Reproducing with AI Across the Expertise Ladder

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

2025-09-01

1 Abstract

We will test whether providing large-language-model (LLM) assistance compresses performance gaps across the expertise ladder in computational reproduction tasks. Individuals stratified by expertise (faculty, postdoc/researchers, PhD, master's, undergraduate) are randomized to AI access (ChatGPT Plus with tools) versus human-only controls. Primary outcomes are: (i) successful reproduction of pre-specified results, (ii) time to first success, and (iii) detection of coding errors (major, minor). We pre-specify outcomes, covariates, identification and estimation strategies, heterogeneity, multiplicity control, and robustness. The design aims to identify whether AI is an equalizer (larger gains among lower tiers) or an amplifier (larger gains among higher tiers).

2 Registration and Funding

This pre-analysis plan will be preregistered on OSF at: OSF prereg link (placeholder). Funded by Open Philanthropy and the Institute for Replication (I4R). Code and replication materials will be mirrored on GitHub upon registration.

3 Background and Rationale

Prior multi-site "AI Replication Games" documented sizable effects of AI assistance on success and speed in reproduction tasks. Building on that foundation, this study turns to the vertical dimension of expertise to assess distributional impacts: whether AI narrows performance gaps by disproportionately lifting less-experienced participants (equalizer) or widens gaps by enabling experts to better leverage tools (amplifier). Understanding this distribution matters for pedagogy, workforce development, and equity in research production: if AI compresses the expertise gradient, training and access policies can prioritize broad inclusion; if it amplifies, training needs to emphasize advanced prompting and tool 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. The design isolates intent-to-treat effects within expertise tiers and allows transparent tests of heterogeneity across tiers and along a continuous experience measure (years of coding). To support interpretability, we pre-specify a compact outcome set (levels, timing, error detection, and referee assessments) and a small list of precision-enhancing controls. We also separate descriptive usage evidence (within the AI arm) from the core estimands to avoid conditioning on post-treatment behavior.

Our design choices address two 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. Second, external validity: by spanning multiple events, software ecosystems (R/Stata/Python), and a broad experience range (undergraduate to professor), we gauge how AI assistance interacts with realistic heterogeneity in tools and backgrounds. These features, combined with preregistration and a limited set of pre-specified estimands, aim to balance credibility with informativeness.

4 Research Questions and Hypotheses

- Primary question: Does AI access increase reproduction success rates relative to human-only controls?
- Distributional question: Do treatment effects vary across expertise tiers, and does AI compress the expertise gradient?

Hypotheses: - H1 (Main effect): Access to AI increases the probability of reproducing pre-specified results and reduces time-to-success. - H2 (Compression): Gains are weakly larger in lower tiers (undergraduates, master's) than in higher tiers (faculty, postdoc), implying a reduced difference across tiers in the AI arm. - H3 (Error detection): AI access increases detection of major and minor coding errors.

5 Experimental Design

We recruit participants across five strata—undergraduates, master's students, PhD students, postdocs/researchers, and professors. Each participant completes an individual, timed, one-day reproduction task using the same software as in the original study (R, Stata, or Python). The working window is seven hours. Participants assigned to the AI arm complete a short onboarding to the tools before the event so that any learning curve is minimized during the timed window.

The study has two arms. In the control arm, participants work without AI assistance and explicitly pledge to refrain from using AI tools during the event. In the treatment arm, participants have access to ChatGPT Plus (GPT-4o or successor, including code interpreter and vision). To preserve internal validity, we combine pledges with random screen checks and ex-post audits of chat logs in the treatment arm. Any deviations are documented and, if material to inference, handled with pre-specified per-protocol and instrumental-variable sensitivity analyses.

Assignment is randomized 1:1 between AI and control within each expertise stratum. Randomization is pre-generated and locked before events, and we report balance checks. All analyses include event-by-software fixed effects to absorb site- and tooling-specific differences, and we report realized cell sizes and any deviations (such as no-shows) before analysis.

We target approximately N 300 participants (60 per stratum) across multiple events. With baseline gaps between undergraduates and professors of roughly 15–20 percentage points in success rates and an AI-induced compression of about 40% of that gap, simulations indicate at least 80% power to detect the key interaction at = 0.05. Standard errors are clustered at the event–software level, and we will report finalized assumptions and achieved power prior to locking the analysis.

Randomization and allocation concealment. Randomization is stratified by expertise tier (UG, MA, PhD, PD, P) with 1:1 allocation within each stratum and event. The assignment list is pre-generated by script with a fixed random seed (recorded in the registry) and stored as a read-only file with a timestamped hash. Allocation is concealed from participants and graders until the event begins; the onsite coordinator reveals assignments at check-in. No-shows remain in their assigned arm for ITT; replacements are permitted only before the event starts and are re-randomized using the same stratum-specific seed. Any deviations (swaps across arms or late changes) are documented prior to analysis.

Missingness and analysis sets. The primary analysis set is ITT: all randomized individuals with any outcome data. Outcomes are never imputed. For success and error counts, nonresponse leads to missing outcomes and the observation is excluded from that specific regression but remains for other outcomes. For timing, non-successes are treated as right-censored at the session cap (420 minutes) in survival analyses; we do not impute minutes in OLS. Covariates: categorical controls (tier, software, prior ChatGPT familiarity) include an explicit "Missing" category if needed; continuous controls (years of coding) use median imputation within stratum with a missingness indicator added to the model. We report outcome and covariate missingness by arm and verify robustness to listwise deletion.

Power assumptions. Table 1 summarizes the design inputs used for prospective power calculations; we will freeze any updates to these assumptions prior to registry lock.

Table 1: Prospective power: core design inputs (pre-lock)

Quantity	**Value**
Participants (N)	300 ($\approx$ 60 per tier)
Allocation	1:1 within tier
Clusters (SE)	Event × software (10–20)
Baseline success (control)	0.50 (finalize pre-lock)
Detectable effect on success	$\approx$ 10 pp @ 80% power, $\alpha = 0.05$
Minutes SD (cap)	$\approx$ 60 (420-minute cap)
Multiplicity (primary family)	Holm over four outcomes
Small-cluster inference	Wild-cluster bootstrap if clusters $<$ 30 (9,999 reps)

6 Outcomes and Measurement

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 the quality of their referee report. 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—usage intensity via prompts/files/images/words. 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 2: Outcomes and measurements (primary and secondary).

Category	Outcome	Type	Measurement / Assessment
Primary	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	Qualitative	Appropriate vs. Not appropriate; human judges: A. Brodeur, J. Aparicio, D. Mikola; AI judge separately.
	Referee — score	0–5	0–5 (higher is better); human score is the average 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	Number of prompts (usage logs / self-report; AI arm only).
	Prompt usage — length (AI arm)	Continuous	Length of prompts (usage logs / self-report; 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 3) 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).

Table 3: 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]
Score (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]
Score (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.

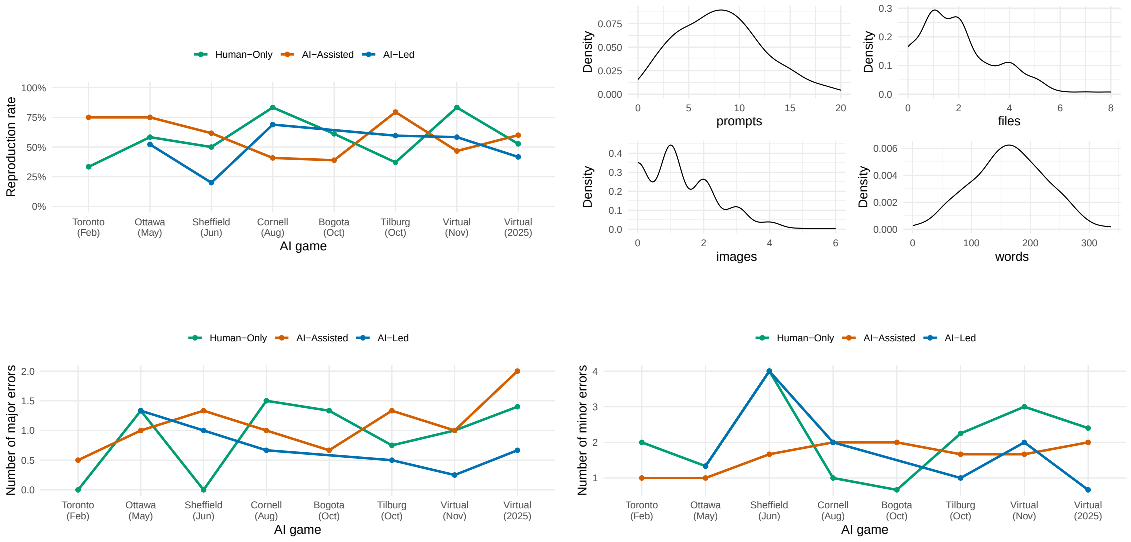

Figure 1: Primary outcomes (levels): reproduction and errors, plus usage context. Notes: Four-panel layout with (top-left) reproduction rates (raw), (top-right) prompt distribution (usage), (bottom-left) major errors (raw), (bottom-right) minor errors (raw). Difference-style plots are intentionally omitted.

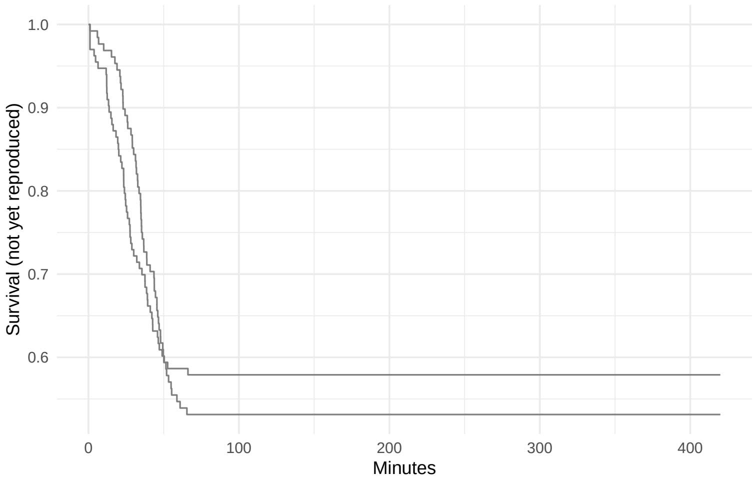

Figure 2: Time-to-success: Kaplan–Meier by arm. Notes: Survival curves stratified by treatment (Control vs. ChatGPT+).

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)

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 4: 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	log(1 + prompts / files / images / words)	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

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

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

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; 0–5 referee score) we use OLS; and for counts (minor/major errors) we fit a Poisson GLM with a log link:

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

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 5) shows means (standard deviations) by arm and Welch tests for the difference in means.

Table 5: 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 usage intensity (prompts/files/images/words) 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

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 6) 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 6: Main effects across outcomes (pre-analysis layout). Standard errors clustered by event–software.

	(1)	(2)	(3)	(4)	(5)	(6)
	Reproduction	Minutes	Minor	Major	At least 1 check	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.

We next examine referee-report outcomes (Table 7), which connect to the communication and assessment dimension of reproduction. We pre-specify models for the binary Appropriateness indicator and the 0–5 Score, each 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 8 and 9; replacing tier with a continuous years-of-coding interaction) to complement the tier-based analysis and trace a dose–response along experience.

Table 7: Referee report outcomes: human and AI assessments. Standard errors clustered by event–software.

	(1) Appropriate (human)	(2) Score 0–5 (human)	(3) Appropriate (AI)	(4) Score 0–5 (AI)
AI-Assisted	0.034	0.288	0.084	0.330*
	(0.149)	(0.218)	(0.134)	(0.191)
	[-0.274; 0.342]	[-0.162; 0.738]	[-0.193; 0.362]	[-0.066; 0.725]
AI × Master's	-0.163	-0.177	0.022	-0.404
	(0.210)	(0.374)	(0.157)	(0.376)
	[-0.599; 0.272]	[-0.951; 0.597]	[-0.303; 0.346]	[-1.182; 0.375]
AI × PhD	0.088	-0.026	-0.150	-0.073
	(0.160)	(0.249)	(0.265)	(0.273)
	[-0.243; 0.420]	[-0.540; 0.488]	[-0.698; 0.399]	[-0.638; 0.492]
AI × Postdoc	0.112	0.151	-0.230	0.149
	(0.254)	(0.377)	(0.244)	(0.387)
	[-0.413; 0.637]	[-0.628; 0.931]	[-0.735; 0.275]	[-0.651; 0.949]
AI × Professor	-0.086	0.268	0.147	0.121
	(0.186)	(0.407)	(0.179)	(0.394)
	[-0.471; 0.299]	[-0.573; 1.110]	[-0.222; 0.517]	[-0.693; 0.935]
Controls	✓	✓	✓	✓
Mean of dep. var	0.464	3.199	0.418	3.212
p-val (Monotonic compression)	0.977	0.713	0.955	0.789
Obs.	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).
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

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.

11 Timeline and Deliverables

We plan five one-day events across partner institutions within the academic year. Each event follows the same protocol: pre-generated randomization within tiers, a standardized onboarding for the AI arm, and a seven-hour work window. We lock randomization and materials ahead of time and document any deviations (no-shows, substitutions) prior to analysis.

After the fifth 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 onboarding consistent 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 8 and 9), 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 10) 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.

Table 8: Main outcomes with years-of-coding moderator (interaction with AI).

	(1)	(2)	(3)	(4)
	Reproduction (yrs)	Minutes (yrs)	Minor (yrs)	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 9: Referee outcomes with years-of-coding moderator (interaction with AI).

	(1)	(2)	(3)	(4)
	Appropriate (yrs)	Score (yrs)	Appropriate (AI, yrs)	Score (AI, yrs)
AI-Assisted	0.099	0.143	0.053	0.136
	(0.117)	(0.223)	(0.098)	(0.212)
	[-0.143; 0.340]	[-0.319; 0.605]	[-0.149; 0.255]	[-0.301; 0.574]
AI × Years of coding	-0.015	0.037	-0.003	0.030
	(0.021)	(0.041)	(0.019)	(0.041)
	[-0.059; 0.029]	[-0.048; 0.121]	[-0.041; 0.036]	[-0.054; 0.115]
Controls	✓	✓	✓	✓
Mean of dep. var	0.464	3.199	0.418	3.212
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 10: 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.