (0–5): Fidelity of the paper summary. 0=Missing or incorrect; 1=Material inaccuracies; 2=Partial coverage with errors; 3=Correct but thin; 4=Comprehensive and precise; 5=Comprehensive, precise, and highlights key quantitative details. - Literature placement (0–5): Ability to situate the reproduced result within the relevant literature. Same anchors as above (0=missing/incorrect through 5=exceptional and well justified), focusing on comparative framing and contextualization. - Weakness diagnosis (0–5): Identification and prioritization of substantive weaknesses or threats to validity. Anchors follow the same scale, emphasizing whether weaknesses are material, supported, and prioritized. - Recommendations (0–5): Specificity and actionability of suggested fixes, robustness checks, or next steps. Anchors follow the same 0–5 scale with 5 meaning clear, prioritized, and

feasible guidance. - Clarity (0–5): Organization, tone, and readability of the write-up. Anchors again follow the 0–5 scale with 5 indicating concise, professional communication. - Overall (0–5): Holistic assessment of the referee report quality (not a mechanical average of components). Anchors use the same 0–5 scale.

Human graders apply these definitions blind to treatment and discipline; disagreements are reconciled via consensus. The same rubric is prompted to the AI judge for the AI-referee outcomes.

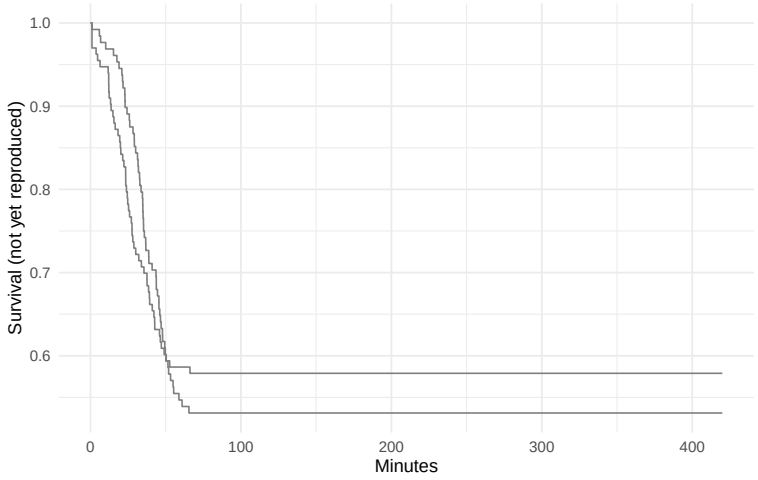

Figure 2: Time-to-success: Kaplan–Meier by arm. Notes: Survival curves stratified by treatment (Control vs. ChatGPT+).

We will interpret Figure 1 as the primary visualization of level outcomes (H1 and H3) and Figure 2 as complementary evidence on timing. We omit difference-style and cumulative milestone panels to reduce redundancy and focus attention on the preregistered estimands.

7 Controls (Covariates and Stratification)

Table 5 summarizes the covariates and fixed effects we pre-specify for precision and design alignment. We use a compact, stable set that mirrors the randomization scheme and absorbs systematic heterogeneity without overfitting.

We include a small, pre-specified set of controls to improve precision and absorb systematic differences that are not of direct interest. Stratification by expertise tier (Undergraduate, Master's, PhD, Postdoc, Professor) reflects our design and is included as dummies in the main specification so that treatment effects are identified within tier; the interaction with treatment captures heterogeneous effects along the expertise ladder. Event and article fixed effects absorb site- and task-specific differences. Software indicators (R/Stata/Python) capture baseline workflow differences across toolchains. Finally, self-reported years of coding and prior AI familiarity improve precision and help stabilize estimates across events.

Two variables play a dual role as moderators in pre-specified secondary analyses: (i) years of coding (interacted with treatment), and (ii) within-AI usage measures (prompts, files, images, words) which we study only in the AI arm to characterize under-/over-use patterns. These moderators are always treated as covariates in the main models; the secondary analyses are reported separately and do not change the main estimands.

Table 5: Controls, fixed effects, and moderators (pre-specified).

Variable	Role	Type	Coding_or_Levels	Notes
Expertise tier	Stratification; control	Categorical	Undergrad, Master's, PhD, Postdoc, Professor	Tier dummies in main specs; interacted with AI (heterogeneity).
Event	Fixed effect	Categorical	One FE per event	Absorbs site/time differences (not a parameter of interest).
Article	Fixed effect	Categorical	One FE per task/article	Absorbs task-specific difficulty/fit (not a parameter).
Software	Control	Categorical	R, Stata, Python	Preferred software indicator (workflow baseline).
Years of coding	Control; moderator	Continuous	Self-reported years	Improves precision; interacted with AI in secondary (yrs×AI).
Prior ChatGPT familiarity	Control	Categorical	None, Some, Heavy	Self-reported familiarity with ChatGPT/AI tools.
Usage (AI arm)	Moderator (AI only)	Continuous	Self-reported prompts/files/images/words (asinh)	Secondary/appendix within AI arm; not a control in main ITT.
Clustering	Estimation setting	—	SE clustered by event×software	Variance estimation (not a control).

8 Statistical Analysis Plan

This section lays out our estimands and modeling approach for the two preregistered dimensions of interest—vertical (expertise tier) and horizontal (out-of-discipline, OOD)—and clarifies how outcome types map to link functions and fixed-effect structure. We begin with intent-to-treat (ITT) specifications and then describe the interaction terms that capture compression: AI × tier for vertical and AI × OOD for horizontal. Event×article fixed effects absorb site/tool and task heterogeneity; standard errors cluster at the event×software level, and when the number of clusters is modest we complement conventional inference with wild-cluster bootstrap. We control multiplicity separately within the vertical and horizontal families using Holm's method at = 0.05.

Let Y_i denote a pre-specified outcome (success; minutes to success; error counts; referee-report assessments). We estimate ITT effects in a tier-interaction framework that allows the AI effect to vary with expertise. Formally,

$$Y_i \;=\; \beta_0 \;+\; \beta_1 A_i \;+\; \sum_s \gamma_s \,\mathbb{1}\{\mathrm{Tier}_i = s\} \;+\; \sum_s \delta_s \, A_i \,\mathbb{1}\{\mathrm{Tier}_i = s\} \;+\; X_i'\theta \;+\; \lambda_{e(i)\times s(i)} \;+\; \varepsilon_i,$$

where A_i is the treatment indicator, $\lambda_{e(i)\times s(i)}$ are event-by-software fixed effects, and X_i collects the pre-specified controls (years of coding and prior AI familiarity). For binary outcomes (success; appropriate referee) we estimate linear probability models; for continuous outcomes (minutes; referee component and overall 0–5 scores) we use OLS; and for counts (minor/major errors) we fit a Poisson GLM with a log link:

$$\mathbb{E}[Y_i \mid \cdot] \;=\; \exp\!\left(\beta_0 + \beta_1 A_i + \sum_s \gamma_s \,\mathbb{1}\{\mathrm{Tier}_i = s\} + \sum_s \delta_s \, A_i \,\mathbb{1}\{\mathrm{Tier}_i = s\} + X_i'\theta + \lambda_{e(i)\times s(i)}\right).$$

Horizontal compression estimands. Let D_i denote an indicator for receiving a task outside the participant's primary discipline (OOD). We test whether AI reduces the OOD penalty by augmenting the main models with D_i and $A_i \times D_i$: $Y_i = \cdots + \beta_2 D_i + \beta_3 (A_i \times D_i) + \varepsilon_i$. The coefficient β_3 captures horizontal compression (a smaller penalty, or a gain, when outside). Article indicators absorb discipline effects because tasks carry a single discipline tag.

Interpretation and outcome-specific notes. In all tables, the "AI-Assisted" row reports the ITT contrast for participants in the baseline tier or inside-discipline case; the interaction rows report how that contrast varies along tiers (vertical) or OOD (horizontal). Minutes are modeled via OLS and visualized with Kaplan–Meier curves; we will also report nonparametric log-rank tests and emphasize right-censoring at 420 minutes. For counts, we fit Poisson models and check over-dispersion (deviance/df); if material, we will re-estimate with a negative binomial link as a robustness check.

Multiplicity. We pre-register two primary families and control the family-wise error rate within each using Holm's method at $\alpha = 0.05$: (i) Vertical family: the four primary outcomes in the tier-interaction models (AI main and AI×tier); (ii) Horizontal family: the four primary outcomes in the OOD-interaction models (AI main and AI×OOD). Referee outcomes and years-of-coding heterogeneity are secondary.

To assess the success of randomization and support the modeling choices, we report balance on all individual-level controls used in the specification. The table below (Table 6) shows means (standard deviations) by arm and Welch tests for the difference in means.

Table 6: Balance of Participant Characteristics by Arm

Variable	Human-Only	AI-Assisted	Human-Only vs AI-Assisted
Years of coding	4.456 (2.768)	5.031 (2.918)	-0.576 [0.103]
Tier: Undergraduate	0.273 (0.447)	0.203 (0.404)	0.070 [0.184]
Tier: Master's	0.188 (0.392)	0.165 (0.373)	0.022 [0.642]
Tier: PhD	0.266 (0.443)	0.248 (0.434)	0.018 [0.747]
Tier: Postdoc	0.133 (0.341)	0.180 (0.386)	-0.048 [0.291]
Tier: Professor	0.141 (0.349)	0.203 (0.404)	-0.062 [0.182]
Software: R	0.336 (0.474)	0.293 (0.457)	0.043 [0.460]
Software: Stata	0.359 (0.482)	0.338 (0.475)	0.021 [0.723]
Software: Python	0.305 (0.462)	0.368 (0.484)	-0.064 [0.278]
Prior ChatGPT familiarity: None	0.430 (0.497)	0.338 (0.475)	0.091 [0.130]
Prior ChatGPT familiarity: Some	0.305 (0.462)	0.368 (0.484)	-0.064 [0.278]
Prior ChatGPT familiarity: Heavy	0.266 (0.443)	0.293 (0.457)	-0.028 [0.621]

Note: Means and standard deviations in parentheses by arm; difference column shows Human-Only − AI-Assisted and two-sided Welch p-values in brackets. All variables are individual-level controls used in the models.

For time-to-event (minutes to success), we present nonparametric Kaplan–Meier curves by arm and report the log-rank test for equality of survival functions. In all models, we use heteroskedasticity-robust standard errors clustered at the event–software level. We report coefficient estimates with 95% confidence intervals for the main effect β_1 and the tier interactions δ_s, and we conduct the pre-specified compression test on the interaction terms. As a sensitivity check, we will also provide wild-cluster bootstrap p-values when the number of clusters is modest.

Heterogeneity and secondary analyses follow two paths. First, we replace tier dummies with a continuous moderator (years of coding) interacted with treatment to trace a dose–response. Second, within the AI arm only, we relate outcomes to self-reported usage intensity (prompts/files/images/words, entered as inverse hyperbolic sine transforms) to characterize under- and over-use; these are descriptive and do not alter the main ITT estimands. We also consider event-order interactions to gauge learning across events.

9 Results

For clarity and symmetry, we present vertical (tier-based) estimates first, followed by parallel horizontal (AI × out-of-discipline) estimates using the same outcomes and table layout. Figures

and descriptives mirror this sequence.

We present results in three parts that map directly to the research questions. First, we report intent-to-treat effects of AI access on the primary outcomes, with expertise-tier interactions to quantify distributional patterns (equalizer vs. amplifier). Second, we evaluate referee-report outcomes to capture communication and assessment quality, paralleling the main design with human and AI judges. Third, we summarize core robustness checks that probe alternative definitions and model choices. Throughout, standard errors are clustered at the event–software level, and the coefficients are displayed with confidence intervals and a pre-specified compression test for the interaction terms.

The main table (Table 7) provides a compact view of the four primary outcomes: reproduction, minutes to success, and counts of minor and major errors. The AI coefficient speaks to H1 (average effect), while the interactions across tiers speak to H2 (compression). We interpret effect magnitudes jointly rather than in isolation, looking for coherence across levels, timing, and error detection. The corresponding Kaplan–Meier curves (Figure 2) provide a complementary view of H1 for time-to-success, and are discussed alongside these estimates.

Table 7: Vertical results: main effects across outcomes (tier interactions; pre-analysis layout). Standard errors clustered by event–software.

	(1) Reproduction	(2) Minutes	(3) Minor	(4) Major	(5) At least 1 check	(6) At least 2 checks
AI-Assisted	-0.051	17.508	0.388	0.154	-0.062	0.165
	(0.163)	(63.960)	(0.369)	(0.779)	(0.111)	(0.097)
	[-0.389; 0.287]	[-114.803; 149.819]	[-0.335; 1.110]	[-1.374; 1.681]	[-0.292; 0.169]	[-0.036; 0.366]
AI × Master's	0.103	-41.193	-0.216	1.216	0.433**	-0.054
	(0.183)	(70.801)	(0.412)	(1.441)	(0.203)	(0.174)
	[-0.276; 0.483]	[-187.656; 105.269]	[-1.022; 0.591]	[-1.608; 4.041]	[0.013; 0.852]	[-0.413; 0.305]
AI × PhD	0.013	-11.386	-0.030	-1.055	0.372	-0.136
	(0.216)	(84.650)	(0.608)	(1.050)	(0.221)	(0.157)
	[-0.433; 0.459]	[-186.498; 163.726]	[-1.223; 1.162]	[-3.113; 1.003]	[-0.085; 0.828]	[-0.459; 0.188]
AI × Postdoc	-0.073	22.807	-1.358**	-0.074	0.106	-0.019
	(0.249)	(95.412)	(0.582)	(1.028)	(0.194)	(0.142)
	[-0.588; 0.441]	[-174.568; 220.182]	[-2.498; -0.219]	[-2.089; 1.940]	[-0.295; 0.508]	[-0.313; 0.275]
AI × Professor	0.043	-14.535	-0.834	0.078	0.208	-0.157
	(0.166)	(64.533)	(0.601)	(1.251)	(0.183)	(0.153)
	[-0.299; 0.386]	[-148.032; 118.961]	[-2.012; 0.345]	[-2.373; 2.530]	[-0.170; 0.586]	[-0.474; 0.159]
Controls	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.444	247.101	0.421	0.176	0.410	0.157
p-val (Monotonic compression)	0.711	0.696	0.773	0.795	0.760	0.782
Obs.	261	261	261	261	261	261

Note: Standard errors in parentheses; confidence intervals in brackets.
Controls: Event × article FE; years of coding; software; prior AI familiarity.
Compression (monotonic): one-sided p-value for increasing effects across tiers (baseline: Undergraduate).
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Definition of robustness columns: " 1 check" indicates at least one qualifying robustness check per the pre-specified rubric; " 2 checks" indicates at least two qualifying checks.

We will interpret the AI coefficient as H1, the set of interaction terms as H2, and compare magnitudes across outcomes in the same layout to assess coherence.

To complement the vertical (tier-based) heterogeneity, we report horizontal compression by interacting AI with an indicator for being outside one's discipline (OOD). The table below mirrors the main layout and reports the AI main effect and the AI×OOD interaction for the same outcomes.

Table 8: Horizontal results: AI × outside-of-discipline across outcomes. Standard errors clustered by event–software.

	(1) Reproduction	(2) Minutes	(3) Minor	(4) Major	(5) At least 1 check	(6) At least 2 checks
AI-Assisted	-0.131	44.313	0.786**	0.380	0.135	0.129
	(0.122)	(47.343)	(0.349)	(0.494)	(0.117)	(0.088)
	[-0.384; 0.123]	[-53.623; 142.248]	[0.102; 1.469]	[-0.587; 1.348]	[-0.107; 0.377]	[-0.053; 0.311]
AI × Outside-discipline	0.162	-60.246	-1.256***	-0.625	0.061	-0.069
	(0.170)	(65.944)	(0.439)	(0.825)	(0.134)	(0.129)
	[-0.189; 0.514]	[-196.662; 76.170]	[-2.117; -0.396]	[-2.242; 0.992]	[-0.217; 0.339]	[-0.335; 0.197]
Controls	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.444	247.101	0.421	0.176	0.410	0.157
Obs.	261	261	261	261	261	261

Note: Standard errors in parentheses; confidence intervals in brackets.
Controls: Event × article FE; years of coding; software; prior AI familiarity.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

We next examine referee-report outcomes (Table 9), which connect to the communication and assessment dimension of reproduction. We pre-specify models for the binary Appropriateness indicator alongside each of the six 0–5 component/overall scores, with results reported separately for human and AI assessors but estimated on the same right-hand side as the main specification. These results shed light on whether AI support changes not just success and speed, but also the quality of participants' evaluation of evidence and errors (H3), and whether patterns mirror the tier-based compression observed in the main outcomes. In the Appendix, we present years-based versions of the main and referee tables (Tables 11 and 12; replacing tier with a continuous years-of-coding interaction) to complement the tier-based analysis and trace a dose–response along experience.

Table 9: Vertical results: referee report outcomes (human and AI assessments). Standard errors clustered by event–software.

	(1) Appropriate (human)	(2) Summary 0–5 (human)	(3) Literature 0–5 (human)	(4) Weakness 0–5 (human)	(5) Recommendations 0–5 (human)	(6) Clarity 0–5 (human)	(7) Overall 0–5 (human)	(8) Appropriate (AI)	(9) Summary 0–5 (AI)	(10) Literature 0–5 (AI)	(11) Weakness 0–5 (AI)	(12) Recommendations 0–5 (AI)	(13) Clarity 0–5 (AI)	(14) Overall 0–5 (AI)
AI-Assisted	0.034	0.288	0.288	0.288	0.288	0.288	0.288	0.084	0.330*	0.330*	0.330*	0.330*	0.330*	0.330*
	(0.149)	(0.218)	(0.218)	(0.218)	(0.218)	(0.218)	(0.218)	(0.134)	(0.191)	(0.191)	(0.191)	(0.191)	(0.191)	(0.191)
	[-0.274; 0.342]	[-0.162; 0.738]	[-0.162; 0.738]	[-0.162; 0.738]	[-0.162; 0.738]	[-0.162; 0.738]	[-0.162; 0.738]	[-0.193; 0.362]	[-0.066; 0.725]	[-0.066; 0.725]	[-0.066; 0.725]	[-0.066; 0.725]	[-0.066; 0.725]	[-0.066; 0.725]
AI × Master's	-0.163	-0.177	-0.177	-0.177	-0.177	-0.177	-0.177	0.022	-0.404	-0.404	-0.404	-0.404	-0.404	-0.404
	(0.210)	(0.374)	(0.374)	(0.374)	(0.374)	(0.374)	(0.374)	(0.157)	(0.376)	(0.376)	(0.376)	(0.376)	(0.376)	(0.376)
	[-0.599; 0.272]	[-0.951; 0.597]	[-0.951; 0.597]	[-0.951; 0.597]	[-0.951; 0.597]	[-0.951; 0.597]	[-0.951; 0.597]	[-0.303; 0.346]	[-1.182; 0.375]	[-1.182; 0.375]	[-1.182; 0.375]	[-1.182; 0.375]	[-1.182; 0.375]	[-1.182; 0.375]
AI × PhD	0.088	-0.026	-0.026	-0.026	-0.026	-0.026	-0.026	-0.150	-0.073	-0.073	-0.073	-0.073	-0.073	-0.073
	(0.160)	(0.249)	(0.249)	(0.249)	(0.249)	(0.249)	(0.249)	(0.265)	(0.273)	(0.273)	(0.273)	(0.273)	(0.273)	(0.273)
	[-0.243; 0.420]	[-0.540; 0.488]	[-0.540; 0.488]	[-0.540; 0.488]	[-0.540; 0.488]	[-0.540; 0.488]	[-0.540; 0.488]	[-0.698; 0.399]	[-0.638; 0.492]	[-0.638; 0.492]	[-0.638; 0.492]	[-0.638; 0.492]	[-0.638; 0.492]	[-0.638; 0.492]
AI × Postdoc	0.112	0.151	0.151	0.151	0.151	0.151	0.151	-0.230	0.149	0.149	0.149	0.149	0.149	0.149
	(0.254)	(0.377)	(0.377)	(0.377)	(0.377)	(0.377)	(0.377)	(0.244)	(0.387)	(0.387)	(0.387)	(0.387)	(0.387)	(0.387)
	[-0.413; 0.637]	[-0.628; 0.931]	[-0.628; 0.931]	[-0.628; 0.931]	[-0.628; 0.931]	[-0.628; 0.931]	[-0.628; 0.931]	[-0.735; 0.275]	[-0.651; 0.949]	[-0.651; 0.949]	[-0.651; 0.949]	[-0.651; 0.949]	[-0.651; 0.949]	[-0.651; 0.949]
AI × Professor	-0.086	0.268	0.268	0.268	0.268	0.268	0.268	0.147	0.121	0.121	0.121	0.121	0.121	0.121
	(0.186)	(0.407)	(0.407)	(0.407)	(0.407)	(0.407)	(0.407)	(0.179)	(0.394)	(0.394)	(0.394)	(0.394)	(0.394)	(0.394)
	[-0.471; 0.299]	[-0.573; 1.110]	[-0.573; 1.110]	[-0.573; 1.110]	[-0.573; 1.110]	[-0.573; 1.110]	[-0.573; 1.110]	[-0.222; 0.517]	[-0.693; 0.935]	[-0.693; 0.935]	[-0.693; 0.935]	[-0.693; 0.935]	[-0.693; 0.935]	[-0.693; 0.935]
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.464	3.199	3.199	3.199	3.199	3.199	3.199	0.418	3.212	3.212	3.212	3.212	3.212	3.212
p-val (Monotonic compression)	0.977	0.713	0.713	0.713	0.713	0.713	0.713	0.955	0.789	0.789	0.789	0.789	0.789	0.789
Obs.	261	261	261	261	261	261	261	261	261	261	261	261	261	261

Note: Standard errors in parentheses; confidence intervals in brackets.
Controls: Event×article FE; years of coding; software; prior AI familiarity.
Compression (monotonic): one-sided p-value for increasing effects across tiers (baseline: Undergraduate).
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 10: Horizontal results: referee outcomes (AI × outside-of-discipline). Standard errors clustered by event–software.

	(1) Appropriate (human)	(2) Summary 0–5 (human)	(3) Literature 0–5 (human)	(4) Weakness 0–5 (human)	(5) Recommendations 0–5 (human)	(6) Clarity 0–5 (human)	(7) Overall 0–5 (human)	(8) Appropriate (AI)	(9) Summary 0–5 (AI)	(10) Literature 0–5 (AI)	(11) Weakness 0–5 (AI)	(12) Recommendations 0–5 (AI)	(13) Clarity 0–5 (AI)	(14) Overall 0–5 (AI)
AI-Assisted	-0.047	0.271*	0.271*	0.271*	0.271*	0.271*	0.271*	0.095	0.228	0.228	0.228	0.228	0.228	0.228
	(0.112)	(0.139)	(0.139)	(0.139)	(0.139)	(0.139)	(0.139)	(0.126)	(0.158)	(0.158)	(0.158)	(0.158)	(0.158)	(0.158)
	[-0.279; 0.186]	[-0.016; 0.557]	[-0.016; 0.557]	[-0.016; 0.557]	[-0.016; 0.557]	[-0.016; 0.557]	[-0.016; 0.557]	[-0.166; 0.356]	[-0.100; 0.555]	[-0.100; 0.555]	[-0.100; 0.555]	[-0.100; 0.555]	[-0.100; 0.555]	[-0.100; 0.555]
AI × Outside-discipline	0.119	0.078	0.078	0.078	0.078	0.078	0.078	-0.092	0.087	0.087	0.087	0.087	0.087	0.087
	(0.129)	(0.221)	(0.221)	(0.221)	(0.221)	(0.221)	(0.221)	(0.168)	(0.206)	(0.206)	(0.206)	(0.206)	(0.206)	(0.206)
	[-0.147; 0.385]	[-0.379; 0.535]	[-0.379; 0.535]	[-0.379; 0.535]	[-0.379; 0.535]	[-0.379; 0.535]	[-0.379; 0.535]	[-0.439; 0.255]	[-0.339; 0.513]	[-0.339; 0.513]	[-0.339; 0.513]	[-0.339; 0.513]	[-0.339; 0.513]	[-0.339; 0.513]
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.464	3.199	3.199	3.199	3.199	3.199	3.199	0.418	3.212	3.212	3.212	3.212	3.212	3.212
Obs.	261	261	261	261	261	261	261	261	261	261	261	261	261	261

Note: Standard errors in parentheses; confidence intervals in brackets.
Controls: Event × article FE; years of coding; software; prior AI familiarity.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Finally, we probe robustness to alternative outcome definitions and specifications. For compactness, the main results table also reports robustness columns (1 and 2 checks), which implement the pre-specified threshold variations. We expect sign and order-of-magnitude stability, with shifts that are interpretable given the alternative codings. These checks complement the design-based safeguards (stratification, fixed effects, pre-specified controls) and help establish that the main conclusions are not an artifact of a particular functional form or threshold.

10 Data Management, Documentation, and Ethics

We will publish de-identified participant-level data, code, and grading rubrics on OSF and GitHub upon acceptance (view-only earlier when necessary). Sensitive logs will be redacted according to the consent form. The study will obtain institutional ethics approval prior to data collection. Any deviations from protocol will be preregistered before accessing outcome data.

Static study materials live inside this repository and will be version-locked before launch: curated task bundles in `Papers/`, the pre-game training deck in `Pre game/`, and the reporting workbook in `Reports/`. Repository history provides provenance for any updates (for example, refreshed replication packages or revised slides); the commit hash distributed to sites is recorded alongside the randomization seed. The same directory structure will be mirrored in the OSF archive so external teams can audit exactly what participants received.

Audio recordings and notes from the post-pilot focus groups are stored on encrypted I4R drives with filenames keyed to anonymous participant IDs and session arm (AI vs. control). Transcription software with local processing is used where feasible; otherwise, trained staff transcribe manually. Only the qualitative analysis team accesses the raw audio. De-identified excerpts linked to codebook categories are released alongside the quantitative replication package after redacting personally identifiable information and any mention of other participants' outputs.

11 Timeline and Deliverables

We plan six one-day events across partner institutions within the academic year, spaced to avoid overlap and allow consistent staffing. Events are pre-announced with a shared protocol: pre-generated randomization within tiers and a seven-hour work window. Each individual may participate in at most one event; duplicate registrations are blocked at check-in. We lock randomization and materials ahead of time and document any deviations (no-shows, substitutions) prior to analysis.

Event and article selection. Events are hosted by partner institutions with capacity for proctoring and secure data handling. Articles are selected from the I4R pipeline to span the three disciplines and software ecosystems (R/Stata/Python), drawing on the catalog documented in Brodeur et al. (2025). Selection prioritizes clarity of a focal result and a mix of tasks with and without known coding pitfalls; each article is tagged to a single discipline before the lock. Because assignments draw randomly from this pool, we do not target a fixed share of tasks with seeded errors within an event. We aim for temporal balance across the academic year to minimize tool-version drift; any updates to AI tooling are logged.

After the sixth event, we finalize the preregistration lock and freeze all code paths before accessing outcome data. The analysis phase proceeds in two stages. First, we produce the pre-specified main results and figures, checking internal coherence and documenting data lineage. Second, we generate the pre-specified secondary and appendix tables to illuminate mechanisms and robustness. All outputs are cross-validated against the preregistered estimands and data checks.

Deliverables include a public replication archive (de-identified individual-level data, code, and grading rubrics), a pre-analysis report that summarizes the locked design and main estimands, and a manuscript integrating results and interpretation. We aim to share preliminary results with partners quickly after the final event and proceed to manuscript submission once the replication archive is complete.

12 Limitations

Despite stratified randomization and event-by-software fixed effects, external validity remains a key limitation. Participating institutions, topics, and software ecosystems may not reflect the broader population of replication exercises or research teams. We minimize site-specific artifacts by controlling for event–software cells and by standardizing instructions and grading rubrics, but context still matters for both the baseline rates and the scope for AI assistance.

Measurement and compliance present additional challenges. Although we combine pledges, spot checks, and audits to monitor AI usage, some noncompliance in the control arm or heterogeneous usage quality in the treatment arm is inevitable. We pre-specify strategies to document and, when necessary, bound any bias (e.g., per-protocol and IV sensitivity), but these strategies trade robustness for different assumptions. Grading, while rubric-based, can also admit residual subjectivity; we address this with clear definitions, double-checks, and rater consensus when needed.

Finally, the evolving nature of AI tools introduces temporal drift. Model updates can affect both capability and interface, potentially shifting the level and composition of gains even with identical prompts. We log model versions and timing, keep baseline instructions identical across events, and emphasize design features (e.g., within-tier randomization) that stabilize inference. Nevertheless, any broader extrapolation should consider how quickly the technology landscape changes and whether the tasks studied here generalize to other domains or longer-horizon research workflows.

13 Appendix

These additional analyses extend and contextualize the main results without altering the primary estimands. We examine moderators beyond the tier-based interactions: a continuous years-of-coding interaction that traces a dose–response (Tables 11 and 12), and within-AI usage intensities (prompts, files, images, words) that help characterize under- and over-use. We also provide usage-by-tier summaries in the AI arm (Table 13) to illustrate whether intensity aligns with the observed treatment heterogeneity. These tables are not substitutes for the main ITT estimands; they are meant to clarify mechanisms and the consistency of patterns.

13.1 AI Referee Prompt

The AI grader receives a standalone instruction each time it evaluates a submission. The block below reproduces the full prompt, including the required output structure and scoring anchors.

Role and objective You are "AI Referee," an expert evaluator for the AI Replication Games. Your task is to read a participant's reproduction package and produce referee-style assessments that mirror the human rubric.

Context - You are judging a replication of a published empirical result. - The participant had access to the original paper, its replication package, and a seven-hour work window. - The materials you receive include the participant's referee report (plain text), a README describing the reproduction steps, a log of detected errors, and any supporting tables or figures they produced.

General instructions - Work independently—do not assume facts that are not present in the materials. - If critical evidence is missing, record that as a weakness and reflect it in the scores. - Base every judgment on the participant's reasoning and evidence; parroting the source paper without evaluation should be penalized. - When assigning numeric scores, use the full 0–5 range

(no half points). - For each scored dimension, write a short justification (two to four sentences) that cites the relevant portion of the participant's submission (quote, paraphrase, or reference to a table or figure).

Evaluation tasks 1. Appropriateness (binary). Return "Appropriate" if the participant correctly characterizes whether the reproduced evidence supports the original claim, flags material discrepancies, and grounds the conclusion in the submitted code or outputs. Otherwise return "Not appropriate". 2. Summary accuracy (0–5). Score the fidelity of the participant's summary of the target paper and reproduced result. 3. Literature placement (0–5). Score how well the participant situates the study within the broader literature, including whether comparisons to related work are accurate and meaningful. 4. Weakness diagnosis (0–5). Score identification and prioritization of substantive weaknesses or threats to validity in the reproduced analysis. 5. Recommendations (0–5). Score the specificity, feasibility, and usefulness of suggested robustness checks or next steps. 6. Clarity (0–5). Score organization, tone, and readability of the write-up. 7. Overall assessment (0–5). Provide a holistic rating that reflects the report's usefulness to an editor deciding on publication, not a mechanical average of the components.

Anchors for 0–5 scales (apply to tasks 2–7) - 0 = Missing, wholly incorrect, or incoherent. - 1 = Substantially inaccurate, with major misunderstandings that would mislead an editor. - 2 = Partially correct but incomplete or containing notable inaccuracies that reduce usefulness. - 3 = Adequate: correct core points with limited depth or weak justification. - 4 = Strong: accurate, well-supported, and actionable with only minor omissions. - 5 = Exceptional: precise, comprehensive, and offering original insight or particularly valuable guidance.

Output format Return JSON with the following fields: `appropriateness` ("Appropriate" or "Not appropriate"), `summary_accuracy`, `literature_placement`, `weakness_diagnosis`, `recommendations`, `clarity`, `overall`, `justifications`, and `notes`. The `justifications` object must contain keys for each of the seven tasks above whose values are the explanatory paragraphs. Include a single string in `notes` describing any missing information, suspected hallucinations, or uncertainties.

```
{
  "appropriateness": "Appropriate",
  "summary_accuracy": 4,
  "literature_placement": 3,
  "weakness_diagnosis": 5,
  "recommendations": 4,
  "clarity": 3,
  "overall": 4,
  "justifications": {
    "appropriateness": "...",
    "summary_accuracy": "...",
    "literature_placement": "...",
    "weakness_diagnosis": "...",
    "recommendations": "...",
    "clarity": "...",
    "overall": "..."
  },
  "notes": "..."
```

```
}
```

Quality control - Double-check that every field is filled. - If evidence is missing for a dimension: - set the score to the literal value `null`, - explain the omission in the notes field, and - flag the gap in the corresponding justification. - Do not invent references or cite external material. - Use only the materials provided with the submission. - Return only the JSON object—no additional prose after the closing brace.

Table 11: Main outcomes with years-of-coding moderator (interaction with AI).

	(1) Reproduction (yrs)	(2) Minutes (yrs)	(3) Minor (yrs)	(4) Major (yrs)
AI-Assisted	0.071	-33.509	0.523	-0.265
	(0.121)	(47.197)	(0.419)	(0.701)
	[-0.179; 0.321]	[-131.143; 64.125]	[-0.298; 1.344]	[-1.640; 1.110]
AI × Years of coding	-0.022	8.676	-0.115	0.059
	(0.019)	(7.318)	(0.077)	(0.134)
	[-0.061; 0.018]	[-6.462; 23.815]	[-0.267; 0.037]	[-0.203; 0.321]
Controls	✓	✓	✓	✓
Mean of dep. var	0.444	247.101	0.421	0.176
Obs.	261	261	261	261

Note: Standard errors in parentheses; confidence intervals in brackets.
Controls: Event × article FE; software; prior AI familiarity.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 12: Referee outcomes with years-of-coding moderator (interaction with AI).

	(1) Appropriate (human) (yrs)	(2) Summary 0–5 (human) (yrs)	(3) Literature 0–5 (human) (yrs)	(4) Weakness 0–5 (human) (yrs)	(5) Recommendation 0–5 (human) (yrs)	(6) Clarity 0–5 (human) (yrs)	(7) Overall 0–5 (human) (yrs)	(8) Appropriate (AI) (yrs)	(9) Summary 0–5 (AI) (yrs)	(10) Literature 0–5 (AI) (yrs)	(11) Weakness 0–5 (AI) (yrs)	(12) Recommendation 0–5 (AI) (yrs)	(13) Clarity 0–5 (AI) (yrs)	(14) Overall 0–5 (AI) (yrs)
AI-Assisted	0.099	0.143	0.143	0.143	0.143	0.143	0.143	0.053	0.136	0.136	0.136	0.136	0.136	0.136
	(0.117)	(0.223)	(0.223)	(0.223)	(0.223)	(0.223)	(0.223)	(0.098)	(0.212)	(0.212)	(0.212)	(0.212)	(0.212)	(0.212)
	[-0.143; 0.340]	[-0.319; 0.605]	[-0.319; 0.605]	[-0.319; 0.605]	[-0.319; 0.605]	[-0.319; 0.605]	[-0.319; 0.605]	[-0.149; 0.255]	[-0.301; 0.574]	[-0.301; 0.574]	[-0.301; 0.574]	[-0.301; 0.574]	[-0.301; 0.574]	[-0.301; 0.574]
AI × Years of coding	-0.015	0.037	0.037	0.037	0.037	0.037	0.037	-0.003	0.030	0.030	0.030	0.030	0.030	0.030
	(0.021)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.019)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)
	[-0.059; 0.029]	[-0.046; 0.121]	[-0.046; 0.121]	[-0.046; 0.121]	[-0.046; 0.121]	[-0.046; 0.121]	[-0.046; 0.121]	[-0.041; 0.036]	[-0.054; 0.115]	[-0.054; 0.115]	[-0.054; 0.115]	[-0.054; 0.115]	[-0.054; 0.115]	[-0.054; 0.115]
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.464	3.199	3.199	3.199	3.199	3.199	3.199	0.418	3.212	3.212	3.212	3.212	3.212	3.212
Obs.	261	261	261	261	261	261	261	261	261	261	261	261	261	261

Note: Standard errors in parentheses; confidence intervals in brackets.
Controls: Event × article FE; software; prior AI familiarity.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 13: Usage by expertise tier in the AI arm (mean and standard deviation).

Tier	Prompts	Files	Images	Words	N
	7.111	1.852	1.444	145.519	
Undergraduate	(4.136)	(1.292)	(1.577)	(52.871)	27
	7.318	2.045	1.091	159.591	
Master's	(3.797)	(2.035)	(1.065)	(68.693)	22
	9.303	1.667	1.333	174.303	
PhD	(3.836)	(1.514)	(1.242)	(60.473)	33
	10.250	1.958	1.625	160.167	
Postdoc	(5.015)	(1.301)	(1.663)	(59.031)	24
	8.185	1.889	1.481	177.593	
Professor	(4.288)	(1.847)	(1.503)	(75.488)	27

14 References

References render from `references.bib`. We will cite prior AI Replication Games and related methodology upon registration finalization.

Brodeur, Abel, Ghina Abdul Baki, Juan P. Aparicio, Bruno Barbarioli, Lenka Fiala, Derek Mikola,

and David Valenta. 2025. “AI-Replication Games.” I4R-DP195. Institute for Replication. https://econstor.eu/bitstream/10419/308508/1/I4R-DP195.pdf.