

Generative AI boosts productivity across the expertise ladder

Juan P. Aparicio^{1,2}, David Valenta^{1,2}, Derek Mikola^{1,2}, Ghina Abdul Baki^{1,2}, Bruno Barbarioli^{1,2}, Luna Fazio^{1,2}, Lenka Fiala^{1,2}, Rohan Alexander⁴, Vincent Arel-Bundock¹¹, Bradley Baker¹³, Patrick Davidson¹, Uğurcan Dündar¹⁶, Evelina Gavrilova⁶, Sebastian Gechert³, Amélie Gourdon-Kanhukamwe¹⁸, Alexi Gugushvili¹⁰, Tomas Havranek⁵, Yu-Shiuan Huang⁹, Alexandra Jabbour⁸, Daryl Lee¹⁷, Alexandru Marcoci⁷, Xavier Martin¹², Jost Sieweke¹⁵, Lukas Wallrich¹⁹, Bo Yao²⁰, Myra Yazbeck¹, Jihae You¹⁴, Patrick Chou¹⁸, David Shanks¹⁷, Joseph Tao-yi Wang²¹, Taleen Abraham⁴, Bo Ai²², Farida Akhtar²³, Shumi Akhtar²⁴, Mahmoudou Alioum¹, Said Alqassabi¹⁷, James Alster¹⁷, Micah Altman²⁵, David Angenendt²⁶, Zahra Arjmandi-Lari²⁷, Rebecca Ashong²⁸, Amadeus Bach²⁹, Francisco Bahamonde-Birke¹², Amina Banda³⁰, Teodora Barbu¹⁷, Söhnke M. Bartram³¹, Monica Beeder³², Louis-Philippe Beland³³, Kyriakos Birmpas³⁵, Matthew Blair³⁶, Jack Blumenau¹⁷, Michelle Blunt¹⁸, Johannes Brehm³⁷, Ryan Briggs³⁸, Egor Bronnikov³⁹, Martin Brun⁴⁰, Stephan Bruns⁴¹, Martin Bulla⁴², Navona Calarco⁴, Andrea Calef¹⁷, Giulio Giacomo Cantone⁴³, Natalia Carrillo¹², Edward Chen⁴, Shi Chen³⁶, Yan Y. T. Chen²¹, Zirou Chen⁴⁴, Alec Christie⁴⁵, Navigué Cissé¹, Felix Clouth¹², Jason Collins⁴⁶, Nikolai Cook⁴⁸, Dylan Mercury Cooper¹⁷, Maité Crespo García⁶³, Vasileios Vasilopoulos Cristogiannopoulos¹², Christian Czymara⁴⁹, Thang Dao⁵⁰, Rainier de la Cruz¹, Andrea de Varda²⁵, Kodjo Boris Decker¹, Binyam Afewerk Demena³⁴, Lien Denoo¹², Nency Dhameja⁵¹, Helena Dick²⁶, Max Dieters¹², Veronica Diveica⁵², Trang Dong⁴¹, Andrea Dsouza⁵³, Erwan Dujeancourt⁵⁴, Thibaut Duprey⁵⁵, Mahmoud Elsherif⁵⁶, Ngozi Chioma Ewurum⁵⁷, Kayson Fakhar⁷, Min Fang⁵⁸, Karina Feng⁴, Guidon Fenig¹, Mojtaba Firouzjaeiangalougah⁵⁹, Jack Fitzgerald¹⁵, Emma Footitt¹⁷, Rachel Forshaw⁶⁰, Joris Frese⁶¹, Yin Hei Joelle Fung¹⁷, Sebastian Gallegos⁶², Aurora Lixinhao Gao⁷, Diogo Galdes⁶⁴, Lukas-Daniel Gerhards⁷, Constanza González⁶⁵, Javier Gonzalez⁶⁶, Samantha Grewal⁴, Henri Gruhl⁶⁷, Andy Sabrina Guerrier¹, Karena Guo⁶⁸, Florian Habermann⁶⁴, Marton Hablicsek⁶⁹, Jonathan Hall⁷⁰, Olle Hammar⁷¹, Charlotte Hammer⁷, Adam Hardaker⁷², Christina Sarah Hauser⁷³, Mengxi He¹, Sophie C. F. Hendrikse⁷⁴, David E. Henshall⁷⁵, Alana Hindiye⁷², Senan Hogan-Hennessy⁷⁶, Gareth Hollands¹⁷, William Hope⁴⁸, Amy Huang⁴, Harry Huang⁷, Yuhuan Huang⁷⁷, Zien Huang¹⁷, Jason Huh⁷⁸, Héloïse Hulstaert⁷⁹, Zlatomira Ilchovska⁸⁰, Chris-Gabriel Islam⁸¹, Niklas Itterheim⁸², Joakim Jansson⁸³, Ewa Jarosz⁸⁴, Bingchen Jiang¹⁷, Hayoung Jo¹, Stanley Jones⁸⁵, Valon Kadriu⁷², Eylul Kantar¹⁸, Klára Kantová⁸⁶, Elif Kasgoz⁷, Edmund Kelly⁸⁷, Dongwoo Kim¹⁴, Rudolf Kirchner⁸⁸, Valentin Klotzbücher⁸⁹, Johannes Koenig⁹⁰, Slawomir Konopa⁹¹, Shiko Kudo²¹, Nikolay Kudrin³⁶, Philipp Kulpmann⁹², Konstantin Kyriai²⁷, Simon Ladouce⁷, Po-Hsun Lai²¹, Ryan M.C. Law⁷, Agnes Lenglin⁴, Felix Lewandowski⁹³, Chen Li¹, Shuangyi Li¹⁷, Zhendong Li⁹⁴, Jhen Tang Lin²¹, Piotr Lipinski⁷, Sichang Liu¹², Yuanfei Liu⁷, Ana Elisa Lopez-Miranda⁴, Edoardo Loru⁹⁵, Katharina Loter¹², Chenhui Lu⁹⁶, Phyllis Lun⁴, Martina Lušková⁵, Lala Ma⁹⁷, Meray Maddah⁹⁸, Colin Mang⁹⁹, Josh Martin¹⁸, Theresa Mayrhofer¹⁰⁰, Anthony McCann⁴, Tommy Mcconnell⁷, Fergus McCormack⁷, Ryan McWay¹⁰¹, Beichen Meng¹⁷, Sisi Meng⁷⁶, Simon Merz¹⁰², Diego Mestanza¹⁰³, Meleme Gardine Metch¹, Sumit Mishra¹⁰⁴, Mina Movahedi¹⁰⁵, Lucija Muehlenbachs³⁰, Kondwani Kajera Mughogho¹⁰⁶, Niklas Murken¹⁰⁷, Shinichi Nakagawa¹⁰⁸, Anna Neumann¹⁰⁹, Kim Nguyen¹¹⁰, Linh Ha Nguyen¹¹¹, Son Kien Nguyen⁸⁰, Tri Dung Nguyen¹¹², Tuan Nguyen¹¹³, Viet Nguyen¹¹⁴, Bartłomiej Nowak¹¹⁵, Melvin-Ikaze Nzabonimpa¹, Daneal O'Habib⁵⁵, Victor Okoro¹, Tim Ölkens¹¹⁶, Adrian Ongteco⁴, Andreas Orland⁹², Ömer Özak¹¹⁷, Mikael Paaso¹¹⁸, Pai Chieh-An²¹, Alex Pâtă¹², Marlou N Perquin⁷, Linh Pham¹¹⁹, Anh Phan Cong⁴⁷, Ulrike Phielers¹⁶, Emily Pimenta⁹⁹, Franz Prante³, Dilaanishi Prasad¹⁸, Andi Prasetya⁹⁵, Mathis Preti¹²⁰, Peter Pütz¹²¹, Quan Qi¹²², Ayesha Quddusi⁴, Martin Queißner¹²³, Heiko Rachinger¹²⁴, Letizia Raminelli¹²⁵, Lyba Razzaq¹²⁶, James Read-Tannock¹²⁷, Victoria Reibenspiess¹²⁸, Vivien Reske⁷², Noah Ripstein⁴, Remi Royer¹, Nicolas Rudolf¹²⁹, Orkun Saka¹³⁰, Obaida Salah⁷², Chiara Saponaro¹²⁵, Uma Sarkar⁷, Georg Sator¹²⁷, Martijn Schoenmakers¹³¹, Yannick Schwinger⁷², Raffaello Seri¹³², Rosaly Severijns¹¹¹, Tianbo Shen³⁶, Paul Sibille¹³³, Noah Sidoli¹, Louise Simeonov¹⁷, Noah Spencer⁴, Marie Stadel¹², Georgiana Alexia Stancescu¹², Stefan Staubli¹³⁴, Xiaoyu Sun¹, Scott D. Swain¹³⁵, Noura Sylla¹, Kwasi Tabiri¹³⁶, Tan Yi Swee⁷, Ebnul Tanim³⁰, Georgy Tarasenko⁷⁶, Wijdan Tariq¹³⁷, Mariam Tarraf¹³⁶, Celine Teo⁹⁹, Dylan Thurgood¹³⁸, Tonghui Tian¹³⁶, Emmanuel Tolani²⁶, Estherina Trachtenberg⁸⁸, Tran Hoang Phuong Linh⁵², Amy Tsai²¹, Annie Tubadji¹³⁹, Sparsh Tyagi²⁶, Jakaria Uddin⁷, Ondřej Uldrijan¹⁴⁰, Thi Minh Tuyet Vo¹³⁰, Sofia Anna Voicu¹², Mohammad Wahdan¹⁷, Ruhani Walia⁵⁵, Alexander Wang¹⁴¹, Jo Ling Wang²¹, Zhaoyi Wang⁴, Joe Watson⁷, Matt Webb¹³⁶, Christoph S. Weber¹⁴², Maximilian Weber¹⁴³, Noah Weisswange⁷², Christian Westheide¹⁴⁴, Patrick Wilson¹³⁶, Marcin Wroński¹⁴⁵, Victor Y. Wu¹⁴⁶, Yuexin Xi⁴, Kayla Xiao¹, Chris Xie⁵⁵, Hemu Xu¹⁷, Pranav Yadav¹², Xiliaili Yaolebasi¹⁴⁷, Wendi Yi⁶, Jin Young Yoon³⁶, Lu Yu¹⁷, Tahreen Zahra¹³⁶, Aline Zayat¹, Fanqi Zeng⁸⁷, Huaiyue Zhang¹⁷, Ruining Zhang⁷, Ziwei Zhao¹⁴⁸, Alan Jun Zhu⁴, Jingxuan Zou²⁶, and Abel Brodeur^{1,2}.

Affiliations are listed in Online Appendix Section [S18](#).

Abstract. Generative AI is improving rapidly, but evidence on whether these advances translate into broad productivity gains across skill levels remains limited. We randomly assigned access to generative AI to 286 researchers, ranging from undergraduates to senior academics, during a seven-hour research assignment spanning economics, political science, psychology, and management. Participants reproduced published results, searched for errors, proposed additional analyses, and wrote referee reports. AI assistance increased successful reproduction, the discovery of verified errors, the production of expert-approved robustness checks, and the quality of referee reports. The gains span computational and written work. The results contrast sharply with our experiment two years earlier, in which AI-assisted teams performed similarly to human-only teams on most outcomes. Positive computational interactions suggest larger gains for graduate students and senior researchers than for undergraduates. Generative AI assistance now appears capable of raising productivity across much of the expertise ladder and across multiple dimensions of scientific work.

Significance. Generative AI is advancing rapidly, raising the question of whether its productivity benefits now extend broadly across levels of expertise. In a randomized experiment with researchers ranging from undergraduates to senior academics, AI assistance improved performance across multiple dimensions of scientific work, including reproducing published results, detecting errors, proposing additional analyses, and writing referee reports. Gains appear in both computational and written work, with suggestive evidence of larger computational gains among more advanced researchers. Strikingly, these results contrast with our experiment two years earlier, when AI-assisted teams performed similarly to human-only teams on most outcomes. The findings suggest that the productivity reach of generative AI has expanded rapidly across both tasks and the expertise ladder.

Generative AI has improved rapidly, but evidence on its effects on knowledge-worker productivity is necessarily backward-looking. Experiments conducted only a few years ago evaluated substantially less capable models and tools than those available today. This creates an unusual challenge for understanding the economic consequences of AI: conclusions about where AI helps, which tasks it can perform, and which workers benefit most may change on a much shorter timescale than is typical for new technologies [1–3]. The central question is therefore no longer simply whether generative AI can raise productivity, but how broadly those gains now extend across tasks and levels of expertise.

Existing experiments provide compelling evidence that generative AI can improve performance in writing, customer support, consulting, and software development [4–7]. Many of the earliest studies also suggested that the largest gains accrued to less experienced or initially lower-performing workers, raising the possibility that generative AI could compress productivity differences between workers [4, 5, 8]. Other evidence points in the opposite direction. AI advice produced more favorable effects among stronger-performing entrepreneurs in Kenya [9], while recent experiments show that AI can help professionals perform unfamiliar technical tasks or combine expertise across domains [10, 11]. These findings suggest that AI may substitute for some forms of expertise while complementing others [12–15].

An equally important possibility has received less attention: the distribution of AI’s benefits may itself change rapidly as the technology improves. Capabilities that once provided little advantage in demanding analytical work may become useful as models gain stronger reasoning, coding, file-handling, and tool-use capabilities [16, 17]. Coding agents introduced in 2025, for example, made it increasingly possible to delegate not only the generation of code but also file editing, command execution, debugging, and iterative implementation. If technological progress expands the set of tasks on which AI is useful, productivity gains may increasingly reach workers throughout the expertise distribution rather than being concentrated among particular groups.

We study this question in a randomized experiment with 286 participants across economics, political science, psychology, and management. Each participant received a published study and its research materials, then had seven hours to complete four tasks. They tried to computationally reproduce a key result, find coding errors in the authors’ code, propose two robustness checks, and write a referee report explaining the article’s strengths, weaknesses, and possible improvements. AI access was randomly assigned. Participants with access could decide when and how to use generative AI alongside standard

statistical software and other research tools. The design therefore measures AI assistance across several components of real analytical work, complementing software-engineering benchmarks such as SWE-bench that evaluate language models on resolving issues in existing codebases [18].

AI assistance improved performance across this broad set of research activities. Participants with AI were more likely to computationally reproduce the target result, identified more coding errors, proposed more expert-approved robustness checks, and wrote referee reports that received higher evaluations. The gains therefore extend beyond code generation or writing alone. They appear in executing an existing analysis, diagnosing problems, developing new analyses, and communicating an assessment. AI-assisted participants also reached key computational milestones earlier during the seven-hour assignment.

The breadth of these gains is particularly notable when compared with our earlier AI Replication Games experiment [19]. In events conducted in 2024, AI-assisted teams performed similarly to human-only teams on most primary outcomes and found fewer major errors. In the present experiment, AI-assisted individuals improve on the same core tasks of reproduction, error detection, and robustness analysis, as well as on the newly added referee-report task. Participants, assignments, team structure, and some measurement procedures differ across experiments, so the comparison does not isolate model improvements. The change in the pattern of treatment effects is nevertheless striking. In essentially the same class of demanding scientific work where AI assistance provided little measurable benefit two years earlier, access now produces broad gains.

Our randomization also allows us to test whether the benefits are broadly distributed across the expertise ladder. Gains are not confined to junior or senior researchers, and the evidence does not support widening the gap between assignments inside and outside a participant’s discipline. There is suggestive evidence that graduate students and senior researchers benefit more than undergraduates on computational tasks.

Focus-group discussions provide a possible explanation: participants often used AI to implement analyses they had chosen, troubleshoot code, turn their own ideas into written text, or obtain explanations, while retaining responsibility for deciding which suggestions were useful and which should be rejected. This pattern is consistent with observations of programmers using code-generating assistants both to accelerate an established plan and to explore options when the next step is unclear [20].

Our findings contribute to a broader literature on technological change and the organization of skilled work [1–3, 21–26].

Much of the debate has focused on whether AI substitutes for expertise and therefore disproportionately benefits less-skilled workers, or complements expertise and amplifies existing advantages. Our results suggest that this framing may be incomplete. As AI capabilities expand, the first-order effect may increasingly be that more workers at many levels of expertise can accomplish more. Whether particular groups gain more remains important, especially for the organization of junior work and training, but it is secondary to a broader shift in the reach of AI assistance. In our setting, generative AI has moved from providing limited measurable productivity gains to improving output across multiple tasks and much of the expertise ladder.

Procedures

We conducted five AI Replication Games: a pilot in Ottawa in fall 2025, followed by events in Toronto, Cambridge, and London in spring 2026. London hosted separate events at University College London and King's College London. The analyzed sample contains 286 participants: 37 undergraduates, 117 master's or PhD students, and 132 professors or doctorate-holding researchers, including postdoctoral researchers. Participants came from economics, political science, psychology, and management and were randomly assigned one of 20 published empirical studies from these fields. Each participant received the article and its replication materials and worked individually for seven hours.

The experiment varied two dimensions of the working environment. First, we randomly assigned participants to work either with or without generative AI. AI access was randomized within event and academic tier. Second, for graduate students and senior researchers, the protocol also assigned articles either within or outside the participant's self-reported discipline. An economist, for example, could receive an economics article or an article from political science, psychology, or management. The undergraduate group included final-year students and students about to complete, or who had just completed, their penultimate year. All undergraduates were assigned a study within their own discipline. The design therefore lets us separately examine whether AI assistance raises output across career stages and whether its usefulness depends on prior disciplinary expertise.

Each participant completed four tasks. They first attempted to computationally reproduce a designated result from the study. They then searched the authors' code and data for coding errors, proposed up to two robustness checks, and wrote a referee report assessing the article's strengths, weaknesses, and possible improvements. These tasks span several forms of scientific work: executing an existing analysis, diagnosing problems, developing new analyses, and evaluating and communicating a scientific argument. Participants were instructed to begin with reproduction, but performance on the tasks was assessed separately. Failing to reproduce the target result, for example, did not prevent a participant from identifying an error, proposing a useful robustness check, or writing a strong referee report.

Participation was voluntary and carried no performance-based compensation. Participants were offered coauthorship to the paper independently of assignment or success. A common orientation introduced the tasks and reporting requirements, and a standardized workbook recorded the resulting work (*SI Appendix*).

The analysis includes 286 participants: 146 in the AI-assisted group and 140 in the human-only group. AI-assisted participants received provisioned ChatGPT Plus access and could decide when and how to use generative AI alongside the other tools available to them. They could read the article themselves, inspect the code and data, run statistical software locally, ask AI for assistance, evaluate its responses, and decide what ultimately entered their submission. Such evaluation can require substantial effort: an early Copilot usability study documented obstacles to interpreting generated code, adapting it, and diagnosing its faults [27]. Human-only participants had access to the same ordinary research materials, documentation, and statistical software but pledged not to use generative AI.

Exit responses show substantial separation between the analyzed conditions. We matched exit-survey responses for 143 of the 146 AI-assisted participants; all reported using an AI tool. All 140 analyzed human-only respondents identified their condition as human without AI. Of the 143 treated respondents, 116 selected ChatGPT, including 44 who also selected another tool. The remaining 27 reported other tools without selecting ChatGPT. The intervention therefore measures access to AI assistance in practice rather than the effect of one fixed model.

We organize the experiment into two overlapping studies of expertise. The vertical study examines career stage among participants working within their own discipline: undergraduates, graduate students, and professors or other doctorate-holding researchers. The undergraduates had already spent several years studying their discipline and therefore represent an early stage of professional formation rather than complete novices. The horizontal study asks whether AI assistance works differently when graduate and senior researchers are assigned a study inside versus outside their own discipline. Because undergraduates were always assigned within field, they do not enter this second comparison.

This design also differs from our earlier AI Replication Games experiment [19]. In 2024, the experiment used within-field teams randomized among human-only, AI-assisted, and AI-led conditions. Our cross-study comparison uses the human-only and AI-assisted conditions. The Online Appendix documents the designs and measurement differences.

Finally, four recorded focus groups from the Ottawa pilot event provide qualitative evidence on how participants actually worked with or without AI. The discussions included participants from both treatment conditions and both within- and outside-discipline assignments. We use these accounts to illuminate choices that final submissions cannot reveal, including how participants delegated implementation, sought explanations, evaluated AI suggestions, and decided what to retain or reject.

Four tasks and seven measures

We measure productive output using seven outcomes from four tasks. Participants were asked to computationally reproduce a main result, detect coding errors or data irregularities, design and implement robustness checks, and write an evaluation of the study. Together, these tasks capture several components of empirical research work. We distinguish outcomes related primarily to computational work from those related to analytical proposals and written assessment, while recognizing that the activities overlap: for example, developing and implementing a robustness check can itself require substantial coding.

1. Computational reproducibility

Participants first attempted to computationally reproduce a designated result from the assigned article using the authors' replication materials. They reported whether their output matched the designated numerical result or figure. This task captures completion of an existing analysis and does not require identifying a problem in it. The binary outcome records participant-reported reproduction.

2. Detect coding errors or data irregularities

Participants inspected the supplied code and materials for coding errors and data irregularities, recorded candidate errors, and explained the evidence for each claim. Reviewers assessed those claims against the code, data, and output. Only accepted claims enter the discovery counts; unsupported claims do not.

Four reviewers (lead authors: J.P.A., D.V., D.M. and L.F.) classified accepted errors as minor or major according to their potential importance for the study's results. A major error was one judged capable of affecting the sign, statistical significance, or substantive interpretation of a result. This task therefore yields two outcomes: the number of accepted minor errors and the number of accepted major errors identified by each participant.

Counts are based on accepted error reports rather than distinct underlying error groups, so multiple accepted reports associated with the same underlying problem remain separate observations. The measures capture successful diagnosis rather than the number of suspicions raised or the amount of code inspected. In supplementary analyses, we also examine claims that reviewers rejected and known error groups that participants failed to identify, helping distinguish successful error detection from simply submitting more candidate errors.

3. Propose and implement robustness checks

Participants next proposed up to two robustness checks intended to assess whether the study's conclusions survived reasonable sensitivity analysis. Examples include using an alternative set of control variables or measuring a variable differently. Participants described the motivation for each proposal, the analytical change, and any numerical results.

Twenty designated expert graders, five in each discipline, evaluated proposals and reports about studies in their field. The

graders included professors and doctorate-holding researchers. They served as field specialists, assessing submissions against the published study and explicit criteria. They only graded articles from their own discipline. An approved proposal had to satisfy all four criteria: a clear motivation, a test not already performed in the paper, feasible implementation, and relevance to the empirical strategy. Approval thus means that an expert grader in the same field judged the proposed analysis useful and sufficiently specified. It is neither the participant's own assessment nor a reward for proposing a test that overturns the original result.

We record whether participants produced two approved proposals and whether both included numerical results. The second measure captures numerical reporting, not independently verified implementation or correctness. The Online Appendix also reports results using a threshold of at least one approved proposal.

4. Write a report

Finally, participants wrote a report akin to a referee report. They were asked to explain the study's contribution, identify weaknesses in its argument or evidence, and recommend improvements. The task therefore required more than summarizing the article: reports had to connect its assessment to the study's actual analysis and provide recommendations that an author could act on.

The same twenty designated expert graders assigned each report an overall score from zero to five and separately evaluated summary accuracy, placement in the literature, diagnosis of weaknesses, actionable recommendations, and clarity. We use two main outcomes: the overall score and an indicator for whether every component received a score of at least three, which we call an appropriate report. Requiring all components to meet this threshold prevents strong presentation alone from compensating for weak substantive evaluation, while the overall score provides a complementary holistic assessment.

The four tasks are related but not interchangeable. Finding an error may motivate a robustness check, and a robustness check may become part of the report. Measuring each product separately allows us to observe how AI assistance affects different contributions made by the same researcher.

We retain all outcomes in their original units. This keeps the treatment effects directly interpretable for each type of output without imposing a common scale on reproduction success, error discoveries, robustness proposals, and expert assessments.

Results

We examine average output gains first, then ask who gains more. Table 1 presents the vertical study of academic rank; Table 2 presents the horizontal study of disciplinary match. In each table, Panel A estimates a common AI effect, while Panel B allows gains to differ across expertise groups. Both tables retain all main outcomes in their original units. AI assistance significantly improves every outcome.

Table 1: Vertical study: AI access and academic rank.

	Computational work			Proposals and reports			
	(1) Reproduction (pp)	(2) Minor errors (count)	(3) Major errors (count)	(4) Approved proposals ≥ 2 (pp)	(5) Approved + numerical result ≥ 2 (pp)	(6) Appropriate report (pp)	(7) Overall score (0–5)
A. Without interactions							
AI-assisted	13.7** (5.7)	0.57*** (0.22)	0.14** (0.07)	15.6*** (5.2)	12.4** (5.1)	19.1*** (5.1)	0.92*** (0.22)
Graduate	24.4** (9.6)	0.20 (0.22)	0.10 (0.08)	16.1** (7.0)	12.5* (6.6)	–5.7 (7.7)	0.75** (0.30)
Professor/researcher	24.1** (9.7)	–0.07 (0.21)	0.20** (0.09)	20.4*** (7.0)	18.1*** (6.9)	–2.4 (8.2)	0.77** (0.31)
Sample mean	0.708	0.500	0.109	0.224	0.199	0.183	2.50
Control-group mean	0.643	0.265	0.051	0.153	0.143	0.102	2.08
Observations	202	202	202	201	201	202	202
B. With interactions							
Graduate	9.3 (12.9)	–0.15 (0.32)	–0.00 (0.06)	16.8** (8.4)	12.5 (7.8)	–4.7 (9.8)	0.77* (0.46)
Professor/researcher	10.1 (12.9)	–0.19 (0.30)	0.06 (0.08)	23.6*** (8.2)	23.0*** (8.1)	–1.1 (10.7)	0.68 (0.46)
AI-assisted	–8.7 (16.3)	0.21 (0.35)	–0.05 (0.10)	18.6** (9.3)	16.1* (8.7)	20.9 (13.1)	0.86 (0.54)
AI × graduate	27.2 (17.6)	0.65 (0.53)	0.19 (0.14)	–1.1 (12.4)	0.3 (11.6)	–1.7 (14.7)	–0.03 (0.62)
AI × professor/researcher	27.0 (18.5)	0.21 (0.41)	0.28* (0.16)	–6.3 (13.3)	–9.6 (12.8)	–2.6 (16.1)	0.18 (0.62)
Sample mean	0.708	0.500	0.109	0.224	0.199	0.183	2.50
Control-group mean	0.643	0.265	0.051	0.153	0.143	0.102	2.08
Observations	202	202	202	201	201	202	202

Notes: OLS; robust standard errors in parentheses. Binary coefficients are percentage points (pp); sample and control-group means retain original units and use each column's estimation sample. All models include event, paper, and participant-discipline fixed effects. Both panels use within-discipline assignments and separate graduate and professor/researcher indicators, with undergraduates as the reference. Graduates include master's and doctoral students; senior researchers are professors or doctorate-holding researchers. Panel A estimates a common AI effect. In Panel B, the AI row gives the undergraduate gain; adding the relevant interaction gives that career group's gain. Rank rows in Panel B describe human-only gaps. Stars denote unadjusted * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

AI Assistance raises productivity in technical execution and diagnosis

AI assistance improved performance on both computational reproduction and error detection. In the within-discipline sample, access to AI increased the probability of successfully reproducing the assigned result by 13.7 percentage points. Across both treatment arms, 70.8% of participants in this sample reproduced the focal result, so assistance increased completion even when successful reproduction was already relatively common.

AI-assisted participants also identified more verified coding errors. On average, assistance increased the number of accepted minor errors by 0.57 and major errors by 0.14 per participant. These measures include only error claims that were subsequently verified against the original code, data, and output. Major errors were uncommon: at least one accepted major error was identified in 13 of the 202 within-discipline reports. Taken together, the results show that AI improved both the execution of a script and the diagnosis of problems within it.

Gains extend to analytical proposals and written work

The productivity gains were not confined to computational tasks. AI assistance also improved participants' ability to propose additional analyses and evaluate the assigned article. In the within-discipline sample, access to AI increased the probability of proposing two expert-approved robustness checks by 15.6 percentage points. The probability of producing two approved checks with numerical results increased by 12.4 percentage points. The latter outcome combines expert approval with numerical reporting; it does not independently verify that the proposed analysis was implemented correctly.

These results are not specific to the relatively demanding threshold of two approved proposals. In supplementary within-discipline models, AI increased the probability of producing at least one approved robustness check by 21.9 percentage points and at least one approved check with numerical results by 17.1 percentage points (*SI Table S23*). AI assistance therefore increased both the probability of producing some useful additional analysis and the probability of reaching the more demanding two-proposal threshold.

Participants with AI access also produced reports that received higher ratings from human experts. The average overall score increased by 0.92 points on the zero-to-five scale, while the probability of producing an appropriate report increased by 19.1 percentage points. An appropriate report had to meet the minimum standard on every assessed dimension, including diagnosis of weaknesses and actionable recommendations as well as clarity. The increase therefore cannot be explained by improvements in writing quality alone.

The same broad pattern appears in the sample used to study disciplinary match. Among graduate and senior researchers who could be assigned either within or outside their field, AI increased successful reproduction by 21.1 percentage points and raised the average referee-report score by 0.93 points. The other five outcomes also significantly move in the favorable direction. Because the vertical and horizontal studies share participants, these are not independent replications of the treatment effect. Rather, they show that the broad productivity gains remain visible in the samples used to examine both dimensions of expertise.

Little evidence that AI assistance benefits specific career stages

We next ask whether the gains from AI differ across career stages. Table 1, Panel B, compares graduate students and senior researchers separately with undergraduates, restricting attention to participants working within their own discipline. Each interaction measures whether AI changes the performance gap between the indicated group and undergraduates relative to the gap observed without AI.

The computational outcomes show a consistent descriptive pattern. All six interactions are positive, indicating larger estimated AI gains for graduate students and senior researchers than for undergraduates. For reproduction, the estimated gain from AI is 27.2 percentage points larger for graduate students and 27.0 points larger for senior researchers. The estimated interactions for minor- and major-error discoveries point in the same direction.

We interpret this pattern cautiously. The individual interaction estimates are imprecise, and the results do not establish that senior researchers benefit more than graduate students. Table S6 directly compares AI gains between these two groups. They instead provide suggestive evidence that more advanced training or experience may complement AI on computational tasks. This possibility is consistent with a setting in which users must not only obtain AI-generated code or suggestions but also direct the work, recognize useful outputs, and evaluate whether the resulting analysis is correct.

The pattern is less systematic for robustness proposals and referee reports. Interaction estimates vary in sign, providing little evidence that AI systematically benefits one career stage more than another on these outcomes. Other measures of expertise likewise do not reveal a common gradient: in supplementary models, gains in approved and qualifying proposals decline with years of coding experience, while computational outcomes show no consistent relationship with coding years (*SI Table S8*). The main conclusion is therefore broad productivity

improvement across career stages, accompanied by (weak) suggestive evidence that graduate and senior researchers may gain more on some computational work. *SI Table S2* reports the estimated AI effect separately for each academic tier.

The reproduction interactions are also consistently positive when individual papers or events are omitted. The corresponding interactions for error counts are less stable, particularly in more restrictive samples, and some change sign. *SI Figure A3* reports the full set of component-level estimates across the defined specification families, both including and excluding master's students. Online Appendix Section S13 provides the corresponding model definitions and participant-influence checks.

No evidence that AI gains differ across field assignments

The horizontal study asks whether AI gains differ between assignments inside and outside a researcher's discipline. Table 2, Panel B, restricts attention to graduate and senior researchers, who were eligible for both. The model includes variables for AI assistance and whether the participant is working with an article within or outside their field, while the interaction term gives the additional gain outside field. We find that the interaction estimates vary in sign and show no clear pattern of larger gains inside or outside field.

We find no evidence that AI widens the performance gap between assignments within and outside a participant's field. This comparison concerns the reach of expertise across specializations, and complements the career-stage comparison within a common disciplinary setting.

The focus groups help clarify what "outside field" means in practice. Participants working on studies from another discipline sometimes encountered unfamiliar substantive questions while still recognizing familiar statistical methods, coding structures, or empirical techniques. One outside-field participant described little need for AI to bridge the disciplinary boundary, whereas others reported using AI to better understand the article or to help develop the report. Disciplinary match therefore captures only one dimension of familiarity with the assigned work: researchers may lack substantive expertise in a field while still bringing transferable methodological or technical knowledge to the task.

Computational progress during the assignment

The timing records show both whether participants reached computational milestones and when they reached them. Figure 1 follows three milestones: successful reproduction, the first verified coding error of either severity, and the first verified major coding error. Across both treatments, by the end of the day, 18 of 93 eligible AI-assisted participants had found a major error, compared with 6 of 84 human-only participants. The curves also show earlier progress in reproduction and finding an error of either severity. Each milestone has its own eligible sample, as reported in the figure.

We summarize timing by using the recorded milestone time for participants who reach it and 420 minutes for those who do not. In the unadjusted comparisons, AI assistance lowers this measure by 61.7 minutes for reproduction, 47.1 minutes for the

Table 2: Horizontal study: AI access and disciplinary match.

	Computational work			Proposals and reports			
	(1) Reproduction (pp)	(2) Minor errors (count)	(3) Major errors (count)	(4) Approved proposals ≥ 2 (pp)	(5) Approved + numerical result ≥ 2 (pp)	(6) Appropriate report (pp)	(7) Overall score (0–5)
A. Without interactions							
AI-assisted	21.1*** (4.9)	0.58*** (0.18)	0.16** (0.07)	14.2** (5.6)	15.1*** (5.4)	20.1*** (4.6)	0.93*** (0.19)
Outside discipline	−6.9 (6.9)	−0.42** (0.17)	−0.08 (0.14)	4.6 (7.2)	2.4 (7.3)	−7.5 (7.1)	−0.08 (0.29)
Sample mean	0.767	0.422	0.169	0.270	0.234	0.173	2.65
Control-group mean	0.669	0.194	0.065	0.202	0.161	0.089	2.23
Observations	249	249	249	248	248	249	249
B. With interactions							
Outside discipline	−11.0 (8.3)	−0.28* (0.17)	−0.08 (0.11)	4.3 (8.9)	−3.8 (8.6)	−8.7 (7.4)	−0.07 (0.36)
AI-assisted	18.4*** (5.6)	0.67*** (0.25)	0.17* (0.09)	14.0** (6.5)	11.0* (6.2)	19.3*** (5.6)	0.93*** (0.23)
AI × outside discipline	8.9 (10.8)	−0.30 (0.28)	−0.02 (0.20)	0.7 (12.5)	13.5 (12.1)	2.7 (9.8)	−0.01 (0.40)
Sample mean	0.767	0.422	0.169	0.270	0.234	0.173	2.65
Control-group mean	0.669	0.194	0.065	0.202	0.161	0.089	2.23
Observations	249	249	249	248	248	249	249

Notes: OLS; robust standard errors in parentheses. Binary coefficients are percentage points (pp); sample and control-group means retain original units and use each column's estimation sample. All models include event, paper, and participant-discipline fixed effects. Both panels include graduate and senior researchers assigned within or outside their discipline, controlling for rank, coding years, a missing-coding-experience indicator, and prior GPT familiarity. Panel A estimates a common AI effect. In Panel B, the AI row gives the within-field gain; adding the interaction gives the outside-field gain. The outside-discipline row in Panel B describes the human-only gap. Stars denote unadjusted * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

first verified error, and 30.0 minutes for the first verified major error. The adjusted comparisons have the same direction (*SI Table A1*). These differences combine earlier progress with a greater likelihood of reaching the milestone; they are not estimates of labor time saved.

All three timing estimates retain their direction when individual papers or events are omitted, although their precision varies (*SI Table S24*).

Comparison to 2024 AI games

Our results differ from the first waves of AI Replication Games experiment [19]. In the 2024 events, AI-assisted teams performed similarly to human-only teams on most primary outcomes and found fewer major errors. Here, assisted participants improve across all outcomes, including the new set of report outcomes. Both experiments used published studies and a seven-hour window. The contrast therefore extends beyond the task that we added.

We directly compare the two experiments in Figure 2. This figure compares AI-assisted and human-only reproduction progress after accounting for papers, events, and participant composition within each study. In the current within-field sample, AI reduces time without completion by about 51 minutes over the seven-hour window, assigning the full 420 minutes to noncompleters. This combines earlier completion with a greater chance of completing. Excluding undergraduates preserves the pattern. The earlier experiment shows no clear

reduction on this measure. AI assistance's advantage now persists through much of the day.

Several explanations are plausible. First, the two studies differ along several dimensions, including team structure (*SI Appendix Section SI2*).

The available tools changed as well. Coding agents released in 2025 broadened opportunities to delegate file editing, command execution, and iterative work [28, 29]. While the 2024 experiment already allowed file uploads, Python execution, and internet access, changes in models and tools are plausible contributors to the improved performance of AI-assistance treatment arm. Neither experiment, however, isolates the contribution of a particular capability or interface.

Tools and the people using them can improve together [30]. In a retrospective discussion of the earlier experiment, one participant described both greater personal skill with AI and a more capable model [19]. The current debriefs likewise describe participants choosing what to delegate, providing context, and revising requests after inspecting the output. These accounts suggest that useful assistance depends on how people organize work around the available technology [20]. The experiments measure assistance in successive working environments; they do not separate user learning from technological progress. *SI Appendix Section SI6* documents the source accounts.



Figure 1: Computational progress during the seven-hour assignment. Share of eligible participants reaching each milestone over time. A true error is a claim accepted by the error reviewers; the middle panel includes minor and major errors, and the right panel includes major errors only. Labels show discoveries or completions relative to each sample size. Error samples require a paper with at least one accepted error of the relevant severity in the pooled adjudication record, plus usable timing records. Participants who do not reach the milestone remain in the sample through 420 minutes. The curves describe the eligible samples. *SI Table A1* reports timing estimates, *SI Table S24* checks individual paper and event exclusions, and *SI Figure S2* separates field assignments.

Focus groups reveal the work around AI assistance

The focus groups conducted right after the conclusion of the AI pilot games in Ottawa provide a closer view of how participants incorporated AI into their work. They do not establish how common particular views were. Instead, they illustrate where AI entered the research workflow: implementing participant-chosen analyses, troubleshooting technical problems, evaluating candidate errors, developing written assessments, and filling gaps in knowledge. *SI Table A3* provides more detail on the sessions and timestamps.

Several accounts describe selective delegation rather than wholesale transfer of the task to AI. One participant chose a robustness check, supplied the relevant data and existing code, and asked AI to write and revise the implementation until the code ran. Another received an AI-suggested check but did not implement it. This distinction echoes an account from our earlier experiment, in which an AI-assisted participant described receiving generic robustness suggestions and returning to the paper to develop the actual checks. Across these examples, AI could help translate an analytical decision into executable work, but deciding what analysis was useful remained a separate step.

AI was also used to overcome concrete technical obstacles. One participant reported that code failed because required software packages were missing; AI identified the dependencies and supplied installation commands. Another viewed similar installation advice as excessive for a simple request, while a third found AI more useful when presented with a specific error message than when asked broadly to search for errors.

These accounts suggest that assistance was often most useful when participants could formulate a well-defined problem for the tool.

Verification remained distinct from generation. Participants described checking suspected errors against the original code and rejecting AI suggestions that did not survive their own inspection. One participant spent substantial time investigating suggestions that were ultimately discarded, leaving less time for robustness analysis. AI could therefore expand the set of candidate problems a researcher considered without eliminating the need to determine whether a problem was real. The path from an AI response to a submitted finding continued to involve human judgment.

The same pattern appeared in writing. One outside-field participant performed familiar computational work without AI but used it to turn personal notes into referee-report prose. Others described using AI to restructure ideas they already knew how to express substantively, then revising generated language when it did not preserve the intended meaning. By contrast, another participant found that an AI-generated report merely summarized the article without identifying useful criticisms or recommendations. This distinction mirrors our outcome measures: clearer prose can improve communication, but a strong report also requires diagnosis and actionable recommendations.

Participants also used AI to compensate for gaps in familiarity. A self-described earlier-stage student sought explanations of methodological concepts and help evaluating a proposed robustness check, while an outside-field participant asked why an unfamiliar article used particular methods. AI assistance

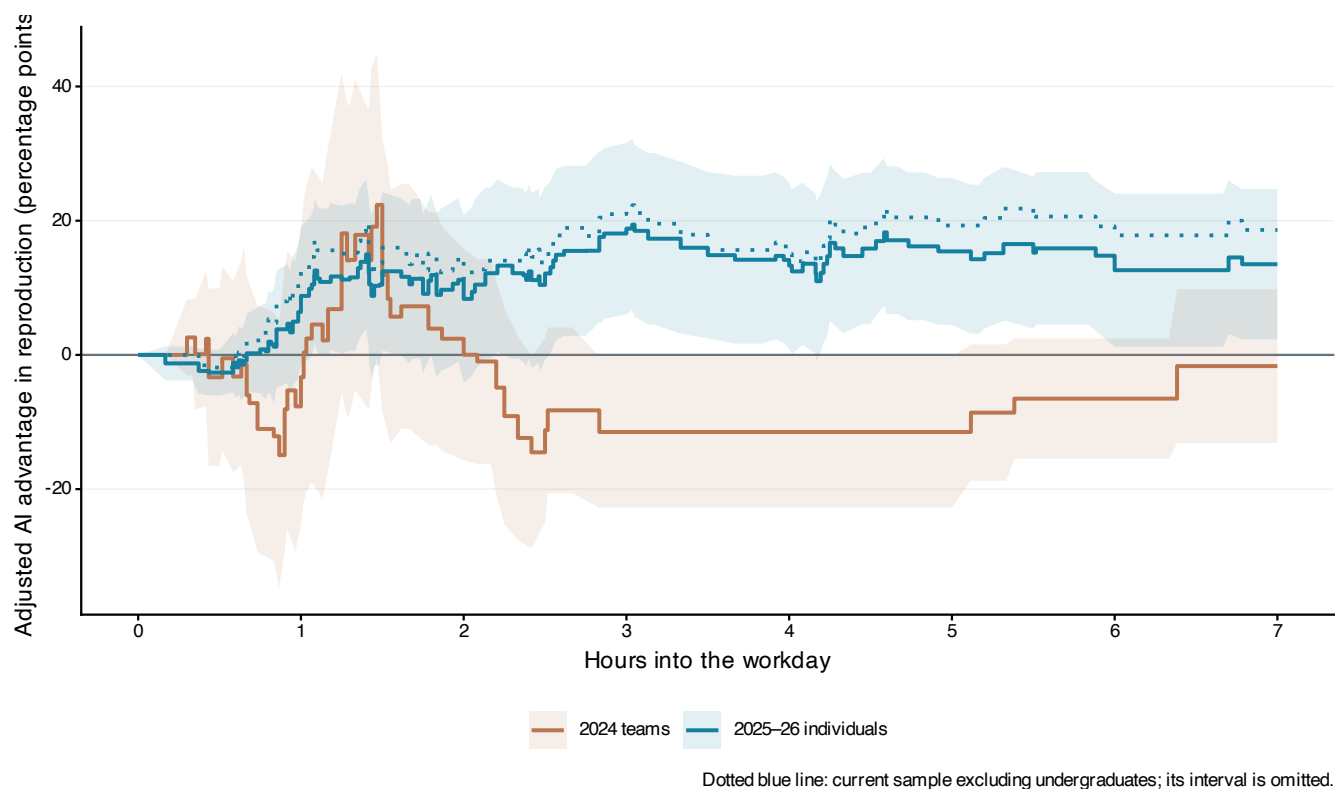


Figure 2: The AI advantage across the workday. Lines show adjusted differences in the probability of reproducing the assigned result by each elapsed minute, comparing AI-assisted with human-only participants. Separate regressions include paper and event fixed effects. The 2024 sample contains 68 teams and controls for team size, maximum and minimum academic position, and attendance. The current sample contains 200 within-field individuals with usable timing records and controls for academic rank. Shading gives pointwise 95% robust confidence intervals, not a simultaneous band. The dotted line excludes the 37 undergraduates, leaving 163 participants; its interval is omitted for readability. The timing sample differs from the 202-observation main-regression sample because it requires usable timing records. Noncompletion contributes 420 minutes to the restricted-time summary. *SI Figures A1* and *A2* report observed progress and final-output comparisons.

therefore entered both the application of existing expertise and the acquisition of information that participants lacked.

Human-only participants, meanwhile, were not necessarily working without assistance of any kind. One described disabling an AI coding assistant embedded in an editor, while others sought help from the internet. The comparison condition therefore reflects conventional access to documentation, whereas the treatment added an immediately available source of implementation, explanation, and feedback.

Taken together, the debriefs portray AI less as a substitute for the researcher than as an additional layer in the workflow. Participants could delegate execution, troubleshooting, explanation, or drafting, but they still had to decide what to ask, determine whether an answer was useful, and choose what ultimately entered their analysis or report. These accounts do not identify the mechanism behind the treatment effects, but they illustrate how broad productivity gains can arise while researcher expertise and judgment remain consequential.

Discussion

Generative AI assistance increased what participants accomplished during a common working day across several dimen-

sions of scientific work. These gains were visible across career stages. The breadth of the effects is especially notable in light of our 2024 experiment, in which AI-assisted teams performed similarly to human-only teams on most outcomes. The two experiments differ in participants, articles, team structure, and available tools, so the contrast cannot be attributed to model progress alone. Nevertheless, in a closely related class of demanding analytical tasks, AI assistance now contributes to substantially more of the work.

AI assistance also changed participants' experience of the work. Assisted participants reported greater enjoyment of the event and each task, and rated their expected performance more highly. Coding and proposing robustness checks showed the largest enjoyment differences, about 20 points each. These responses cover the four spring events and complement the measured output gains: participants accomplished more and experienced the work more positively (*SI Table S27*).

The results also speak to an ongoing question about who benefits most from generative AI. Several early experiments in writing, customer support, and consulting found larger gains among workers with lower initial performance or less experience [4–6]. Other evidence points to complementarity

between AI and stronger prior performance or expertise [9, 31]. Our results do not reveal a single expertise gradient. Average gains are broad, while the computational outcomes provide weak suggestive evidence that graduate students and senior researchers may benefit more than undergraduates. Broad productivity gains and convergence in performance are therefore distinct possibilities: AI assistance can raise output throughout the expertise distribution without necessarily reducing differences between workers.

Even within computational work, existing evidence does not imply a universal relationship between AI and expertise [32]. AI has accelerated bounded programming tasks and increased output among professional developers, whereas experienced open-source developers working on familiar repositories were slower with early-2025 tools [7, 33, 34]. Our assignment combines several activities that are difficult to separate: understanding an empirical study, navigating an existing codebase, executing and modifying analyses, and deciding whether a suspected problem is substantively important.

The focus groups help illustrate how those activities interacted with AI assistance. Participants often delegated implementation, troubleshooting, explanation, or drafting while retaining control over what to pursue and what to accept. A participant could choose a robustness check and use AI to implement it, or arrive with an assessment and use AI to express it more clearly. At the same time, participants rejected proposed errors, discarded unhelpful suggestions, and described generated referee reports that summarized an article without offering useful criticism. AI could therefore reduce the cost of turning an idea into code or prose without eliminating the need to decide whether the resulting work was useful [35].

This distinction between generation and evaluation is consistent with a broader literature on human–AI complementarity. Strong AI performance does not automatically translate into strong combined human–AI performance [36–39], and evaluating an AI response can itself require knowledge and effort [40–43]. Our qualitative evidence does not identify a causal mechanism, and cannot be reliably linked to career stage or individual outcomes. Participants also used AI to learn unfamiliar concepts and methods, showing that assistance can help acquire missing knowledge as well as apply existing expertise.

These findings raise questions about how junior work and training may change as AI becomes more capable. Tasks traditionally assigned to junior researchers often serve two purposes: they contribute to production and they provide opportunities to acquire skills. If AI reduces the cost of execution for experienced workers, the productive role of some junior tasks may change even when their training value remains [44–47]. Our experiment does not measure hiring, wages, career progression, or learning over time. It therefore motivates studying how entry-level work and training adapt rather than predicting a particular labor-market response.

Several limitations qualify the results. The experiment covers a single seven-hour assignment among volunteers in four academic disciplines. We measure completed and assessed output within that window, not output per hour of effort,

monetary returns, repeated use, later learning, or the downstream scientific value of the work. Finally, the comparison with our 2024 experiment is informative but not longitudinal: participants, papers, team structure, and some measurement procedures differ across studies.

Within these limits, the main result is broad. 2026 generative AI improves performance across several components of demanding analytical work and across much of the expertise ladder. As the technology changes, understanding its productivity effects will therefore require tracking not only whether AI works, but also how the range of tasks and workers for which it is useful continues to expand.

Materials and Methods

Participants, assignment, and analysis samples

The experiment compares AI-assisted and human-only work across academic ranks and fields. Participants completed the same four tasks described above during a seven-hour session. The pre-analysis plan and its amendment are available in the [GitHub repository](#); the *SI* documents the deposit history and departures from the planned procedures.

AI access was randomly assigned under a protocol specifying assignment within event and academic tier. The analysis includes 286 participants across five events: Ottawa (43), Toronto (63), Cambridge (87), University College London (40), and King’s College London (53). There are 37 undergraduates, 117 graduate students, and 132 professors or doctorate-holding researchers. The undergraduate group included final-year students and students about to complete, or who had just completed, their penultimate year. Graduate students include master’s and PhD students; the senior category includes postdoctoral researchers, faculty, and other doctorate-holding researchers. Each participant worked on one of 20 articles, five from each of four journals: *American Economic Journal: Applied Economics*, *American Journal of Political Science*, *Psychological Science*, and *Management Science*.

The vertical and horizontal studies use overlapping samples. The vertical study (Table 1) compares academic ranks among the 203 participants working within their discipline. After dropping one participant who is alone in a fixed-effect category, these models contain 202 observations for computational and referee-report outcomes and 201 for robustness outcomes. The horizontal study (Table 2) compares within-field and outside-field assignments among 249 graduate and senior researchers, including 83 assigned outside their field. Only these ranks were eligible for randomized outside-field assignment; undergraduates always received an inside-field article. These models contain 249 observations for computational and referee-report outcomes and 248 for robustness outcomes.

Disciplinary match compares the participant’s self-reported primary discipline with the assigned article’s field. The protocol specified random allocation of articles and outside-field assignments among graduate and senior participants. Panel B therefore tests whether the gain from AI differs across field assignments; it does not estimate the causal effect of moving a

participant to another field.

The main analysis follows communicated AI access. One participant received AI-assisted instructions from an organizer despite a human-only entry in the saved randomization sheet. We classify that participant with the AI-assisted group. The original-record and leave-one-out sensitivity checks appear in *SI Table S7*; the vertical sample is unaffected because this participant worked outside their discipline.

Outcome assessment

Participants reported whether their output reproduced the designated numerical result or figure from the assigned article. This measure records their assessment of the match, rather than an independent expert audit.

Four reviewers (lead authors J.P.A., D.V., D.M., and L.F.) assessed candidate coding errors and data irregularities against the cited code, data, and output. Unsupported claims were classified as false positives. Accepted errors were classified as major or minor according to their importance for the focal result. The review guide permits a major classification when correcting the problem could change the result's sign, statistical significance, or substantive interpretation. The *SI* explains how this rule differs from the analysis plan's requirement for demonstrated impact. Main counts sum accepted error records; counting distinct underlying defects preserves the direction of the average effects and career-stage interactions (*SI Table S20*). False negatives count known error groups that the participant did not identify.

Twenty designated expert graders, including professors and doctorate-holding researchers, were assigned one article each within their field, with five graders per field. Each assessed proposals and reports about that article using the grading rubric. The retained sample contains 442 check reviews and 275 referee-report reviews, all matched to the grader's paper and field. The *SI* documents the materials supplied to graders and the separate reproduction and error-review procedures. The current review cards omit explicit treatment and participant identifiers. The *SI* distinguishes the current interface from what can be established about historical blinding.

Expert-approved robustness checks meet four criteria: clear motivation, feasibility, novelty relative to the original article, and relevance to the validity of its empirical strategy. Approved checks with numerical results additionally require a digit in a result or coefficient field, screened for explicit nonexecution or year-only entries. We therefore measure numerical reporting rather than independently verified execution or correctness.

Experts graded the report along several dimensions: summary accuracy, literature placement, weakness diagnosis, recommendations, and clarity, plus an overall score. Appropriateness requires every component to score at least three. This constructed outcome differs from the direct binary judgment specified in the pre-analysis plan. Each item has one final expert assessment in the analysis. The archive does not contain repeated independent ratings, so it cannot be used to measure agreement between graders.

All submitted reports have expert grades. Eleven participants submitted no referee text and receive zeros; these are

distinct from expert-assigned scores. One participant's retained file contains only the report sheet, leaving their robustness outcomes unknown. We exclude that observation only from robustness models. Assigning the unknown binary outcomes either zero or one leaves the AI assistance effects positive (*SI Table S17*).

Estimation and expertise comparisons

Main estimates use ordinary least squares with robust standard errors. Models include separate event, article, and participant-discipline fixed effects. The vertical study models (*Table 1*) use within-discipline assignments and separate graduate and senior indicators, with undergraduates as the reference group. The interacted models include an AI assistance interaction for each of these two groups. Horizontal study models (*Table 2*) use eligible graduate and senior researchers and include rank, coding years, an indicator for missing coding-experience responses, and prior GPT familiarity. Full specifications, alternative estimators, and multiplicity sensitivity for the seven outcomes appear in the *SI*.

For participant i , the vertical model is

$$Y_i = \alpha + \beta A_i + \gamma_G G_i + \gamma_R R_i + \delta_G A_i G_i + \delta_R A_i R_i + \lambda_{e(i)} + \mu_{p(i)} + \nu_{d(i)} + \varepsilon_i.$$

Here A_i indicates AI access, G_i graduate status, and R_i senior-researcher status. Undergraduates are the reference group. The model uses within-field assignments. The fixed effects capture event e , assigned article p , and home discipline d .

Among graduate and senior researchers, the horizontal model is

$$Y_i = \alpha + \beta A_i + \theta O_i + \delta A_i O_i + X_i' \eta + \lambda_{e(i)} + \mu_{p(i)} + \nu_{d(i)} + \varepsilon_i,$$

where O_i indicates an outside-field assignment. The controls X_i contain rank, coding experience, an indicator for missing coding-experience responses, and prior GPT familiarity. In each main table, Panel A omits the AI interactions and Panel B includes them.

The planned event-by-article fixed effects created singleton cells and reduced sample retention. Separate fixed effects retain more participants. Some binary-response models also exhibited separation, and some count models had unstable estimates. We therefore use OLS for the main comparisons and report the alternative specifications and model diagnostics in the Online Appendix.

Interactions compare treatment gains across academic tiers and field assignments. Sixteen reports lack coding years; horizontal models use event-by-tier medians with an overall fallback, plus a missingness indicator. Complete-case checks instead omit participants without recorded coding years and re-estimate the same models. Coding experience is measured before the task. Supplementary pooled coding-experience analyses and their complete-case checks are presented in the *SI*.

Component interactions and sensitivity analysis

The pre-analysis plan specified the vertical and horizontal expertise questions. The main tables report seven outcomes separately, with graduate students and senior researchers each compared with undergraduates. Main-table tests are two-sided and unadjusted for multiple comparisons. The *SI* reports component-only adjustments across the fourteen vertical interactions and seven horizontal interactions, separately and together.

The component specification curves are supplementary analyses informed by the baseline results. They vary article and event inclusion, overlap restrictions, fixed effects, treatment slopes, and baseline covariates. Each family is estimated with and without master's students. The analysis record lists every attempted model. Participant-deletion checks are reported separately. These exercises describe sensitivity to analytical choices rather than the frequency of independent replications.

Timing and the analysis record

Timing uses elapsed minutes from event-specific scheduled starts. Restricted time equals the observed milestone time or 420 minutes for a non-event. We examine the first accepted error of either severity and the first accepted major error separately. An error sample includes participants. The *SI* reports sample rules, timing estimates, and article and event exclusions.

Recorded focus groups

Four recorded small-group debriefs from Ottawa document how participants approached the work. Recorded introductions describe twelve participants across sessions totaling approximately 129 minutes. We coded timestamped passages around how participants directed work, used assistance, and checked answers. We added themes that emerged from the discussions, retained contrary accounts, and distinguished the day's experience from prior practice and speculation. The analysis used AI assistance and was informed by the quantitative results. The recordings include participant introductions, but the transcripts do not reliably identify the speaker of every later contribution. We therefore do not link individual passages to academic rank or experimental outcomes. These accounts develop hypotheses about AI-assisted work without estimating thematic prevalence or causal mediation. The *SI* documents sources, coding, and the outstanding consent and human source-review requirements.

Data and materials

The pre-analysis plan and amendment are available in the [GitHub repository](#). A selected Appendix follows the references, containing the central supporting figures and tables. A separate *SI* contains detailed methods, further robustness checks, full model variants, grading instructions, and the protocol record. The public data and code release remains to be completed.

References

- [1] David H. Autor. Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3):3–30, 2015. doi: 10.1257/jep.29.3.3.
- [2] Daron Acemoglu and Pascual Restrepo. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30, 2019. doi: 10.1257/jep.33.2.3.
- [3] David H. Autor. Applying AI to rebuild middle class jobs. Working Paper 32140, National Bureau of Economic Research, 2024.
- [4] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586.
- [5] Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044.
- [6] Fabrizio Dell'Acqua, Edward McFowland, III, Ethan R. Mollick, et al. Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2):403–423, 2026. doi: 10.1287/orsc.2025.21838.
- [7] Kevin Zheyuan Cui, Mert Demirel, Sonia Jaffe, Leon Musolf, Sida Peng, and Tobias Salz. The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers. *Management Science*, 2026. doi: 10.1287/mnsc.2025.00535. Articles in Advance, published February 27, 2026.
- [8] Anil R. Doshi and Oliver P. Hauser. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290.
- [9] Nicholas G. Otis, Rowan Clarke, Solène Delecourt, et al. The uneven impact of generative artificial intelligence on entrepreneurial performance: Evidence from a field experiment in kenya. *Management Science*, 2026. doi: 10.1287/mnsc.2024.06909. URL <https://pubsonline.informs.org/doi/10.1287/mnsc.2024.06909>. Articles in Advance, published July 10, 2026.
- [10] Emma Wiles, Lisa Kraye, Mohamed Abbadi, et al. GenAI as an exoskeleton: Experimental evidence on knowledge workers using GenAI on new skills. Working Paper 4944588, SSRN, 2024.
- [11] Fabrizio Dell'Acqua, Charles Ayoubi, Hila Lifshitz, et al. The cybernetic teammate: A field experiment on generative AI and teamwork. *Organization Science*, 37(4):1217–1242, 2026. doi: 10.1287/orsc.2025.20702. URL <https://pubsonline.informs.org/doi/10.1287/orsc.2025.20702>.
- [12] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8:2293–2303, 2024. doi: 10.1038/s41562-024-02024-1.
- [13] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5:46–57, 2023. doi: 10.1038/s42256-022-00593-2.

- [14] Keren J. Kanarik, Wojciech T. Osowiecki, Yu (Joe) Lu, Dipongkar Talukder, Niklas Roschewsky, Sae Na Park, Mattan Kamon, David M. Fried, and Richard A. Gottscho. Human-machine collaboration for improving semiconductor process development. *Nature*, 616:707–711, 2023. doi: 10.1038/s41586-023-05773-7.
- [15] Alex Davies, Petar Veličković, Lars Buesing, et al. Advancing mathematics by guiding human intuition with AI. *Nature*, 600: 70–74, 2021. doi: 10.1038/s41586-021-04086-x.
- [16] Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651:914–919, 2026. doi: 10.1038/s41586-026-10265-5.
- [17] Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, et al. An AI system to help scientists write expert-level empirical software. *Nature*, 654:909–916, 2026. doi: 10.1038/s41586-026-10658-6.
- [18] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024. URL <https://arxiv.org/abs/2310.06770>.
- [19] Abel Brodeur, David Valenta, Alexandru Marcoci, et al. AI-assisted teams outperform AI-led teams but not human-only teams in assessing research reproducibility in quantitative social science. *Proceedings of the National Academy of Sciences*, 123(22):e2524747123, 2026. ISSN 1091-6490. doi: 10.1073/pnas.2524747123.
- [20] Shraddha Barke, Michael B. James, and Nadia Polikarpova. Grounded Copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages*, 7(OOPSLA1), 2023. doi: 10.1145/3586030. URL <https://doi.org/10.1145/3586030>.
- [21] David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4): 1279–1333, 2003. doi: 10.1162/003355303322552801.
- [22] Daron Acemoglu and David Autor. Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics*, volume 4B, pages 1043–1171. Elsevier, 2011. doi: 10.1016/S0169-7218(11)02410-5.
- [23] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308, 2024. doi: 10.1126/science.adj0998.
- [24] Guillermo Cruces, Diego Fernández Mejjide, Sebastian Galiani, Ramiro H Gálvez, and María Lombardi. Does generative ai narrow education-based productivity gaps? evidence from a randomized experiment. Technical Report 34851, National Bureau of Economic Research, 2026.
- [25] Leonardo Gambacorta, Han Qiu, Shuo Shan, and Daniel M Rees. Generative ai and labour productivity: A quasi experiment on coding. *Journal of Financial Stability*, 84:101543, 2026. doi: 10.1016/j.jfs.2026.101543.
- [26] Morgan R. Frank, David Autor, James E. Bessen, Erik Brynjolfsson, Manuel Cebrian, David J. Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, Dashun Wang, Hyejin Youn, and Iyad Rahwan. Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116:6531–6539, 2019. doi: 10.1073/pnas.1900949116.
- [27] Priyan Vaithilingam, Tianyi Zhang, and Elena Glassman. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7. Association for Computing Machinery, 2022. doi: 10.1145/3491101.3519665. URL <https://doi.org/10.1145/3491101.3519665>.
- [28] Anthropic. Claude 3.7 Sonnet and Claude Code. Research preview announcement, February 24 2025. URL <https://www.anthropic.com/news/claude-3-7-sonnet>.
- [29] OpenAI. Introducing Codex. Research preview announcement, May 16 2025. URL <https://openai.com/index/introducing-codex/>.
- [30] Minkyu Shin, Jin Kim, Bas van Opheusden, and Thomas L. Griffiths. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, 120:e2214840120, 2023. doi: 10.1073/pnas.2214840120.
- [31] Feiyang Yu, Alex Moehring, Oishi Banerjee, Tobias Salz, Nikhil Agarwal, and Pranav Rajpurkar. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine*, 30:837–849, 2024. doi: 10.1038/s41591-024-02850-w.
- [32] Simone Daniotti, Johannes Wachs, Xiangnan Feng, and Frank Neffke. Who is using AI to code? global diffusion and impact of generative AI. *Science*, 391:831–835, 2026. doi: 10.1126/science.adz9311.
- [33] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirel. The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv preprint arXiv:2302.06590, 2023.
- [34] Joel Becker, Nate Rush, Elizabeth Barnes, and David Rein. Measuring the impact of early-2025 AI on experienced open-source developer productivity, 2025. Version 2, July 25, 2025.
- [35] Lingfei Wu and Bogdan Vasilescu. AI raises the productivity bar. *Science*, 391:763–764, 2026. doi: 10.1126/science.aef5239.
- [36] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the most accurate AI the best teammate? Optimizing AI for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414, 2021. doi: 10.1609/aaai.v35i13.17359.
- [37] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pages 1526–1533, 2020. doi: 10.24963/ijcai.2020/212.
- [38] Sebastian Raisch and Kateryna Fomina. Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Academy of Management Review*, 50(2):441–464, 2025. doi: 10.5465/amr.2021.0421. URL <https://journals.aom.org/doi/abs/10.5465/amr.2021.0421>.

- [39] Ethan Goh, Robert Gallo, Jason Hom, et al. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969, 2024. doi: 10.1001/jamanetworkopen.2024.40969.
- [40] Linda J Skitka, Kathleen L Mosier, and Mark Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999. doi: 10.1006/ijhc.1999.0252.
- [41] Gagan Bansal, Tongshuang Wu, Joyce Zhou, et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16, 2021. doi: 10.1145/3411764.3445717.
- [42] Helena Vasconcelos, Matthew Jörke, Madeleine Grundle-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023. doi: 10.1145/3579605.
- [43] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. doi: 10.1145/3449287.
- [44] Gary S. Becker. Investment in human capital: A theoretical analysis. *Journal of Political Economy*, 70(5, Part 2):9–49, 1962. doi: 10.1086/258724.
- [45] Yoram Ben-Porath. The production of human capital and the life cycle of earnings. *Journal of Political Economy*, 75(4, Part 1):352–365, 1967. doi: 10.1086/259291.
- [46] Matthew Beane. Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly*, 64(1):87–123, 2019. doi: 10.1177/0001839217751692.
- [47] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakçı, and Rei Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025. doi: 10.1073/pnas.2422633122.

Appendix

This appendix collects the evidence closest to the main findings. Figures A1 and A2 develop the comparison with the earlier experiment. Figure A3 shows the computational interactions across specifications. Tables A1, A2, and A3 report timing, sample balance, and participant accounts. Detailed methods and further checks follow in the Online Appendix.

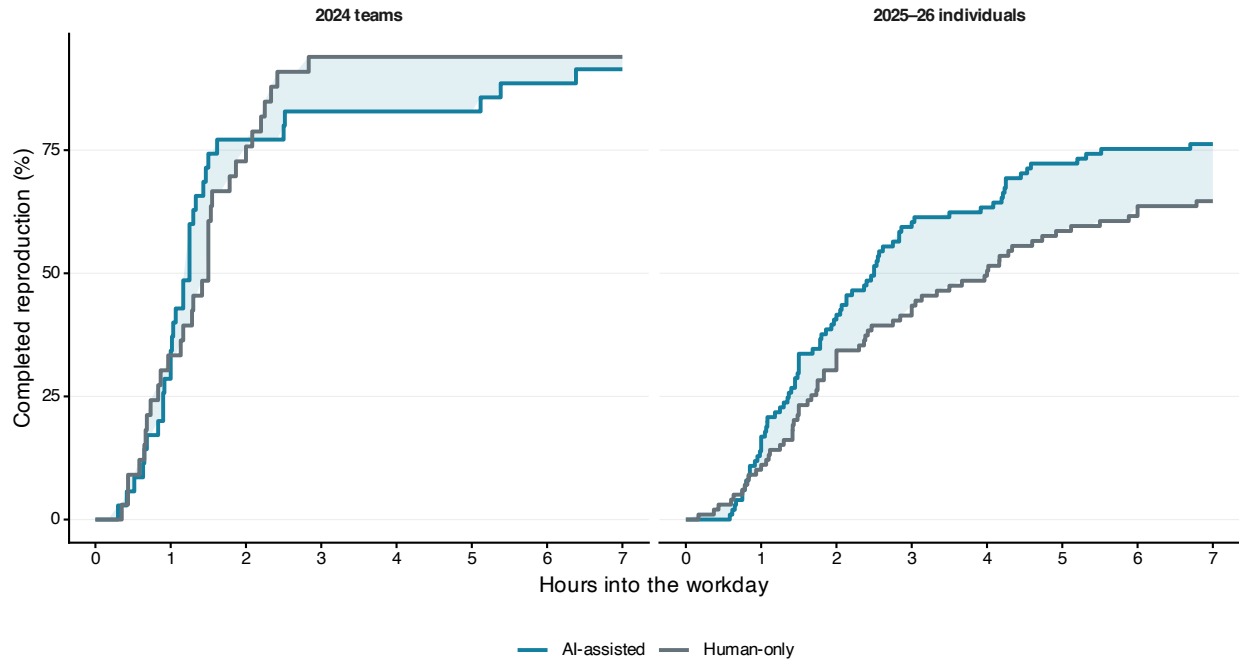


Figure A1: Observed reproduction progress in the two experiments. Curves show the share completing reproduction by each elapsed time. Shading marks the difference between arms. The samples match Figure 2; the curves are unadjusted. Participants who do not complete remain in the denominator throughout the seven-hour window.

Table A1: Timing of three computational milestones.

Milestone	Comparison	N	AI effect (min)	95% CI
Reproduction	Unadjusted	282	-61.7	[-95.4, -27.9]
Reproduction	Within discipline	199	-50.0	[-85.1, -14.9]
Reproduction	Graduate and senior	245	-78.3	[-109.4, -47.3]
First true error	Unadjusted	257	-47.1	[-77.4, -16.8]
First true error	Within discipline	189	-57.9	[-90.0, -25.8]
First true error	Graduate and senior	224	-63.7	[-97.4, -30.1]
First true major error	Unadjusted	177	-30.0	[-57.2, -2.7]
First true major error	Within discipline	125	-33.2	[-64.6, -1.7]
First true major error	Graduate and senior	154	-39.6	[-76.3, -3.0]

Notes: Participants who do not reach the milestone receive 420 minutes; others receive their recorded time. Negative estimates combine earlier progress with a greater chance of reaching the milestone, rather than measuring labor time saved. Within-discipline models adjust for academic rank. Graduate and senior models adjust for field assignment, rank, coding years, missing coding years, and prior GPT familiarity. Both adjusted comparisons include event, paper, and participant-discipline fixed effects. Intervals use robust standard errors. The first true error includes either severity; the first true major error uses major errors only. Error samples require an adjudicated opportunity and usable timing records.

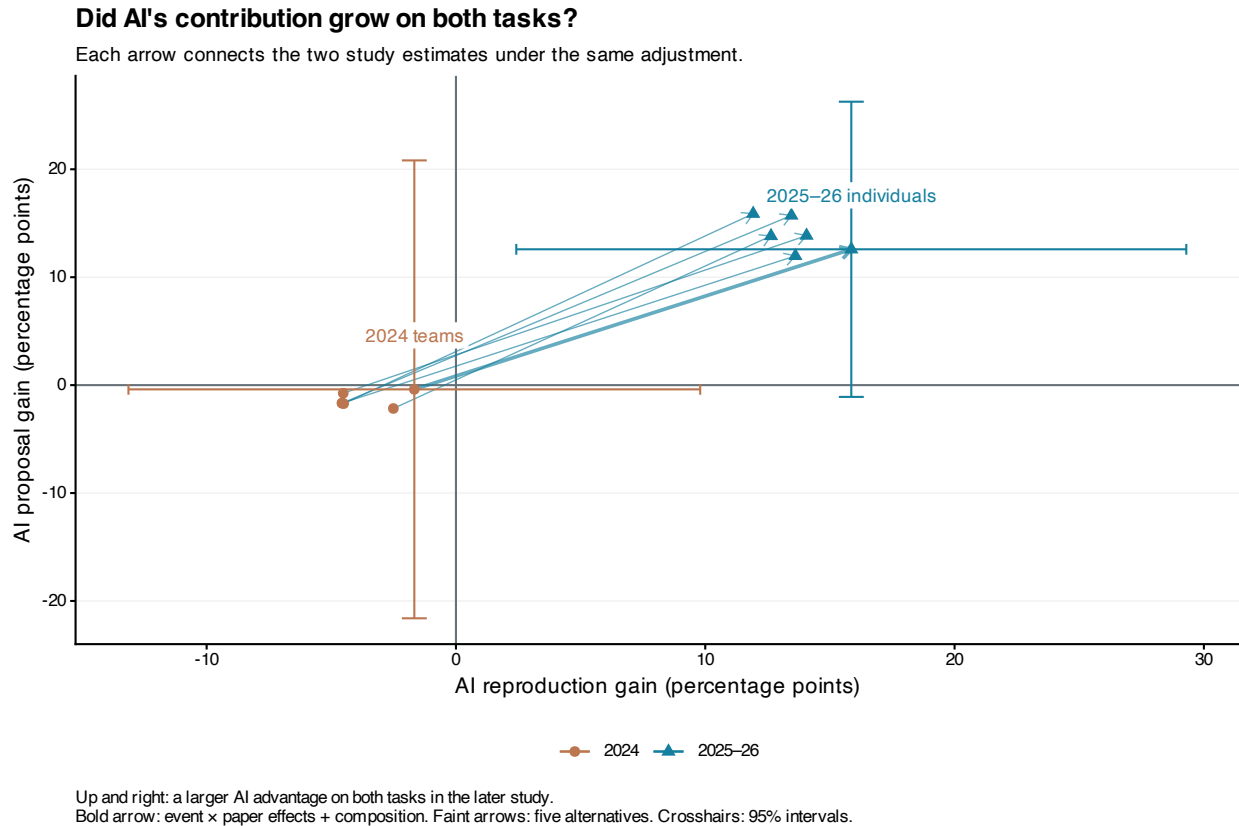
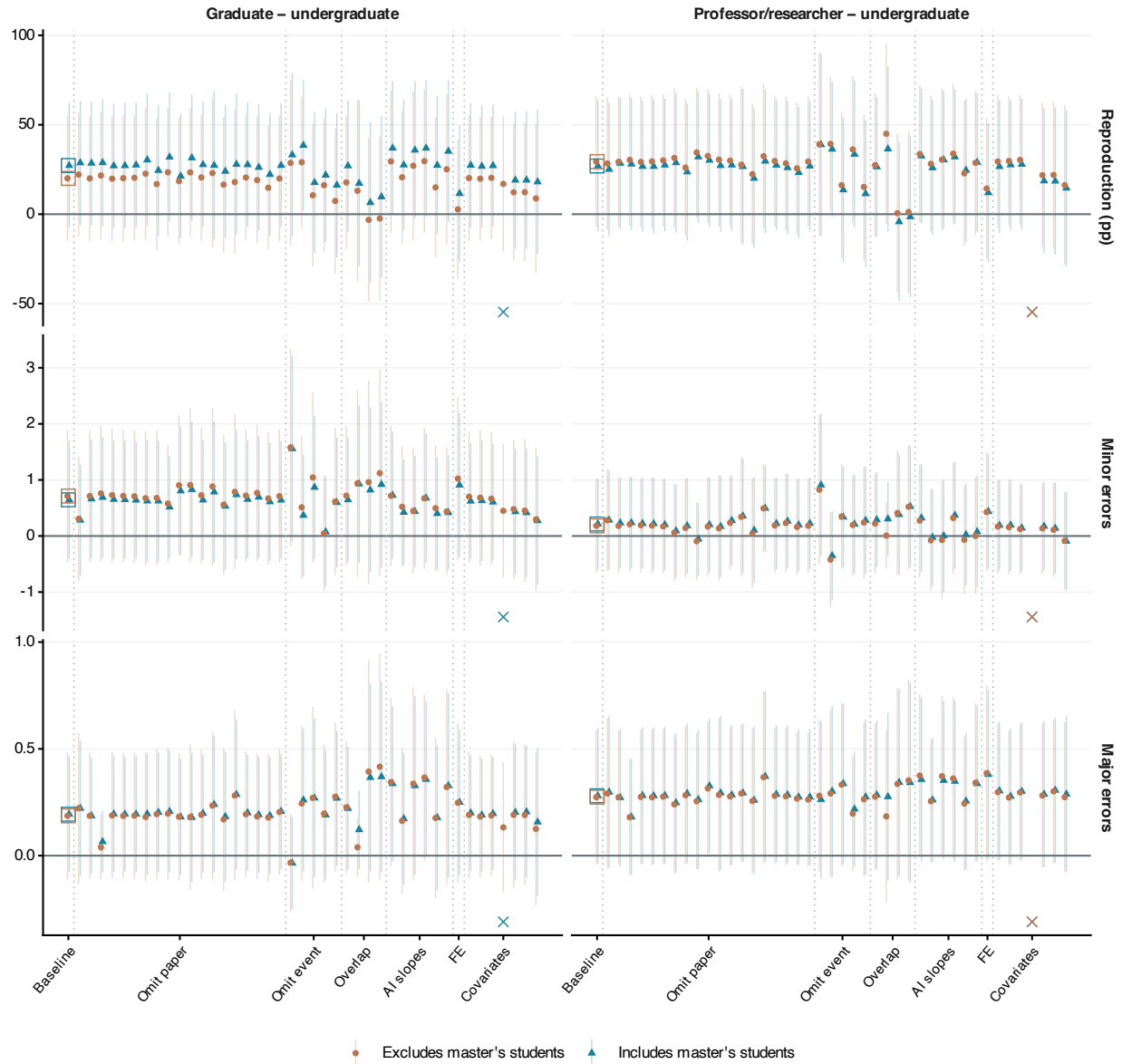


Figure A2: Did AI's contribution grow on both tasks? The horizontal axis gives the AI effect on reproduction and the vertical axis the effect on two approved proposals, both in percentage points. An arrow moving up and right means that assistance added more on both tasks in the later study. Each arrow connects two study estimates under one adjustment set, not the same people over time. The bold path uses event-by-paper fixed effects and study-specific composition controls; crosshairs show marginal 95% confidence intervals for that specification. Faint paths use no covariates, paper effects, event effects, paper plus event effects, or event-by-paper effects. All fits use the two relevant treatment arms. The earlier sample contains 68 teams; current within-field samples contain 202 observations for reproduction and 201 for proposals. These final-output samples differ from the timing samples. Proposal grading and team structure differ across experiments. Paths are dependent sensitivity checks, not independent replications.

Do computational expertise gaps survive alternative specifications?

Above zero: larger AI gains than for undergraduates.



Squares: baseline. Crosses along the lower edge: non-estimable models. Intervals: 95% CIs.
Panels retain sign reversals; reproduction is more directionally stable than error counts.

Figure A3: Do computational expertise gaps survive alternative specifications? The six panels show the additional AI gain for graduates and professors/researchers relative to undergraduates, separately for reproduction and discoveries of minor and major errors. A point above zero means a larger AI gain than for undergraduates. Blue includes master's students; orange excludes them. Squares mark the main specification. Vertical lines are pointwise 95% intervals; crosses at the bottom identify contrasts that cannot be estimated and are not zero estimates. Labeled blocks separate the specification families; dotted lines mark their boundaries. The order is fixed across panels: baseline (1), paper exclusions (2–20), event exclusions (21–25), common-support restrictions (26–29), assignment-specific AI slopes (30–35), event-by-rank fixed effects (36), and baseline-covariate or complete-case alternatives (37–43). All 507 identified structural contrasts from 516 attempts are shown. Nine contrasts are unidentifiable. Reproduction is more directionally stable than the error-count outcomes. These checks preserve separate career groups and use the original component outcomes.

Table A2: Sample balance, missingness, and assignment-to-report coverage**Panel A. Observed pretreatment characteristics**

Variable	Human N	Mean (SD)	AI N	Mean (SD)	AI – H	Welch <i>p</i>
Years coding	129	8.12 (5.77)	141	7.91 (6.48)	-0.20	0.787
Prior AI frequency (0-100)	135	62.53 (25.61)	144	67.17 (24.09)	4.64	0.121
Undergrad share	140	0.11 (0.32)	146	0.14 (0.35)	0.03	0.457
Grad share	140	0.41 (0.49)	146	0.41 (0.49)	0.00	0.948
Researcher share	140	0.48 (0.50)	146	0.45 (0.50)	-0.03	0.573
Field: Economics	140	0.56 (0.50)	146	0.68 (0.47)	0.13	0.026
Field: Management	140	0.06 (0.25)	146	0.07 (0.25)	0.00	0.887
Field: Political Science	140	0.13 (0.34)	146	0.05 (0.23)	-0.07	0.032
Field: Psychology	140	0.25 (0.43)	146	0.19 (0.40)	-0.06	0.237
GPT: Some share	137	0.72 (0.45)	143	0.62 (0.49)	-0.09	0.099
GPT: Heavy share	137	0.28 (0.45)	143	0.38 (0.49)	0.09	0.099

Panel B. Missing outcomes and covariates

Measure	Human missing / 140	AI missing / 146
Coding outcomes (each)	0 (0.0%)	0 (0.0%)
Robustness outcomes (each)	0 (0.0%)	1 (0.7%)
Human referee outcomes (each)	0 (0.0%)	0 (0.0%)
AI referee outcomes (each)	0 (0.0%)	0 (0.0%)
Coding years	11 (7.9%)	5 (3.4%)
GPT familiarity	3 (2.1%)	3 (2.1%)
Prior AI frequency	5 (3.6%)	2 (1.4%)
Exit-survey linkage	0 (0.0%)	3 (2.1%)

Panel C. Saved assignments, analyzed reports, and survey linkage

Event	Arm	Saved	Linked	Reports	Intake	Exit	Overrides
Ottawa	Human	46	16	20	19	20	5
Ottawa	AI	47	22	23	23	20	6
Toronto	Human	49	32	33	32	33	9
Toronto	AI	48	30	30	30	30	7
Cambridge	Human	60	39	40	40	40	3
Cambridge	AI	62	44	47	47	47	4
UCL	Human	48	21	22	22	22	5
UCL	AI	48	16	18	18	18	2
KCL	Human	38	25	25	25	25	2
KCL	AI	41	28	28	28	28	2

Panel D. Human referee content and grading

Arm	Report / grade	N	Unknown grade
Human	No text; rubric assigns zero	7	0
Human	Text and expert review	133	0
AI	No text; rubric assigns zero	4	0
AI	Text and expert review	142	0

Notes: Panel A reports observed means/SDs in the analyzed sample; categorical means are shares and Welch tests are unadjusted for strata/multiplicity. No respondent reports None for GPT familiarity. Panel B leaves unknown grades missing; missingness rates do not establish missing-at-random. Baseline software preference is unverified. Saved assignments comprise 487 participant-event rows and 474 distinct emails, with no within-event email duplicates. Exact event/email matching links 274 of 286 reports; one saved arm differs from the communicated assignment described in Section [SII](#). The 213 saved rows without a retained analyzed report are not automatically dropouts: eligibility, attendance, late changes, and submissions remain unreconciled. Three linked rows lack paper titles, rather than establishing paper switches. Eleven Toronto Other/Unsupported discipline entries were recoded through author-provided corrections recorded on May 19, 2026; independent source verification remains undocumented. Overrides are reconciliation flags, not necessarily reassignments. Human grading covers all 275 submitted referee reports; 11 participants submitted no text and receive rubric zeros. One robustness outcome remains unavailable because the retained participant file contains only the referee-report sheet. Survey linkage does not objectively measure compliance.

Table A3: Participant debriefs: implementation, judgment, and verification

These accounts illustrate how participants used and evaluated AI during the assignment. They provide context for the quantitative findings rather than estimates of how common each experience was.

Panel A. Recorded corpus: Ottawa, October 16, 2025		
Session	Recorded assignment	Participants; duration
FG01	Human-only; within discipline	3; 30:46
FG02	AI-assisted; outside discipline	2; 30:45
FG03	Human-only; outside discipline	3; 44:07
FG04	AI-assisted; mixed discipline	4 (3 within, 1 outside); 23:29
Panel B. Reported experiences and contrasting accounts		
Theme / sources	Reported experience	Counterexample or interpretive limit
Selective delegation E14, E19, E20, E53, E54, E60	AI turned human notes into prose, helped restructure intended points, and wrote code for participant-chosen robustness checks.	An AI-generated check remained unimplemented; one generated report offered summary without useful critique. Running code is not expert approval.
Verification E43, E44, E45, E46, E52	Participants rejected suspected errors after code inspection. One described less time left for robustness work.	Actual error messages or missing dependencies elicited useful repairs; another account judged installation suggestions excessive. Rejected suggestions are not submitted false positives; perceived time costs are not causal estimates.
Practical familiarity E11, E15, E55, E57, E58	Familiar software and methods could transfer across fields, while scientific interpretation still required additional understanding.	Broad field match did not guarantee topic familiarity. A student also reported help with concepts. These accounts do not establish an expertise interaction.
Human substitutes E03, E07, E26, E39, E65	Controls described removing usual AI tools and seeking manual or peer assistance, sometimes without the needed expertise nearby.	Peer help could be useful or unavailable. Some valued unaided work and regarded randomization as fair. These accounts establish neither universal reliance nor compliance.
Task allocation E04, E28, E47, E51, E66	Downloads, reading, and reporting-template uncertainty consumed attention before or alongside analytical work.	Other packages were small, clearly documented, and easy to run. Recalled time and complaints do not replace recorded outcomes or establish protocol breaches.

Notes: All entries are paraphrased accounts, not verified performance outcomes. Participant counts and assignments follow introductions and discussion; the roster remains unverified. Source IDs resolve to recording times and stable transcript-segment ranges in the archived coding workbook. Key passages: E14, FG02 04:21–05:31; E19–E20, FG02 14:26–16:47; E45–E46, FG04 05:22–06:42; E43–E44, FG04 03:57–05:21; E52–E54, FG04 11:16–13:05; E57, FG04 13:47–15:24; E60, FG04 19:11–19:36; E65, FG03 22:25–23:50; E66, FG03 12:21–13:03. E56 is excluded because its transcription is ambiguous. The analysis does not estimate thematic prevalence or treatment effects. The thematic review is described in [Online Appendix Section S16](#).

Online Appendix

The appendix brings together the experimental materials, comparisons across studies, additional analyses, participant debriefs, and protocol record.

OA1. Experimental design and outcome measurement	21
OA2. AI assistance across successive experiments	25
OA3. Expertise and treatment heterogeneity	27
OA4. Robustness of the quantitative findings	36
OA5. Additional outcomes and computational timing	57
OA6. Participant debriefs and reported AI use	65
OA7. Analysis plan, protocol record, and grading instructions	68
OA8. Author group and affiliations	74

SI1. Experimental design and outcome measurement

This section describes the assignments, participants, and expert assessments underlying the seven outcomes.

Research questions and hypotheses

The pre-analysis plan (PAP) lists three research questions and the following hypotheses: (i) H1 posits that access to AI increases research and coding skills—raising the probability of a successful reproduction while lowering major and minor coding errors; (ii) H2 mirrors this logic across disciplines, anticipating that AI shrinks the outside-of-discipline penalty for all outcome variables; H3 posits that AI assistance improves coding and research performance.

The main tables retain seven outcomes; [Table S22](#) reports AI-referee outcomes and [Table S23](#) reports the one-check thresholds. Continuous coding-years moderation was planned as supplementary analysis. Its pooled, tier-adjusted implementation was selected after examining results and is retained here as supplementary evidence. Timing and component decompositions are additional analyses.

Recruitment and exclusions

The Institute for Replication advertised the events through Bluesky, X, direct email, and its event webpage (<https://www.i4replication.org/games>). Participants included final-year undergraduates (or about to complete penultimate year), graduate students (master and PhD), and professors or doctorate-holding researchers. The analysis includes reports with at least one usable outcome. Missing outcomes are not imputed. Missing coding years use event-by-tier medians, with the overall median as fallback, and an additive missingness indicator. Panel B of both main tables controls for prior GPT familiarity using indicators for heavy familiarity and missing responses, with some familiarity as the reference category.

Event schedule

The five events took place in four cities, with separate London events at University College London (UCL) and King's College London (KCL). Their seven-hour windows were Ottawa, October 16, 2025, 09:00–16:00 Eastern; Toronto, April 29, 2026, 09:00–16:00 Eastern; Cambridge, May 11, 2026, 10:00–17:00 London time; UCL, May 13, 2026, 09:00–16:00 London time; and KCL, May 14, 2026, 09:00–16:00 London time. Timing outcomes were converted to elapsed minutes from each event's local start and censored at 420 minutes.

Participant documents and survey

Each participant completed a standardized workbook. Its main sheet recorded identifiers, event, assignment, software, and discipline match. Separate sheets captured coding errors, reproduction status and time, up to two robustness checks, and the referee report. Coding-error entries included the time identified, affected element, explanation, and the participant's proposed severity. The post-event survey recorded treatment confirmation, AI tools used, self-reported AI-use intensity, perceived performance effects, useful or missing features, and open comments.

Assignment and reported AI use

The analysis contains 146 AI-assisted and 140 human-only participants. Matched exit responses are available for 143 and 140, respectively. The assignment classification follows the instructions participants received. One participant was told to use AI by an organizer working from an older sheet, although the saved randomization record listed human-only. Their exit answers consistently describe AI-assisted work. We retain this participant in the AI-assisted group and preserve the original recorded assignment for the sensitivity analysis in [Table S7](#).

Orientation and AI access

The one-hour orientation covered event logistics, the workbook, deliverables, documentation, timestamps, and report expectations. Participants saw an overview of the ChatGPT Plus tools available to the treated arm. AI-assisted participants received provisioned ChatGPT Plus access; human-only participants pledged not to use generative AI. Exact model availability varied with the product interface and date and is not used as an analytical variable.

Coding-error adjudication

Errors and data irregularities reported by participants were reviewed by the lead authors. Reviewers inspected the cited code, data, and output and classified individual candidate-error records. Main counts sum the records retained as true discoveries, rather than counting distinct underlying error groups. Group identifiers are used separately when constructing false negatives. Unsupported claims are classified as false positives. Major and minor labels are the final reviewers' classifications of the defects' importance for the focal result. The reviewer guide permits a major classification when a correction could change its sign, statistical significance, or substantive interpretation. The pre-analysis plan's narrower operational definition required demonstrated impact. The retained labels do not uniformly document a rerun establishing that impact. Explicit group adjudications and the final cleaning rules determine the analysis counts. False negatives count distinct known true-error groups in the assigned article that a participant did not identify.

Expert assignment and assessment materials

Twenty designated expert graders were assigned one paper each within their registered discipline, five per field. Every canonical proposal and referee review matches the grader's assigned paper and field. Fourteen of the expert graders also participated in the experiment, all in the professor/researcher category. A comparison of grader and participant records found no cases of graders assessing their own work. Each proposal and report has one final expert assessment; items were not independently scored by multiple graders.

Proposal cards supply the paper and its PDF, the participant's motivation, targeted element, specification and sample changes, numerical fields, and a matched exhibit image when available. Referee cards supply the paper and submitted report. The twenty graders' role differs from reproduction auditing and the coding-error review workflow. The latter includes candidate claims, evidence pointers, reviewer classifications, and subsequent explicit adjudications. Auxiliary AI assessments in the review workflow are not counted as independent human raters.

Robustness-check adjudication

Expert graders scored each proposed check on four binary criteria: clear motivation, feasibility, novelty relative to the original article, and relevance to the validity of the empirical strategy. A proposal was expert-approved only when it satisfied all four criteria. A qualifying check additionally required numeric content in the result field; successful execution was not independently established by that rule. Main outcomes use the at-least-two thresholds; the individual criteria and other secondary measures appear in [Table S21](#).

Report grading

Expert graders scored summary accuracy, literature placement, weakness diagnosis, actionable recommendations, clarity and style, and an overall assessment on zero-to-five scales. The expert guide defines literature placement as situating the article in relevant research, theory, debate, or contribution. It emphasizes accuracy but does not require a reference-by-reference existence check. The recorded scores therefore measure expert-assessed literature placement, not a systematic citation audit. The binary appropriateness outcome equals one only when each of the five component scores is at least three. This constructed rule prevents a strong average from masking a deficient component. It differs from the PAP's planned direct binary judgment, which was not collected as such. The local AI grader applied the archived rubric once

to redacted narrative fields. Its estimates provide an alternative assessment of report quality in [Table S22](#). Agreement with human-score treatment effects does not independently validate either grading procedure. Sample balance and coverage appear in *SI Table A2*.

Study inventory

The 20 articles were:

Economics, *American Economic Journal: Applied Economics*

1. What Impacts Can We Expect from School Spending Policy? Evidence from Evaluations in the United States
<https://doi.org/10.1257/app.20220279>.
2. The Evolution of Local Labor Markets After Recessions
<https://doi.org/10.1257/app.20220132>.
3. Expected Returns to Crime and Crime Location
<https://doi.org/10.1257/app.20220585>.
4. The Willingness to Pay for a Cooler Day: Evidence from 50 Years of Major League Baseball Games
<https://doi.org/10.1257/app.20220606>.
5. Better Alone? Evidence on the Costs of Intermunicipal Cooperation
<https://doi.org/10.1257/app.20230344>.

Political science, *American Journal of Political Science*

1. Public Participation and Constitutional Compliance
<https://doi.org/10.1111/ajps.70000>.
2. Austerity, Economic Vulnerability, and Populism
<https://doi.org/10.1111/ajps.12865>.
3. A Spatial Rogowski Theory of America's 19th-Century Protectionism
<https://doi.org/10.1111/ajps.12870>.
4. Biased Bureaucrats and the Policies of International Organizations
<https://doi.org/10.1111/ajps.12921>.
5. Expertise Acquisition in Congress
<https://doi.org/10.1111/ajps.12848>.

Psychology, *Psychological Science*

1. Universal Constraints on Linguistic Event Categories: A Cross-Cultural Study of Child Homesign
<https://doi.org/10.1177/09567976221140328>.
2. The Development of Spatial Cognition and Its Malleability Assessed in Mass Population via a Mobile Game
<https://doi.org/10.1177/09567976221137313>.
3. Contactless Real-Time Heart Rate Predicts the Performance of Elite Athletes
<https://doi.org/10.1177/09567976221143127>.
4. Impaired and Spared Auditory Category Learning in Developmental Dyslexia
<https://doi.org/10.1177/09567976231151581>.
5. The Impact of Implicit-Bias-Oriented Diversity Training on Police Officers' Beliefs, Motivations, and Actions
<https://doi.org/10.1177/09567976221150617>.

Management, *Management Science*

1. Attention Constraints and Learning in Categories
<https://doi.org/10.1287/mnsc.2023.4803>.
2. Does Decision Making for Others Close the Gender Gap in Competition?
<https://doi.org/10.1287/mnsc.2023.4861>.
3. Learning in Networks: An Experiment on Large Networks with Real-World Features
<https://doi.org/10.1287/mnsc.2023.4680>.

4. Strategic Inattention in Product Search
<https://doi.org/10.1287/mnsc.2022.03413>.
5. The Risk of Caution: Evidence from an Experiment
<https://doi.org/10.1287/mnsc.2021.4292>.

SI2. AI assistance across successive experiments

The comparison traces changes in the value of AI assistance across two generations of the experiment. The designs share core tasks and a seven-hour work window.

Lineage from the first AI Replication Games experiment

The tasks build on Brodeur et al. [19], which studied 288 participants working in 103 teams during seven events in 2024. Its seven-hour assignment required reproduction, error finding, and robustness proposals. The current experiment retains those task families, examines individuals across academic and disciplinary expertise, and adds referee reports. The seven main outcomes are therefore not identical across papers. Both include reproduction and minor and major discoveries; the earlier tables additionally emphasize one- and two-check proposals and reported implementation. Here the main proposal measures use two-check thresholds, qualification records numerical documentation, and the one-check thresholds remain supplementary.

The four conceptual proposal criteria descend from the first study, while their operational grading differs. The earlier study also accepted correction of a major error followed by rerunning the corrected code as a good check; the present approval rule requires all four criteria for each proposal. The earlier implementation outcomes relied on participant reports, whereas the present qualifying indicator requires numerical content in a result field. Referee components and the holistic report score are additional outcomes. These distinctions preserve the task lineage without treating the two tables as identical measures.

The 2024 events provided GPT-4 and, from June onward, GPT-4o; later events also offered o1-preview and o1-mini without file processing. Current events ran from October 2025 through May 2026, with ChatGPT Plus provisioned and broader tool use reported. Neither round fixes a universally used model version. The earlier human-only reproduction rate was 31 of 33 teams, and every human-only team proposed at least one good check. Different ceilings, participant inputs, studies, and measurement prevent interpreting the cross-round contrast as the isolated effect of replacing one model. The comparison documents performance in successive AI-assisted work settings.

Comparative timing and final-output models

The timing comparison uses only human-only and AI-assisted observations from the earlier experiment [19]. Its 68 teams comprise 33 human-only and 35 AI-assisted teams. The current comparison uses 200 within-field participants with usable timing records; the undergraduate-exclusion sample contains 163. The outcome at each elapsed time is an indicator that the assigned result has been reproduced by that time. Separate ordinary least-squares models include paper and event fixed effects. Earlier-study controls are team size, maximum and minimum academic position, and attendance. Current-study controls are the separate graduate and senior indicators. The confidence intervals are pointwise and use robust covariance estimates with the full model's degrees of freedom.

For the workday summary, the dependent variable is elapsed time to reproduction, capped at 420 minutes, with noncompletion assigned 420. Integrating the adjusted completion gap over the window reproduces the adjusted reduction in this measure. The current estimate is 50.6 minutes, with a 95% interval from 15.7 to 85.6; excluding undergraduates gives 64.4 minutes, with an interval from 28.0 to 100.7. The earlier estimate is a reduction of −22.5 minutes, with an interval from −58.6 to 13.6. These quantities combine completion and speed; they are not estimates of labor time saved. This exploratory comparison does not identify which changes in people, assignments, tools, or work organization account for the cross-study pattern.

The observed curves and final-output adjustment map appear in Appendix Figures A1 and A2.

Table S1: Design and measurement in successive AI experiments

Feature	Earlier experiment	Current experiment
Events	Seven events in 2024	Five events, October 2025–May 2026
Working unit	Teams	Individuals
Timing comparison	33 human-only and 35 AI-assisted teams	200 within-field individuals; 163 excluding undergraduates
Shared tasks	Reproduction, error discovery, robustness proposals	Same task families; referee reports added
Reproduction target	Assigned published result	Assigned published result
Approved proposals	Earlier rubric also accepted major-error correction and rerunning	All four proposal criteria required
Implementation measure	Participant-reported implementation	Approved check with numerical documentation
Tools available	GPT-4, GPT-4o; later o1-preview/o1-mini with restricted file processing	ChatGPT Plus provisioned; broader tools reported
Composition controls	Team size, minimum and maximum academic position, attendance	Separate graduate and senior indicators
Restricted-time AI effect	Reduction of -22.5 minutes; 95% CI $[-58.6, 13.6]$	Reduction of 50.6 minutes; 95% CI $[15.7, 85.6]$

Notes: Sources: Brodeur et al. [19], the current study records, and Section S12. Comparison samples include only human-only and AI-assisted arms. Restricted time assigns 420 minutes to noncompletion; a positive reduction favors assistance. Main Figure 2 uses separate paper and event fixed effects within each study. These are study-specific effects, not a test isolating changes in model capability. Proposal criteria and team structure differ; neither experiment fixes one universally used model version.

SI3. Expertise and treatment heterogeneity

These analyses examine whether AI gains differ across academic ranks, coding experience, and events. Component-specific sensitivity checks retain separate graduate and professor/researcher comparisons with undergraduates.

Statistical models

Primary and secondary outcome estimates use OLS with heteroskedasticity-robust standard errors. The main specifications absorb separate event, assigned-paper, and participant-discipline fixed effects. Appendix alternatives change the controls or fixed effects as indicated in each table. The count-model diagnostics in [Table S14](#) and binary-link checks in [Tables S15–S16](#) are kept separate from the common OLS estimates. The supplementary vertical sample is restricted to inside-discipline assignments and includes separate academic-rank indicators and the stated fixed effects, without additional baseline covariates. Main [Tables 1](#) and [2](#) use the same separate graduate and senior indicators for the seven individual outcomes. The horizontal sample contains graduate students and professor/researchers and controls for tier, years of coding experience, an indicator for missing coding-experience responses, and prior GPT familiarity. Software is not included as a control.

The additive models estimate treatment, expertise, and discipline-match associations without treatment interactions. The interacted vertical model includes treatment-by-graduate and treatment-by-professor/researcher terms. The interacted horizontal model includes treatment-by-outside-discipline. The supplementary coding-experience specification pools all assignments and uses coding years centered at seven and scaled in five-year units. It controls for tier, discipline match, and an indicator for missing coding-experience responses. [Table S8](#) adds AI-by-coding-years, so its AI coefficient refers to seven coding years and its experience coefficient describes human-only participants. These are treatment-effect gradients, not causal effects of accumulating experience. [Table S2](#) reports the AI effect within each academic tier, combining the main treatment coefficient and the corresponding interaction. The confidence intervals account for their covariance. These effects describe each group's gain; [Table 1](#), Panel B, tests differences between academic ranks.

Component specification curves

The computational sensitivity analysis estimates each original outcome separately, retaining separate graduate and professor/researcher interactions relative to undergraduates. We specify 43 model and sample checks for each outcome, both including and excluding master's students. These comprise the main model; each single-paper and single-event exclusion; four restrictions based on representation of career groups and treatment arms; six models allowing AI effects to vary by paper, event, or both; event-by-rank fixed effects; and seven adjustments for coding experience, prior AI familiarity, and complete-case samples. Support restrictions use group and treatment counts rather than outcome values. We additionally omit each participant in turn. Outcomes retain their original units. Pointwise 95% intervals use the same robust variance convention as the main tables. All identified contrasts are displayed, including negative estimates; unidentifiable contrasts are marked. These related specifications are sensitivity checks, not independent replications or a global test of widening. [Appendix Figure A3](#) displays these structural checks. The participant-influence checks follow separately.

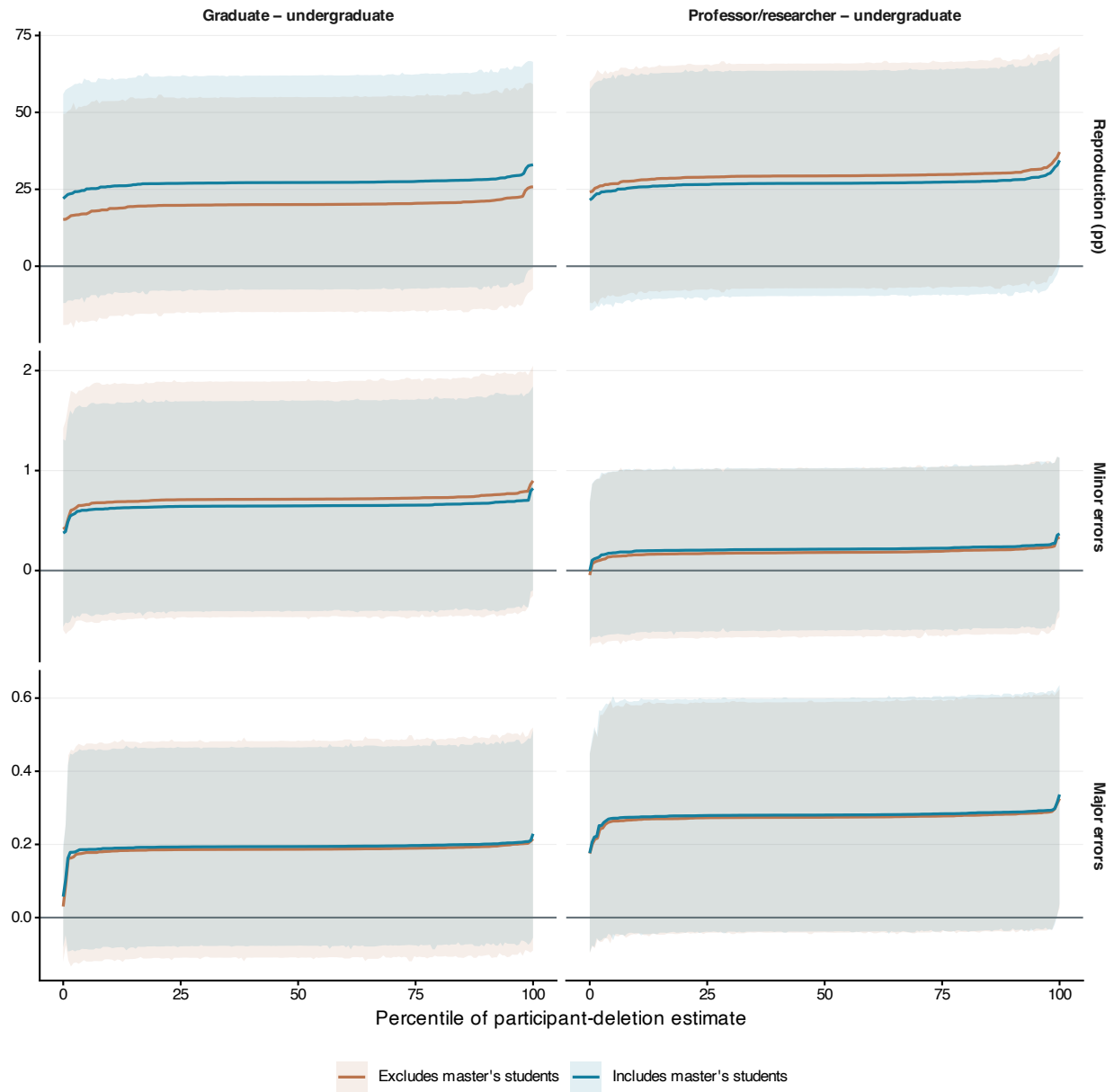


Figure S1: Participant influence. Each curve orders the interaction estimates from every leave-one-participant-out model within its outcome, career comparison, and sample variant. Each contains 202 fits including master's students and 183 excluding them. Shading denotes pointwise 95% intervals. Percentile positions need not correspond to the same participant across curves. All 2,310 identified contrasts are shown.

Table S2: AI effects within each academic tier

Outcome	Undergraduate	Graduate	Professor/researcher
	−8.7 (16.3)	18.6** (8.1)	18.3** (8.6)
Reproduction (pp)	[−40.9, 23.6]	[2.7, 34.5]	[1.3, 35.3]
Human / AI observations	16 / 21	40 / 47	42 / 36
	0.21 (0.35)	0.86** (0.42)	0.43* (0.23)
Minor errors	[−0.47, 0.90]	[0.04, 1.68]	[−0.03, 0.89]
Human / AI observations	16 / 21	40 / 47	42 / 36
	−0.05 (0.10)	0.14** (0.07)	0.23 (0.15)
Major errors	[−0.24, 0.14]	[0.01, 0.28]	[−0.07, 0.52]
Human / AI observations	16 / 21	40 / 47	42 / 36
	18.6** (9.3)	17.4* (9.1)	12.3 (8.9)
Two approved proposals (pp)	[0.3, 36.9]	[−0.6, 35.5]	[−5.2, 29.8]
Human / AI observations	16 / 21	40 / 47	42 / 35
	16.1* (8.7)	16.4* (8.8)	6.5 (8.7)
Two qualifying checks (pp)	[−1.1, 33.3]	[−0.8, 33.7]	[−10.7, 23.7]
Human / AI observations	16 / 21	40 / 47	42 / 35
	20.9 (13.1)	19.2*** (7.1)	18.2** (9.0)
Appropriate report (pp)	[−5.0, 46.7]	[5.1, 33.3]	[0.4, 36.0]
Human / AI observations	16 / 21	40 / 47	42 / 36
	0.86 (0.54)	0.83** (0.33)	1.05*** (0.34)
Overall report score	[−0.20, 1.92]	[0.17, 1.49]	[0.38, 1.72]
Human / AI observations	16 / 21	40 / 47	42 / 36

Notes: Within-discipline sample and the interacted model in main Table 1, Panel B. Each entry adds the AI coefficient to the relevant tier interaction; standard errors use their full covariance. Parentheses contain robust standard errors; brackets contain 95% intervals. Counts below each outcome are the model-retained human-only and AI-assisted observations in that tier. Binary effects are percentage points; counts and the overall score retain their units. Stars are unadjusted: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Within-tier effects and between-tier differences answer distinct questions.

Table S3 examines coding experience rather than academic rank. Panel A uses within-field assignments and imputes missing coding years; Panel B retains complete cases; Panel C pools field assignments. Interactions measure the change in the AI gain per five additional coding years. The AI effect is evaluated at seven years of coding experience.

Table S3: Coding experience replacing academic tier: sample comparisons

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	1+ appr. checks (4)	2+ appr. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	Human approp. (8)	Human overall (9)	AI approp. (10)	AI overall (11)
Panel A. Within discipline: imputed coding years											
Coding years (per 5)	6.84 (4.36) [−1.76, 15.43]	0.05 (0.07) [−0.09, 0.18]	0.00 (0.02) [−0.04, 0.04]	11.30** (4.58) [2.26, 20.33]	9.55*** (3.19) [3.25, 15.85]	11.83*** (4.01) [3.92, 19.74]	9.35*** (3.19) [3.06, 15.64]	6.17* (3.30) [−0.34, 12.69]	0.44*** (0.12) [0.21, 0.67]	10.87*** (4.16) [2.66, 19.08]	0.34*** (0.10) [0.15, 0.53]
AI-assisted	14.07** (5.83) [2.57, 25.57]	0.62*** (0.22) [0.18, 1.06]	0.13* (0.07) [−0.01, 0.27]	22.60*** (6.47) [9.82, 35.38]	15.99*** (5.30) [5.53, 26.46]	17.50*** (6.49) [4.68, 30.32]	12.62** (5.12) [2.51, 22.74]	19.24*** (5.08) [9.22, 29.26]	0.97*** (0.23) [0.53, 1.42]	23.90*** (6.53) [11.00, 36.79]	0.93*** (0.19) [0.56, 1.29]
AI × coding years	−6.70 (5.93) [−18.41, 5.01]	−0.09 (0.13) [−0.34, 0.16]	0.01 (0.03) [−0.05, 0.07]	−12.91** (6.08) [−24.91, −0.90]	−10.69** (4.57) [−19.70, −1.67]	−17.94*** (4.99) [−27.79, −8.10]	−12.09*** (4.11) [−20.20, −3.99]	−9.14** (4.48) [−17.99, −0.29]	−0.35** (0.15) [−0.65, −0.04]	−12.88** (6.22) [−25.16, −0.60]	−0.38** (0.19) [−0.76, 0.00]
Observations	202	202	202	201	201	201	201	202	202	202	202
Mean outcome	70.79	0.50	0.11	47.76	22.39	39.80	19.90	18.32	2.50	57.92	2.76
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Panel B. Within discipline: observed coding years											
Coding years (per 5)	7.56* (4.28) [−0.89, 16.01]	0.01 (0.07) [−0.13, 0.15]	0.00 (0.02) [−0.04, 0.05]	12.05** (4.79) [2.59, 21.51]	9.94*** (3.27) [3.49, 16.40]	11.94*** (4.12) [3.81, 20.07]	9.69*** (3.25) [3.26, 16.11]	5.86* (3.41) [−0.88, 12.60]	0.39*** (0.12) [0.15, 0.62]	10.92** (4.33) [2.37, 19.46]	0.35*** (0.10) [0.15, 0.55]
AI-assisted	17.37*** (5.99) [5.55, 29.20]	0.67*** (0.22) [0.22, 1.11]	0.14* (0.08) [−0.01, 0.29]	22.05*** (6.68) [8.85, 35.24]	16.31*** (5.67) [5.12, 27.50]	15.31** (6.79) [1.91, 28.72]	12.88** (5.51) [1.99, 23.77]	18.95*** (5.27) [8.55, 29.35]	1.02*** (0.23) [0.57, 1.47]	26.27*** (6.77) [12.90, 39.63]	0.99*** (0.19) [0.62, 1.37]
AI × coding years	−8.58 (5.67) [−19.77, 2.61]	−0.07 (0.13) [−0.32, 0.17]	0.01 (0.03) [−0.05, 0.06]	−13.60** (6.28) [−26.00, −1.19]	−11.18** (4.66) [−20.37, −1.98]	−17.41*** (5.15) [−27.58, −7.24]	−12.51*** (4.21) [−20.82, −4.21]	−8.67* (4.59) [−17.73, 0.38]	−0.32** (0.15) [−0.62, −0.02]	−13.59** (6.41) [−26.25, −0.93]	−0.41** (0.19) [−0.79, −0.03]
Observations	188	188	188	187	187	187	187	188	188	188	188
Mean outcome	71.28	0.47	0.12	49.20	23.53	41.18	20.86	18.62	2.52	58.51	2.79
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Years are centered at seven and divided by five. AI-assisted is evaluated at seven coding years. Coding years replaces tier. Panel A adds only a missing-years indicator; Panel B excludes missing years. Neither adds GPT or software controls. Missing years use event-by-tier medians, with the overall median as fallback; outcomes are not imputed.

Table S3: Coding experience replacing academic tier: sample comparisons (continued)

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	1+ appr. checks (4)	2+ appr. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	Human approp. (8)	Human overall (9)	AI approp. (10)	AI overall (11)
Panel C. Pooled assignments: tier replaced by coding years											
Coding years (per 5)	1.51 (3.55) [−5.48, 8.49]	0.01 (0.05) [−0.09, 0.10]	0.01 (0.02) [−0.02, 0.05]	10.69*** (3.90) [3.00, 18.38]	8.03*** (2.85) [2.42, 13.64]	9.66*** (3.48) [2.81, 16.51]	7.31*** (2.67) [2.05, 12.58]	1.96 (2.71) [−3.38, 7.31]	0.29*** (0.11) [0.07, 0.50]	8.58** (3.81) [1.07, 16.08]	0.29*** (0.09) [0.11, 0.47]
AI-assisted	16.43*** (5.05) [6.49, 26.36]	0.52*** (0.15) [0.22, 0.82]	0.15** (0.06) [0.03, 0.27]	19.99*** (5.67) [8.82, 31.16]	16.44*** (4.77) [7.05, 25.83]	20.29*** (5.59) [9.29, 31.29]	17.35*** (4.57) [8.36, 26.34]	19.15*** (4.28) [10.72, 27.57]	0.94*** (0.19) [0.58, 1.31]	23.62*** (5.52) [12.74, 34.50]	0.91*** (0.16) [0.59, 1.22]
AI × coding years	0.87 (4.79) [−8.56, 10.30]	−0.05 (0.09) [−0.22, 0.12]	0.01 (0.04) [−0.06, 0.08]	−10.08* (5.19) [−20.30, 0.14]	−9.21** (3.98) [−17.05, −1.37]	−11.08** (4.65) [−20.24, −1.91]	−9.33** (3.64) [−16.49, −2.17]	−5.39 (3.53) [−12.34, 1.56]	−0.22* (0.14) [−0.49, 0.04]	−8.29 (5.15) [−18.43, 1.84]	−0.26* (0.15) [−0.56, 0.04]
Outside discipline	−5.40 (7.17) [−19.51, 8.71]	−0.38** (0.15) [−0.69, −0.08]	−0.03 (0.12) [−0.26, 0.19]	−6.52 (8.50) [−23.26, 10.21]	6.85 (7.13) [−7.20, 20.90]	−8.33 (8.24) [−24.56, 7.89]	4.81 (7.13) [−9.24, 18.85]	−6.38 (6.80) [−19.78, 7.02]	−0.24 (0.29) [−0.81, 0.33]	−13.74 (8.35) [−30.18, 2.70]	−0.35 (0.23) [−0.81, 0.10]
Observations	286	286	286	285	285	285	285	286	286	286	286
Mean outcome	73.08	0.42	0.15	50.18	24.56	41.05	21.40	17.48	2.57	59.44	2.76
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Years are centered at seven and divided by five. AI-assisted is evaluated at seven coding years. The pooled model adds discipline match and a missing-years indicator, but no tier, GPT, or software controls. It differs from the tier-adjusted supplementary coding-experience specification. Missing years use event-by-tier medians, with the overall median as fallback; outcomes are not imputed.

Table S4: Treatment heterogeneity across events

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	1+ appr. checks (4)	2+ appr. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	Human approp. (8)	Human overall (9)	AI approp. (10)	AI overall (11)
Panel A. Planned linear event-order interaction											
Event order (absorbed)	—	—	—	—	—	—	—	—	—	—	—
AI-assisted at Ottawa	10.63 (8.27) [−5.65, 26.92]	0.55* (0.28) [0.00, 1.11]	0.22** (0.11) [0.01, 0.44]	18.62* (10.46) [−1.97, 39.22]	4.43 (9.22) [−13.72, 22.59]	9.01 (10.43) [−11.53, 29.55]	7.90 (8.57) [−8.97, 24.78]	28.26*** (8.27) [11.98, 44.54]	0.95** (0.37) [0.22, 1.69]	20.51** (10.34) [0.15, 40.86]	0.80*** (0.27) [0.28, 1.33]
AI × event order	2.88 (3.62) [−4.24, 10.01]	−0.02 (0.09) [−0.21, 0.16]	−0.04 (0.04) [−0.12, 0.05]	−0.15 (4.54) [−9.09, 8.78]	5.07 (3.80) [−2.42, 12.55]	4.63 (4.51) [−4.25, 13.51]	3.66 (3.58) [−3.38, 10.71]	−5.47 (3.41) [−12.19, 1.26]	−0.02 (0.15) [−0.30, 0.27]	1.21 (4.19) [−7.04, 9.47]	0.05 (0.10) [−0.15, 0.26]
Observations	286	286	286	285	285	285	285	286	286	286	286
Mean outcome	73.08	0.42	0.15	50.18	24.56	41.05	21.40	17.48	2.57	59.44	2.76
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Panel B. Unrestricted event interactions											
AI-assisted at Ottawa	−4.32 (11.40) [−26.77, 18.14]	0.09 (0.22) [−0.33, 0.52]	0.11 (0.09) [−0.06, 0.29]	14.64 (14.58) [−14.07, 43.36]	−1.55 (12.21) [−25.60, 22.51]	2.95 (13.72) [−24.07, 29.98]	5.35 (10.55) [−15.43, 26.14]	22.83** (10.26) [2.61, 43.05]	0.52 (0.57) [−0.61, 1.64]	11.28 (15.16) [−18.57, 41.13]	0.44 (0.33) [−0.21, 1.10]
AI × Toronto	25.51* (14.62) [−3.29, 54.30]	1.07 (0.69) [−0.28, 2.43]	0.17 (0.19) [−0.22, 0.55]	11.39 (19.19) [−26.41, 49.18]	13.15 (16.58) [−19.52, 45.81]	18.77 (18.58) [−17.82, 55.36]	5.31 (15.11) [−24.45, 35.08]	12.79 (14.42) [−15.60, 41.19]	0.66 (0.68) [−0.68, 2.01]	22.93 (19.03) [−14.56, 60.41]	0.64 (0.50) [−0.34, 1.61]
AI × Cambridge	26.20* (14.75) [−2.85, 55.26]	0.10 (0.27) [−0.44, 0.64]	0.06 (0.14) [−0.21, 0.33]	−1.95 (18.33) [−38.06, 34.17]	20.46 (14.61) [−8.31, 49.23]	11.26 (17.38) [−22.97, 45.49]	13.26 (13.53) [−13.39, 39.90]	−12.91 (13.08) [−38.67, 12.85]	0.65 (0.64) [−0.62, 1.91]	12.84 (17.91) [−22.44, 48.12]	0.83** (0.41) [0.01, 1.65]
AI × UCL	31.73* (17.53) [−2.79, 66.25]	0.73 (0.49) [−0.24, 1.70]	−0.02 (0.22) [−0.44, 0.41]	9.05 (22.22) [−34.72, 52.82]	20.41 (18.28) [−15.60, 56.41]	25.89 (21.46) [−16.37, 68.15]	15.14 (17.17) [−18.68, 48.96]	−15.42 (15.14) [−45.25, 14.40]	0.19 (0.74) [−1.27, 1.66]	−7.37 (20.41) [−47.57, 32.83]	−0.41 (0.62) [−1.63, 0.80]
AI × KCL	16.23 (16.22) [−15.72, 48.18]	0.31 (0.33) [−0.35, 0.97]	−0.06 (0.16) [−0.38, 0.27]	3.46 (20.26) [−36.45, 43.38]	22.73 (16.71) [−10.18, 55.64]	23.06 (19.77) [−15.88, 62.00]	13.76 (15.41) [−16.59, 44.11]	−10.62 (14.94) [−40.04, 18.80]	0.23 (0.69) [−1.13, 1.59]	21.52 (19.49) [−16.86, 59.91]	0.72* (0.43) [−0.12, 1.56]
Observations	286	286	286	285	285	285	285	286	286	286	286
Mean outcome	73.08	0.42	0.15	50.18	24.56	41.05	21.40	17.48	2.57	59.44	2.76
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Controls: tier, discipline match, coding years, missing years, and GPT familiarity. Event order is Ottawa=0, Toronto=1, Cambridge=2, UCL=3, KCL=4. It compares different cohorts/sites, not individual learning. Panel B is an additional functional-form check. Missing years use event-by-tier medians, with the overall median as fallback; outcomes are not imputed.

Table S5: Five-category expertise: amended secondary specification

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	1+ appr. checks (4)	2+ appr. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	Human approp. (8)	Human overall (9)	AI approp. (10)	AI overall (11)
Panel A. Vertical: within-discipline assignments											
Master's (vs undergrad)	−18.18 (18.75) [−55.20, 18.83]	−0.44 (0.40) [−1.24, 0.35]	−0.09 (0.12) [−0.32, 0.14]	−16.23 (17.05) [−49.89, 17.42]	21.67 (14.61) [−7.18, 50.53]	−14.29 (17.92) [−49.68, 21.09]	19.93 (14.24) [−8.19, 48.05]	−13.68 (12.77) [−38.89, 11.52]	0.45 (0.53) [−0.59, 1.50]	1.07 (22.07) [−42.51, 44.65]	0.07 (0.54) [−0.99, 1.13]
PhD student (vs undergrad)	14.98 (13.31) [−11.30, 41.26]	−0.02 (0.33) [−0.68, 0.64]	0.03 (0.06) [−0.09, 0.14]	22.87* (13.50) [−3.78, 49.52]	15.41* (8.94) [−2.23, 33.05]	11.29 (14.19) [−16.72, 39.30]	10.20 (8.21) [−6.01, 26.41]	−3.25 (9.95) [−22.90, 16.39]	0.81 (0.49) [−0.16, 1.78]	15.33 (15.65) [−15.56, 46.22]	0.83* (0.46) [−0.09, 1.74]
Postdoc (vs undergrad)	5.30 (15.06) [−24.43, 35.03]	0.18 (0.43) [−0.66, 1.03]	0.37 (0.31) [−0.24, 0.98]	16.16 (16.34) [−16.11, 48.43]	14.84 (12.61) [−10.05, 39.73]	13.82 (16.78) [−19.30, 46.94]	13.07 (12.88) [−12.36, 38.51]	−27.68** (13.19) [−53.71, −1.64]	−0.22 (0.59) [−1.39, 0.94]	11.70 (19.31) [−26.42, 49.82]	0.58 (0.55) [−0.49, 1.66]
PhD researcher (vs undergrad)	11.09 (13.69) [−15.93, 38.12]	−0.25 (0.30) [−0.84, 0.35]	−0.01 (0.07) [−0.16, 0.13]	31.80** (14.78) [2.62, 60.98]	25.67*** (9.15) [7.60, 43.74]	26.12* (14.64) [−2.79, 55.03]	25.11*** (9.06) [7.22, 43.00]	5.64 (10.82) [−15.73, 27.01]	0.89* (0.48) [−0.05, 1.84]	27.09* (15.16) [−2.83, 57.01]	0.86* (0.46) [−0.05, 1.77]
AI-assisted	−8.92 (16.61) [−41.72, 23.88]	0.26 (0.36) [−0.45, 0.98]	−0.05 (0.09) [−0.23, 0.14]	26.71* (16.00) [−4.88, 58.30]	18.39* (9.36) [−0.09, 36.88]	27.03* (15.80) [−4.17, 58.23]	15.81* (8.83) [−1.61, 33.24]	20.02 (13.19) [−6.02, 46.06]	0.84 (0.54) [−0.23, 1.92]	26.09 (17.10) [−7.67, 59.85]	0.97* (0.56) [−0.13, 2.06]
AI × master's	58.57** (24.74) [9.73, 107.41]	0.36 (0.53) [−0.69, 1.40]	0.25 (0.18) [−0.09, 0.60]	24.95 (23.93) [−22.30, 72.20]	−7.48 (21.54) [−50.01, 35.06]	24.42 (25.09) [−25.13, 73.96]	−3.49 (21.19) [−45.34, 38.35]	−2.90 (19.00) [−40.40, 34.60]	0.16 (0.78) [−1.39, 1.70]	8.86 (29.19) [−48.77, 66.50]	0.38 (0.74) [−1.08, 1.84]
AI × PhD student	20.44 (17.91) [−14.92, 55.80]	0.70 (0.59) [−0.47, 1.87]	0.17 (0.15) [−0.12, 0.47]	0.20 (18.66) [−36.65, 37.05]	0.76 (13.29) [−25.47, 26.99]	−5.19 (19.00) [−42.71, 32.33]	1.26 (12.42) [−23.26, 25.79]	1.32 (15.30) [−28.87, 31.52]	−0.02 (0.65) [−1.30, 1.26]	−2.44 (21.00) [−43.90, 39.03]	−0.29 (0.62) [−1.51, 0.94]
AI × postdoc	49.88** (21.80) [6.85, 92.91]	0.18 (0.61) [−1.03, 1.40]	−0.09 (0.37) [−0.82, 0.64]	−9.75 (28.91) [−66.83, 47.33]	−3.53 (22.13) [−47.23, 40.17]	−6.44 (28.56) [−62.84, 49.96]	−0.06 (21.75) [−43.00, 42.88]	−2.11 (17.35) [−36.37, 32.14]	0.88 (0.80) [−0.71, 2.46]	14.79 (26.24) [−37.02, 66.59]	0.46 (0.73) [−0.97, 1.90]
AI × PhD researcher	20.25 (19.65) [−18.54, 59.05]	0.13 (0.45) [−0.76, 1.02]	0.38** (0.18) [0.02, 0.73]	−17.00 (20.95) [−58.37, 24.37]	−6.40 (13.99) [−34.03, 21.23]	−30.82 (19.82) [−69.96, 8.31]	−11.73 (13.55) [−38.49, 15.02]	−0.20 (17.04) [−33.84, 33.44]	0.04 (0.65) [−1.25, 1.33]	−10.40 (20.91) [−51.68, 30.89]	−0.02 (0.66) [−1.33, 1.29]
Observations	202	202	202	201	201	201	201	202	202	202	202
Mean outcome	70.79	0.50	0.11	47.76	22.39	39.80	19.90	18.32	2.50	57.92	2.76
Additional controls	—	—	—	—	—	—	—	—	—	—	—
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. No additional controls. Before outcome/FE exclusions, human/AI counts are: undergrad 16/21, master's 8/12, PhD 33/35, postdoc 9/9, PhD researcher 33/27. The April 16 amendment specifies these categories; public deposit timing is disclosed in Materials and Methods.

Table S6: Formal treatment-heterogeneity tests and multiplicity checks**Panel A. Equality and ordered treatment-effect tests**

Outcome	Grad – UG	Res. – grad	Equal p	Decr. p	Incr. p	AI $\times$ out.
Comp. rep.	27.21 (17.62)	-0.26 (11.98)	0.283	0.938	0.509	8.90 (10.80)
Minor errors	0.65 (0.53)	-0.43 (0.47)	0.476	0.886	0.821	-0.30 (0.28)
Major errors	0.19 (0.14)	0.09 (0.16)	0.172	0.920	0.296	-0.02 (0.20)
1+ appr. checks	3.08 (17.94)	-19.28 (15.24)	0.443	0.568	0.896	-0.85 (13.69)
2+ appr. checks	-1.15 (12.43)	-5.16 (13.30)	0.884	0.463	0.651	0.73 (12.51)
1+ qual. checks	-0.30 (17.95)	-25.47 (15.38)	0.209	0.493	0.950	16.89 (13.65)
2+ qual. checks	0.31 (11.62)	-9.94 (12.85)	0.694	0.511	0.780	13.53 (12.05)
Human approp.	-1.65 (14.74)	-0.97 (11.84)	0.987	0.467	0.545	2.65 (9.82)
Human overall	-0.03 (0.62)	0.22 (0.48)	0.894	0.675	0.522	-0.01 (0.40)
AI approp.	-0.87 (20.15)	-3.99 (15.32)	0.953	0.483	0.603	5.62 (11.90)
AI overall	-0.18 (0.60)	0.26 (0.40)	0.804	0.742	0.620	0.02 (0.32)

Panel B. Coding-years moderation: complete cases and BH adjustment

Outcome	Full	Complete	Full p	CC p	BH full	BH CC
Comp. rep.	0.49 (4.72)	-1.44 (4.58)	0.917	0.753	0.917	0.816
Minor errors	-0.06 (0.09)	-0.05 (0.09)	0.485	0.541	0.593	0.661
Major errors	0.02 (0.03)	0.01 (0.04)	0.606	0.816	0.667	0.816
1+ appr. checks	-10.49** (5.14)	-11.41** (5.25)	0.042	0.031	0.116	0.084
2+ appr. checks	-9.32** (3.89)	-9.48** (3.97)	0.017	0.018	0.063	0.070
1+ qual. checks	-11.23** (4.66)	-11.20** (4.75)	0.017	0.019	0.063	0.070
2+ qual. checks	-9.36*** (3.56)	-9.44** (3.64)	0.009	0.010	0.063	0.070
Human approp.	-5.37 (3.56)	-5.63 (3.61)	0.133	0.121	0.183	0.166
Human overall	-0.24* (0.13)	-0.24* (0.13)	0.069	0.078	0.132	0.143
AI approp.	-8.31* (4.93)	-8.31* (5.03)	0.093	0.100	0.147	0.156
AI overall	-0.27* (0.15)	-0.29* (0.15)	0.072	0.052	0.132	0.114

Notes: Binary effects are percentage points; parentheses contain robust SEs. Panel A compares treatment effects, not outcome levels. Equality is a two-restriction Wald F test. Decreasing/increasing tests take the maximum of two one-sided component p -values using the fitted covariance; ordered effects alone do not establish gap closure. Panel B reports AI-by-years effects per five years; full hierarchies appear in [Table S8](#). BH adjusts eleven outcomes separately within each sample, not all appendix tests. It is exploratory. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ are unadjusted. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Vertical has no controls; Horizontal controls: academic tier, coding years, an indicator for missing coding-experience responses, and GPT familiarity (None/Some, Heavy, Missing). Panel B retains tier, discipline match, and missing years where varying.

SI4. Robustness of the quantitative findings

The following checks vary the sample, controls, fixed effects, functional form, and treatment of the unavailable robustness submission. Table [S12](#) compares the main estimates with models controlling for baseline proficiency in R, Stata, Python, and Matlab. The comparison holds each estimation sample fixed.

Prespecified analyses and sensitivity checks

The appendix groups all eleven outcome constructs within the same specification tables. The non-coding analyses planned in PAP Tables 9 and 10 appear in the last four columns of main Tables [1](#) and [2](#), with one-check thresholds in Table [S23](#) and individual human and AI report assessments in Table [S22](#). Table [S23](#) places one- and two-check robustness thresholds side by side. Tables [S3–S6](#) report the planned experience, event-order, five-category, and formal compression analyses. Event order compares different cohorts rather than individual learning.

Tables [S8–S11](#) combine related complete-case, population, assignment-link, covariate, and fixed-effect checks without changing their distinct samples. The pooled coding-experience complete-case checks retain tier controls and seven-year centering. Tables [A2–S26](#) report balance, missingness, grading coverage, and survey descriptions. Table [S28](#) records unresolved measurement and implementation departures. These disclosures distinguish estimated analyses from verified protocol adherence. Table [S7](#) compares the main classification with the saved randomization record and with omitting the corrected observation. Average AI effects remain positive across all seven outcomes. The participant worked outside their field, so the vertical estimates are identical.

Table S7: Sensitivity to the assignment-record correction

A. All assignments: adjusted average effects			
Outcome	Communicated access	Recorded assignment	Omit corrected record
Reproduction (pp)	16.3*** (4.9)	16.0*** (4.9)	16.1*** (4.9)
Minor errors	0.51*** (0.15)	0.51*** (0.16)	0.51*** (0.16)
Major errors	0.15** (0.06)	0.15** (0.06)	0.15** (0.06)
Two approved proposals (pp)	14.4*** (4.9)	14.7*** (5.0)	14.6*** (5.0)
Two qualifying checks (pp)	15.1*** (4.7)	15.4*** (4.7)	15.3*** (4.7)
Appropriate report (pp)	17.5*** (4.3)	17.4*** (4.3)	17.5*** (4.3)
Overall report score	0.92*** (0.18)	0.91*** (0.18)	0.92*** (0.18)
Observations (computational / proposals)	286 / 285	286 / 285	285 / 284
B. Horizontal study: average effects			
Outcome	Communicated access	Recorded assignment	Omit corrected record
Reproduction (pp)	21.1*** (4.9)	20.7*** (4.9)	20.9*** (4.9)
Minor errors	0.58*** (0.18)	0.59*** (0.19)	0.59*** (0.19)
Major errors	0.16** (0.07)	0.16** (0.07)	0.16** (0.07)
Two approved proposals (pp)	14.2** (5.6)	14.5** (5.7)	14.5** (5.7)
Two qualifying checks (pp)	15.1*** (5.4)	15.5*** (5.4)	15.4*** (5.4)
Appropriate report (pp)	20.1*** (4.6)	20.0*** (4.6)	20.1*** (4.6)
Overall report score	0.93*** (0.19)	0.92*** (0.19)	0.93*** (0.19)
Observations (computational / proposals)	249 / 248	249 / 248	248 / 247
C. Horizontal study: outside-field interactions			
Outcome	Communicated access	Recorded assignment	Omit corrected record
Reproduction (pp)	8.9 (10.8)	7.9 (10.9)	8.4 (11.0)
Minor errors	-0.30 (0.28)	-0.29 (0.29)	-0.29 (0.29)
Major errors	-0.02 (0.20)	-0.01 (0.20)	-0.01 (0.20)
Two approved proposals (pp)	0.7 (12.5)	1.9 (12.8)	1.8 (12.8)
Two qualifying checks (pp)	13.5 (12.1)	15.1 (12.3)	15.0 (12.3)
Appropriate report (pp)	2.7 (9.8)	2.5 (10.0)	2.7 (10.0)
Overall report score	-0.01 (0.40)	-0.03 (0.40)	-0.02 (0.41)
Observations (computational / proposals)	249 / 248	249 / 248	248 / 247

Notes: The main analysis follows the AI-assisted instructions received by one participant whose saved randomization entry was human-only. Alternatives use the saved assignment or omit that record. This participant worked outside field, so vertical estimates are unchanged. Panels A and B report average AI effects; Panel C reports the AI-by-outside-field interaction. Binary effects are percentage points. Parentheses contain robust standard errors; stars use unadjusted conventional thresholds. Sample-specific coding-year imputation is recomputed. Referee-outcome sample sizes equal computational sample sizes.

Table S8: Supplementary coding-years specification: complete-case and assignment-link checks

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	1+ appr. checks (4)	2+ appr. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	Human approp. (8)	Human overall (9)	AI approp. (10)	AI overall (11)
Panel A. Additive: current imputation											
Coding years (per 5)	0.64 (2.84)	−0.01 (0.06)	−0.03 (0.04)	4.69 (3.51)	1.33 (2.36)	3.40 (3.32)	0.36 (2.19)	−1.31 (2.54)	0.17** (0.08)	0.99 (3.21)	0.07 (0.11)
AI-assisted	[−4.95, 6.24] 17.08*** (4.83) [7.57, 26.59]	[−0.13, 0.10] 0.51*** (0.15) [0.22, 0.81]	[−0.10, 0.04] 0.16*** (0.06) [0.04, 0.28]	[−2.23, 11.60] 18.50*** (5.70) [7.27, 29.74]	[−3.33, 5.98] 15.25*** (4.85) [5.71, 24.80]	[−3.15, 9.94] 18.44*** (5.64) [7.32, 29.55]	[−3.95, 4.67] 16.15*** (4.62) [7.05, 25.25]	[−6.31, 3.69] 18.26*** (4.22) [9.96, 26.56]	[0.01, 0.32] 0.91*** (0.18) [0.56, 1.27]	[−5.33, 7.31] 22.89*** (5.49) [12.07, 33.70]	[−0.14, 0.29] 0.88*** (0.15) [0.58, 1.19]
Observations	286	286	286	285	285	285	285	286	286	286	286
Mean outcome	73.08	0.42	0.15	50.18	24.56	41.05	21.40	17.48	2.57	59.44	2.76
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Panel B. Additive: observed coding years											
Coding years (per 5)	0.35 (2.79)	−0.03 (0.06)	−0.03 (0.04)	4.64 (3.54)	1.22 (2.38)	3.35 (3.35)	0.21 (2.20)	−1.53 (2.58)	0.15* (0.08)	0.81 (3.24)	0.06 (0.11)
AI-assisted	[−5.14, 5.84] 19.52*** (4.96) [9.76, 29.28]	[−0.15, 0.09] 0.52*** (0.15) [0.22, 0.82]	[−0.10, 0.04] 0.17*** (0.06) [0.05, 0.29]	[−2.33, 11.61] 18.04*** (5.87) [6.48, 29.61]	[−3.47, 5.91] 14.87*** (5.11) [4.80, 24.94]	[−3.24, 9.94] 17.26*** (5.82) [5.79, 28.73]	[−4.13, 4.55] 15.91*** (4.89) [6.27, 25.55]	[−6.61, 3.55] 19.00*** (4.34) [10.45, 27.54]	[−0.01, 0.31] 0.95*** (0.18) [0.59, 1.31]	[−5.57, 7.18] 24.81*** (5.64) [13.69, 35.93]	[−0.15, 0.28] 0.93*** (0.16) [0.63, 1.24]
Observations	270	270	270	269	269	269	269	270	270	270	270
Mean outcome	73.70	0.39	0.15	51.67	25.65	42.38	22.30	17.78	2.59	60.00	2.78
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Panel C. Additive: linked assignment record											
Coding years (per 5)	0.69 (2.85)	0.00 (0.06)	−0.02 (0.04)	—	1.43 (2.38)	—	0.46 (2.22)	−1.26 (2.65)	0.16** (0.08)	—	—
AI-assisted	[−4.92, 6.29] 17.56*** (5.06) [7.60, 27.52]	[−0.12, 0.12] 0.55*** (0.16) [0.25, 0.86]	[−0.09, 0.05] 0.18*** (0.06) [0.05, 0.31]	—	[−3.26, 6.11] 15.82*** (5.11) [5.77, 25.88]	—	[−3.91, 4.84] 15.85*** (4.93) [6.14, 25.56]	[−6.48, 3.95] 18.24*** (4.35) [9.67, 26.82]	[0.00, 0.32] 0.99*** (0.18) [0.63, 1.35]	—	—
Observations	273	273	273	—	272	—	272	273	273	—	—
Additional controls	✓	✓	✓	—	✓	—	✓	✓	✓	—	—
Fixed effects	✓	✓	✓	—	✓	—	✓	✓	✓	—	—

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Years are centered at seven and divided by five. AI-assisted is evaluated at seven coding years. Controls: tier, discipline match, and additive missing years where varying; no GPT or software. Years use event-by-tier/global medians. Complete cases exclude missing years. Linked models cover seven main outcomes only; dashes mark unestimated outcomes. Exact assignment matches and their limitations are documented in [Table A2](#).

Table S8: Supplementary coding-years specification: complete-case and assignment-link checks (continued)

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	1+ appr. checks (4)	2+ appr. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	Human approp. (8)	Human overall (9)	AI approp. (10)	AI overall (11)
Panel D. Interacted: current imputation											
Coding years (per 5)	0.35 (4.03)	0.02 (0.05)	−0.04 (0.03)	10.87** (4.38)	6.82** (3.03)	10.01** (3.95)	5.88** (2.84)	1.87 (3.10)	0.31*** (0.12)	5.92 (4.11)	0.23** (0.10)
	[−7.59, 8.29]	[−0.08, 0.13]	[−0.10, 0.02]	[2.24, 19.49]	[0.86, 12.78]	[2.22, 17.80]	[0.28, 11.48]	[−4.23, 7.97]	[0.08, 0.54]	[−2.17, 14.01]	[0.03, 0.43]
AI-assisted	17.00*** (4.93)	0.52*** (0.16)	0.16** (0.06)	20.27*** (5.67)	16.82*** (4.78)	20.32*** (5.63)	17.72*** (4.57)	19.15*** (4.32)	0.95*** (0.18)	24.27*** (5.50)	0.93*** (0.16)
	[7.30, 26.70]	[0.22, 0.83]	[0.04, 0.28]	[9.09, 31.44]	[7.40, 26.23]	[9.23, 31.41]	[8.72, 26.73]	[10.65, 27.66]	[0.59, 1.31]	[13.43, 35.11]	[0.62, 1.23]
AI × coding years	0.49 (4.72)	−0.06 (0.09)	0.02 (0.03)	−10.49** (5.14)	−9.32** (3.89)	−11.23** (4.66)	−9.36*** (3.56)	−5.37 (3.56)	−0.24* (0.13)	−8.31* (4.93)	−0.27* (0.15)
	[−8.80, 9.78]	[−0.23, 0.11]	[−0.05, 0.08]	[−20.61, −0.37]	[−16.98, −1.66]	[−20.40, −2.06]	[−16.38, −2.35]	[−12.38, 1.64]	[−0.51, 0.02]	[−18.02, 1.40]	[−0.56, 0.02]
Observations	286	286	286	285	285	285	285	286	286	286	286
Mean outcome	73.08	0.42	0.15	50.18	24.56	41.05	21.40	17.48	2.57	59.44	2.76
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Panel E. Interacted: observed coding years											
Coding years (per 5)	1.21 (4.00)	0.00 (0.05)	−0.04 (0.03)	11.35** (4.44)	6.80** (3.04)	9.94** (3.94)	5.76** (2.86)	1.79 (3.15)	0.29** (0.12)	5.72 (4.11)	0.24** (0.10)
	[−6.67, 9.08]	[−0.11, 0.11]	[−0.10, 0.03]	[2.61, 20.09]	[0.80, 12.79]	[2.18, 17.70]	[0.13, 11.39]	[−4.41, 8.00]	[0.06, 0.52]	[−2.37, 13.81]	[0.03, 0.44]
AI-assisted	19.80*** (5.10)	0.53*** (0.16)	0.17*** (0.06)	20.30*** (5.87)	16.75*** (5.07)	19.48*** (5.84)	17.78*** (4.87)	20.10*** (4.46)	1.00*** (0.19)	26.44*** (5.67)	0.99*** (0.16)
	[9.75, 29.86]	[0.22, 0.85]	[0.04, 0.29]	[8.75, 31.86]	[6.76, 26.73]	[7.98, 30.98]	[8.17, 27.38]	[11.31, 28.89]	[0.63, 1.37]	[15.27, 37.61]	[0.68, 1.31]
AI × coding years	−1.44 (4.58)	−0.05 (0.09)	0.01 (0.04)	−11.41** (5.25)	−9.48** (3.97)	−11.20** (4.75)	−9.44** (3.64)	−5.63 (3.61)	−0.24* (0.13)	−8.31* (5.03)	−0.29* (0.15)
	[−10.47, 7.58]	[−0.23, 0.12]	[−0.06, 0.08]	[−21.75, −1.07]	[−17.30, −1.66]	[−20.56, −1.84]	[−16.62, −2.26]	[−12.74, 1.49]	[−0.50, 0.03]	[−18.21, 1.59]	[−0.59, 0.00]
Observations	270	270	270	269	269	269	269	270	270	270	270
Mean outcome	73.70	0.39	0.15	51.67	25.65	42.38	22.30	17.78	2.59	60.00	2.78
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Panel F. Interacted: linked assignment record											
Coding years (per 5)	1.74 (3.99)	0.02 (0.05)	−0.03 (0.03)	—	6.56** (3.07)	—	6.20** (2.96)	1.41 (3.21)	0.32*** (0.12)	—	—
	[−6.13, 9.60]	[−0.08, 0.13]	[−0.10, 0.03]		[0.51, 12.61]		[0.37, 12.02]	[−4.91, 7.74]	[0.08, 0.55]		
AI-assisted	17.86*** (5.17)	0.56*** (0.16)	0.18*** (0.06)	—	17.32*** (5.04)	—	17.52*** (4.86)	19.01*** (4.44)	1.03*** (0.18)	—	—
	[7.68, 28.04]	[0.24, 0.88]	[0.05, 0.30]		[7.39, 27.25]		[7.95, 27.10]	[10.26, 27.76]	[0.67, 1.40]		
AI × coding years	−1.79 (4.70)	−0.04 (0.09)	0.01 (0.04)	—	−8.78** (3.96)	—	−9.82*** (3.70)	−4.56 (3.74)	−0.26* (0.14)	—	—
	[−11.05, 7.47]	[−0.22, 0.15]	[−0.06, 0.09]		[−16.58, −0.98]		[−17.11, −2.52]	[−11.93, 2.81]	[−0.53, 0.00]		
Observations	273	273	273	—	272	—	272	273	273	—	—
Additional controls	✓	✓	✓	—	✓	—	✓	✓	✓	—	—
Fixed effects	✓	✓	✓	—	✓	—	✓	✓	✓	—	—

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Years are centered at seven and divided by five. AI-assisted is evaluated at seven coding years. Controls: tier, discipline match, and additive missing years where varying; no GPT or software. Years use event-by-tier/global medians. Complete cases exclude missing years. Linked models cover seven main outcomes only; dashes mark unestimated outcomes. Exact assignment matches and their limitations are documented in [Table A2](#).

Table S9: Vertical specifications: covariates, population, and assignment linkage

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	1+ appr. checks (4)	2+ appr. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	Human approp. (8)	Human overall (9)	AI approp. (10)	AI overall (11)
Panel A. Within discipline: available covariates											
Grad (vs undergrad)	8.68 (12.60) [−16.21, 33.56]	−0.14 (0.30) [−0.74, 0.46]	0.02 (0.06) [−0.11, 0.14]	12.89 (13.28) [−13.33, 39.12]	14.77* (8.26) [−1.54, 31.08]	4.20 (13.51) [−22.47, 30.87]	10.96 (7.67) [−4.19, 26.11]	−6.95 (10.04) [−26.76, 12.87]	0.66 (0.47) [−0.27, 1.59]	11.63 (15.14) [−18.26, 41.52]	0.65 (0.46) [−0.25, 1.55]
Researcher (vs undergrad)	7.85 (13.35) [−18.50, 34.21]	−0.24 (0.32) [−0.87, 0.39]	0.13 (0.10) [−0.07, 0.33]	22.32 (15.58) [−8.45, 53.09]	18.29** (9.14) [0.24, 36.34]	18.86 (15.24) [−11.23, 48.96]	19.65** (8.92) [2.05, 37.26]	−6.79 (11.73) [−29.95, 16.37]	0.31 (0.51) [−0.69, 1.32]	23.47 (16.10) [−8.33, 55.26]	0.80 (0.52) [−0.22, 1.82]
AI-assisted	−9.50 (16.07) [−41.22, 22.22]	0.26 (0.32) [−0.38, 0.90]	−0.07 (0.10) [−0.26, 0.12]	28.53* (15.87) [−2.80, 59.86]	19.33** (9.18) [1.19, 37.46]	27.68* (15.49) [−2.90, 58.26]	16.32* (8.61) [−0.67, 33.32]	20.33 (13.07) [−5.47, 46.13]	0.93* (0.55) [−0.14, 2.01]	26.71 (16.71) [−6.27, 59.69]	0.98* (0.54) [−0.08, 2.04]
AI × grad	27.14 (17.27) [−6.95, 61.23]	0.60 (0.53) [−0.44, 1.65]	0.20 (0.13) [−0.07, 0.46]	2.58 (17.96) [−32.89, 38.05]	−2.36 (12.52) [−27.08, 22.35]	−0.40 (17.96) [−35.87, 35.07]	−0.58 (11.68) [−23.63, 22.48]	−2.31 (14.88) [−31.69, 27.08]	−0.06 (0.63) [−1.30, 1.18]	−0.08 (20.09) [−39.74, 39.58]	−0.15 (0.59) [−1.31, 1.02]
AI × researcher	27.90 (18.56) [−8.74, 64.55]	0.14 (0.40) [−0.64, 0.92]	0.30* (0.16) [−0.02, 0.62]	−20.45 (20.04) [−60.01, 19.11]	−10.11 (13.25) [−36.27, 16.04]	−28.77 (18.93) [−66.14, 8.60]	−12.48 (12.93) [−38.02, 13.06]	−3.80 (16.29) [−35.96, 28.36]	0.07 (0.62) [−1.15, 1.28]	−6.10 (19.24) [−44.09, 31.89]	0.05 (0.60) [−1.13, 1.24]
Observations	202	202	202	201	201	201	201	202	202	202	202
Mean outcome	70.79	0.50	0.11	47.76	22.39	39.80	19.90	18.32	2.50	57.92	2.76
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Panel B. Within discipline: complete covariate cases											
Grad (vs undergrad)	9.88 (14.10) [−17.97, 37.72]	0.08 (0.28) [−0.47, 0.64]	0.03 (0.07) [−0.11, 0.17]	12.08 (15.41) [−18.36, 42.52]	17.22* (9.42) [−1.38, 35.82]	4.97 (15.58) [−25.82, 35.75]	12.67 (8.72) [−4.55, 29.89]	−6.59 (11.72) [−29.74, 16.56]	0.77 (0.54) [−0.29, 1.83]	14.19 (16.23) [−17.86, 46.25]	0.61 (0.50) [−0.38, 1.60]
Researcher (vs undergrad)	10.85 (14.47) [−17.73, 39.42]	−0.15 (0.28) [−0.69, 0.40]	0.13 (0.11) [−0.07, 0.34]	22.54 (17.62) [−12.27, 57.36]	20.93** (9.59) [1.98, 39.88]	18.60 (17.02) [−15.01, 52.22]	21.81** (9.37) [3.30, 40.31]	−7.44 (12.83) [−32.78, 17.91]	0.29 (0.56) [−0.83, 1.40]	26.50 (16.81) [−6.70, 59.70]	0.81 (0.56) [−0.30, 1.93]
AI-assisted	0.36 (18.33) [−35.84, 36.57]	0.38 (0.33) [−0.26, 1.02]	−0.08 (0.12) [−0.31, 0.16]	27.22 (18.95) [−10.22, 64.65]	20.66* (11.10) [−1.26, 42.58]	24.04 (18.41) [−12.32, 60.39]	17.23 (10.43) [−3.39, 37.84]	18.87 (15.24) [−11.24, 48.98]	0.96 (0.59) [−0.21, 2.13]	31.11* (18.73) [−5.89, 68.11]	1.02* (0.61) [−0.18, 2.22]
AI × grad	19.07 (19.50) [−19.45, 57.60]	0.42 (0.58) [−0.74, 1.57]	0.21 (0.16) [−0.11, 0.52]	2.67 (20.89) [−38.61, 43.94]	−4.15 (14.36) [−32.51, 24.21]	−0.48 (20.65) [−41.27, 40.32]	−1.82 (13.38) [−28.25, 24.62]	−1.80 (17.14) [−35.66, 32.06]	−0.16 (0.68) [−1.51, 1.20]	−3.32 (21.93) [−46.64, 39.99]	−0.12 (0.65) [−1.41, 1.17]
AI × researcher	18.51 (20.83) [−22.63, 59.64]	0.14 (0.41) [−0.68, 0.95]	0.31* (0.17) [−0.03, 0.65]	−20.19 (23.03) [−65.68, 25.31]	−12.47 (14.78) [−41.66, 16.72]	−26.21 (21.68) [−69.02, 16.61]	−14.34 (14.42) [−42.82, 14.15]	−1.85 (18.17) [−37.74, 34.04]	0.19 (0.66) [−1.11, 1.50]	−9.61 (21.40) [−51.89, 32.66]	0.03 (0.67) [−1.29, 1.35]
Observations	188	188	188	187	187	187	187	188	188	188	188
Mean outcome	71.28	0.47	0.12	49.20	23.53	41.18	20.86	18.62	2.52	58.51	2.79
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Controls: coding years, missing years, and original GPT categories (Some, Heavy, Missing). Panel B drops missing years or familiarity. Preferred software is unavailable. Missing years use event-by-tier medians, with the overall median as fallback; outcomes are not imputed.

Table S9: Vertical specifications: covariates, population, and assignment linkage (continued)

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	1+ appr. checks (4)	2+ appr. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	Human approp. (8)	Human overall (9)	AI approp. (10)	AI overall (11)
Panel C. All assignments: controlling discipline match											
Grad (vs undergrad)	9.02 (12.13) [−14.87, 32.91]	−0.14 (0.29) [−0.71, 0.43]	0.06 (0.07) [−0.08, 0.20]	21.55* (11.69) [−1.48, 44.58]	10.73 (7.47) [−3.99, 25.45]	13.26 (11.90) [−10.19, 36.70]	6.73 (7.12) [−7.30, 20.76]	−4.00 (9.25) [−22.22, 14.23]	0.67 (0.43) [−0.18, 1.52]	14.83 (13.40) [−11.57, 41.23]	0.69* (0.41) [−0.13, 1.51]
Researcher (vs undergrad)	5.20 (12.00) [−18.43, 28.82]	−0.17 (0.28) [−0.71, 0.38]	0.10 (0.08) [−0.05, 0.25]	25.09** (12.05) [1.36, 48.83]	13.91* (7.76) [−1.37, 29.19]	16.15 (12.02) [−7.52, 39.83]	10.79 (7.46) [−3.90, 25.48]	−4.10 (9.58) [−22.95, 14.76]	0.44 (0.43) [−0.41, 1.29]	22.09* (12.92) [−3.36, 47.54]	0.71* (0.41) [−0.09, 1.52]
AI-assisted	−8.51 (15.51) [−39.05, 22.04]	0.02 (0.33) [−0.63, 0.67]	0.05 (0.08) [−0.11, 0.20]	25.84* (14.94) [−3.59, 55.27]	12.67 (8.86) [−4.77, 30.12]	30.50** (14.65) [1.65, 59.36]	12.15 (8.57) [−4.72, 29.03]	14.29 (13.18) [−11.66, 40.25]	0.73 (0.51) [−0.28, 1.74]	18.90 (15.74) [−12.10, 49.89]	0.73 (0.53) [−0.32, 1.77]
AI × grad	26.31 (16.42) [−6.02, 58.64]	0.67 (0.46) [−0.23, 1.57]	0.01 (0.12) [−0.22, 0.24]	−5.58 (16.93) [−38.93, 27.77]	2.46 (11.42) [−20.02, 24.95]	−12.75 (16.70) [−45.65, 20.15]	4.60 (11.09) [−17.24, 26.44]	5.14 (14.65) [−23.71, 34.00]	0.01 (0.57) [−1.11, 1.13]	4.31 (18.05) [−31.24, 39.86]	0.06 (0.57) [−1.06, 1.19]
AI × researcher	32.03* (16.61) [−0.68, 64.74]	0.42 (0.35) [−0.27, 1.12]	0.23* (0.12) [−0.01, 0.47]	−10.19 (17.44) [−44.53, 24.16]	4.10 (11.77) [−19.08, 27.28]	−13.37 (16.83) [−46.52, 19.77]	5.08 (11.50) [−17.57, 27.72]	3.96 (14.41) [−24.42, 32.34]	0.36 (0.56) [−0.75, 1.46]	4.51 (17.17) [−29.30, 38.33]	0.28 (0.56) [−0.83, 1.39]
Outside discipline	−9.41 (6.86) [−22.92, 4.11]	−0.42** (0.17) [−0.75, −0.08]	−0.09 (0.12) [−0.31, 0.14]	−6.61 (8.12) [−22.60, 9.39]	5.67 (7.10) [−8.31, 19.65]	−6.99 (8.07) [−22.89, 8.91]	3.63 (7.16) [−10.46, 17.73]	−6.36 (6.77) [−19.69, 6.98]	−0.28 (0.26) [−0.80, 0.24]	−16.52** (7.80) [−31.88, −1.16]	−0.45** (0.22) [−0.88, −0.01]
Observations	286	286	286	285	285	285	285	286	286	286	286
Mean outcome	73.08	0.42	0.15	50.18	24.56	41.05	21.40	17.48	2.57	59.44	2.76
Additional controls	—	—	—	—	—	—	—	—	—	—	—
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. The population expands beyond within-discipline assignments. Discipline match is included, but there are no additional baseline covariates. Undergraduates were all assigned within discipline, so there is no outside-undergrad cell. Compare [Table 1](#).

Table S9: Vertical specifications: covariates, population, and assignment linkage (continued)

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	2+ appr. checks (4)	2+ qual. checks (5)	Human approp. (6)	Human overall (7)
Panel D. Linked assignments: additive							
Grad (vs undergrad)	26.85*** (9.69) [7.73, 45.97]	0.18 (0.23) [−0.26, 0.63]	0.11 (0.09) [−0.06, 0.28]	15.64** (7.23) [1.36, 29.92]	12.23* (6.87) [−1.33, 25.78]	−7.21 (8.10) [−23.20, 8.79]	0.61** (0.30) [0.01, 1.21]
Researcher (vs undergrad)	28.05*** (9.58) [9.14, 46.97]	−0.11 (0.22) [−0.55, 0.32]	0.21** (0.09) [0.04, 0.38]	20.28*** (7.27) [5.92, 34.64]	18.22** (7.25) [3.92, 32.53]	−4.86 (8.62) [−21.88, 12.16]	0.63* (0.32) [−0.01, 1.27]
AI-assisted	14.35** (5.74) [3.03, 25.67]	0.58*** (0.22) [0.14, 1.01]	0.15** (0.07) [0.00, 0.29]	15.35*** (5.45) [4.59, 26.12]	11.94** (5.32) [1.44, 22.44]	20.29*** (5.14) [10.14, 30.43]	1.04*** (0.22) [0.62, 1.47]
Observations	196	196	196	195	195	196	196
Additional controls	—	—	—	—	—	—	—
Fixed effects	✓	✓	✓	✓	✓	✓	✓
Panel E. Linked assignments: interacted							
Grad (vs undergrad)	14.22 (12.16) [−9.79, 38.23]	−0.18 (0.34) [−0.85, 0.48]	0.00 (0.06) [−0.12, 0.13]	16.64* (8.77) [−0.67, 33.95]	12.13 (8.07) [−3.81, 28.06]	−5.50 (10.34) [−25.93, 14.92]	0.70 (0.48) [−0.25, 1.64]
Researcher (vs undergrad)	17.78 (12.07) [−6.05, 41.60]	−0.27 (0.32) [−0.89, 0.36]	0.07 (0.08) [−0.09, 0.23]	25.08*** (8.54) [8.22, 41.95]	24.43*** (8.43) [7.79, 41.08]	−3.70 (11.29) [−25.99, 18.59]	0.62 (0.50) [−0.37, 1.61]
AI-assisted	−3.40 (16.78) [−36.53, 29.72]	0.17 (0.37) [−0.56, 0.90]	−0.04 (0.10) [−0.25, 0.16]	19.74* (10.14) [−0.29, 39.77]	16.50* (9.51) [−2.28, 35.28]	22.52* (13.60) [−4.34, 49.37]	1.10** (0.54) [0.03, 2.17]
AI × grad	22.96 (17.99) [−12.55, 58.48]	0.68 (0.57) [−0.44, 1.81]	0.19 (0.15) [−0.10, 0.48]	−1.48 (13.10) [−27.34, 24.38]	0.72 (12.25) [−23.47, 24.91]	−3.12 (15.24) [−33.21, 26.98]	−0.16 (0.62) [−1.38, 1.06]
AI × researcher	19.61 (19.00) [−17.91, 57.13]	0.29 (0.44) [−0.57, 1.16]	0.27* (0.16) [−0.05, 0.60]	−9.27 (13.87) [−36.66, 18.12]	−12.02 (13.35) [−38.39, 14.35]	−2.21 (16.53) [−34.85, 30.42]	0.02 (0.63) [−1.22, 1.26]
Observations	196	196	196	195	195	196	196
Additional controls	—	—	—	—	—	—	—
Fixed effects	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. No covariates; within-discipline assignments. Assignment-linked models retain exact within-event email matches to saved assignment sheets (273 of 286 reports before model exclusions); this sensitivity requires agreement with the saved treatment entry. Unmatched reports are not established violations. This post hoc documentation check is not an attrition correction.

Table S10: Horizontal specifications: complete covariates and assignment linkage

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	1+ appr. checks (4)	2+ appr. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	Human approp. (8)	Human overall (9)	AI approp. (10)	AI overall (11)
Panel A. Current imputation and missingness indicators											
Outside discipline	−10.97 (8.31) [−27.34, 5.40]	−0.28* (0.17) [−0.62, 0.05]	−0.08 (0.11) [−0.29, 0.13]	−8.34 (10.04) [−28.13, 11.44]	4.29 (8.89) [−13.24, 21.81]	−16.23* (9.65) [−35.25, 2.78]	−3.77 (8.57) [−20.67, 13.13]	−8.71 (7.38) [−23.26, 5.83]	−0.07 (0.36) [−0.78, 0.64]	−14.78 (9.46) [−33.42, 3.87]	−0.31 (0.26) [−0.82, 0.19]
AI-assisted	18.38*** (5.64) [7.26, 29.50]	0.67*** (0.25) [0.18, 1.17]	0.17* (0.09) [0.00, 0.34]	19.14** (7.44) [4.47, 33.82]	13.96** (6.46) [1.22, 26.70]	12.26* (7.39) [−2.32, 26.83]	11.04* (6.22) [−1.23, 23.31]	19.27*** (5.64) [8.15, 30.40]	0.93*** (0.23) [0.47, 1.39]	21.16*** (7.20) [6.96, 35.36]	0.94*** (0.19) [0.56, 1.32]
AI × outside	8.90 (10.80) [−12.39, 30.20]	−0.30 (0.28) [−0.85, 0.26]	−0.02 (0.20) [−0.41, 0.38]	−0.85 (13.69) [−27.84, 26.15]	0.73 (12.51) [−23.93, 25.39]	16.89 (13.65) [−10.01, 43.80]	13.53 (12.05) [−10.23, 37.28]	2.65 (9.82) [−16.70, 22.01]	−0.01 (0.40) [−0.80, 0.78]	5.62 (11.90) [−17.83, 29.07]	0.02 (0.32) [−0.62, 0.66]
Observations	249	249	249	248	248	248	248	249	249	249	249
Mean outcome	76.71	0.42	0.17	53.23	27.02	42.74	23.39	17.27	2.65	61.85	2.84
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Panel B. Observed coding years and GPT familiarity											
Outside discipline	−10.32 (8.72) [−27.51, 6.87]	−0.24 (0.18) [−0.59, 0.11]	−0.08 (0.11) [−0.31, 0.15]	−9.99 (10.48) [−30.64, 10.66]	6.46 (9.67) [−12.61, 25.54]	−20.16** (10.02) [−39.91, −0.41]	−2.10 (9.31) [−20.45, 16.25]	−10.21 (7.26) [−24.52, 4.11]	−0.02 (0.38) [−0.77, 0.73]	−19.63** (9.63) [−38.61, −0.66]	−0.44 (0.28) [−0.98, 0.11]
AI-assisted	18.72*** (5.81) [7.26, 30.18]	0.69*** (0.26) [0.19, 1.20]	0.17* (0.09) [−0.01, 0.34]	18.13** (7.58) [3.19, 33.07]	13.97** (6.68) [0.80, 27.13]	10.10 (7.51) [−4.70, 24.91]	11.06* (6.45) [−1.65, 23.77]	19.16*** (5.71) [7.90, 30.42]	0.98*** (0.24) [0.51, 1.44]	22.93*** (7.33) [8.47, 37.39]	0.99*** (0.19) [0.61, 1.37]
AI × outside	11.07 (11.40) [−11.41, 33.55]	−0.35 (0.29) [−0.91, 0.21]	0.00 (0.22) [−0.43, 0.44]	−3.33 (14.53) [−31.98, 25.33]	−1.45 (13.24) [−27.55, 24.64]	18.05 (14.31) [−10.16, 46.26]	12.24 (12.67) [−12.74, 37.22]	6.54 (9.95) [−13.09, 26.16]	0.02 (0.43) [−0.83, 0.86]	3.22 (12.33) [−21.10, 27.53]	−0.01 (0.34) [−0.67, 0.66]
Observations	238	238	238	237	237	237	237	238	238	238	238
Mean outcome	76.47	0.41	0.17	54.01	27.85	43.46	24.05	17.23	2.64	61.76	2.84
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Horizontal controls: academic tier, coding years, an indicator for missing coding-experience responses, and GPT familiarity (None/Some, Heavy, Missing). Non-undergraduates only. Panel B drops missing years/familiarity. Original Some/Heavy/Missing GPT coding is algebraically equivalent to the current coding because no None responses occur; that duplicate fit is not repeated. Missing years use event-by-tier medians, with the overall median as fallback; outcomes are not imputed.

Table S10: Horizontal specifications: complete covariates and assignment linkage (continued)

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	2+ appr. checks (4)	2+ qual. checks (5)	Human approp. (6)	Human overall (7)
Panel C. Linked assignments: additive							
Outside discipline	−8.54 (7.66) [−23.65, 6.57]	−0.45** (0.20) [−0.85, −0.05]	−0.09 (0.16) [−0.40, 0.21]	7.51 (8.12) [−8.51, 23.53]	4.44 (8.40) [−12.12, 21.00]	−2.81 (8.09) [−18.76, 13.14]	0.02 (0.33) [−0.63, 0.66]
AI-assisted	21.02*** (5.15) [10.87, 31.17]	0.62*** (0.19) [0.24, 1.01]	0.18** (0.07) [0.04, 0.32]	14.66** (5.98) [2.86, 26.46]	14.87** (5.76) [3.51, 26.23]	19.68*** (4.67) [10.47, 28.89]	0.99*** (0.19) [0.61, 1.37]
Observations	238	238	238	237	237	238	238
Additional controls	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓
Panel D. Linked assignments: interacted							
Outside discipline	−14.32* (8.64) [−31.36, 2.72]	−0.34* (0.19) [−0.73, 0.04]	−0.10 (0.12) [−0.34, 0.14]	5.93 (9.87) [−13.54, 25.39]	−2.24 (9.68) [−21.33, 16.86]	−2.52 (8.19) [−18.66, 13.62]	0.08 (0.40) [−0.71, 0.86]
AI-assisted	17.34*** (5.74) [6.02, 28.66]	0.69*** (0.26) [0.18, 1.20]	0.17* (0.09) [0.00, 0.35]	13.65** (6.67) [0.50, 26.79]	10.60 (6.45) [−2.11, 23.32]	19.87*** (5.75) [8.53, 31.20]	1.03*** (0.23) [0.58, 1.47]
AI × outside	13.02 (11.47) [−9.59, 35.64]	−0.24 (0.30) [−0.82, 0.34]	0.02 (0.21) [−0.40, 0.44]	3.57 (13.68) [−23.41, 30.55]	15.06 (13.34) [−11.25, 41.38]	−0.66 (10.12) [−20.60, 19.29]	−0.13 (0.43) [−0.97, 0.71]
Observations	238	238	238	237	237	238	238
Additional controls	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Horizontal controls: academic tier, coding years, an indicator for missing coding-experience responses, and GPT familiarity (None/Some, Heavy, Missing). Non-undergraduates only. Assignment-linked models retain exact within-event email matches to saved assignment sheets (273 of 286 reports before model exclusions); this sensitivity requires agreement with the saved treatment entry. Unmatched reports are not established violations. This post hoc documentation check is not an attrition correction. Missing years use event-by-tier medians, with the overall median as fallback; outcomes are not imputed.

Table S11: Event-by-paper fixed effects: current and available-PAP controls

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	1+ appr. checks (4)	2+ appr. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	Human approp. (8)	Human overall (9)	AI approp. (10)	AI overall (11)
Panel A. Vertical: event-by-paper and home-discipline FE											
Grad (vs undergrad)	21.73 (15.88) [−9.69, 53.14]	−0.25 (0.42) [−1.07, 0.58]	−0.06 (0.09) [−0.23, 0.11]	12.70 (15.95) [−18.86, 44.25]	11.50 (10.18) [−8.63, 31.64]	−1.19 (16.38) [−33.59, 31.22]	6.45 (9.82) [−12.98, 25.88]	−2.76 (12.56) [−27.60, 22.08]	1.01* (0.55) [−0.07, 2.10]	7.49 (17.63) [−27.38, 42.36]	0.53 (0.52) [−0.50, 1.56]
Researcher (vs undergrad)	22.77 (16.30) [−9.48, 55.03]	−0.41 (0.41) [−1.22, 0.40]	0.10 (0.09) [−0.08, 0.28]	18.65 (17.65) [−16.26, 53.56]	10.44 (9.67) [−8.69, 29.56]	9.52 (17.09) [−24.28, 43.32]	11.04 (9.20) [−7.16, 29.24]	−0.76 (13.11) [−26.69, 25.17]	0.64 (0.59) [−0.52, 1.80]	15.98 (17.64) [−18.92, 50.88]	0.64 (0.54) [−0.43, 1.72]
AI-assisted	6.92 (19.93) [−32.50, 46.34]	0.17 (0.46) [−0.74, 1.07]	−0.12 (0.13) [−0.37, 0.13]	23.98 (20.00) [−15.58, 63.53]	11.56 (10.51) [−9.24, 32.36]	19.97 (19.90) [−19.39, 59.33]	9.29 (10.06) [−10.60, 29.19]	25.68 (16.32) [−6.60, 57.95]	0.82 (0.67) [−0.50, 2.15]	15.88 (18.86) [−21.42, 53.18]	0.73 (0.63) [−0.53, 1.99]
AI × grad	8.75 (21.02) [−32.83, 50.33]	0.68 (0.66) [−0.62, 1.99]	0.31 (0.19) [−0.07, 0.70]	−3.53 (22.69) [−48.42, 41.37]	0.23 (13.84) [−27.15, 27.61]	2.04 (22.42) [−42.31, 46.40]	5.04 (13.38) [−21.43, 31.51]	−6.97 (18.20) [−42.97, 29.03]	−0.16 (0.73) [−1.61, 1.29]	9.38 (21.79) [−33.73, 52.49]	−0.01 (0.67) [−1.34, 1.31]
AI × researcher	12.93 (22.78) [−32.12, 57.98]	0.41 (0.54) [−0.66, 1.49]	0.29 (0.18) [−0.07, 0.65]	−9.27 (25.47) [−59.65, 41.12]	2.32 (15.29) [−27.93, 32.57]	−14.49 (24.68) [−63.31, 34.33]	−0.59 (15.11) [−30.49, 29.31]	−2.86 (19.47) [−41.38, 35.66]	0.42 (0.81) [−1.17, 2.02]	9.42 (24.55) [−39.15, 57.98]	0.32 (0.76) [−1.19, 1.83]
Observations	183	183	183	182	182	182	182	183	183	183	183
Mean outcome	69.40	0.52	0.10	46.15	21.43	37.91	19.23	16.94	2.45	57.92	2.77
Additional controls	—	—	—	—	—	—	—	—	—	—	—
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Panel B. Horizontal: event-by-paper and home-discipline FE											
Outside discipline	−12.29 (12.34) [−36.67, 12.09]	−0.20 (0.32) [−0.83, 0.43]	−0.07 (0.21) [−0.48, 0.35]	−23.15 (14.50) [−51.80, 5.50]	5.57 (11.33) [−16.82, 27.95]	−19.60 (14.87) [−48.99, 9.78]	10.00 (11.24) [−12.21, 32.21]	−4.58 (10.48) [−25.28, 16.11]	0.14 (0.51) [−0.88, 1.15]	−20.26 (14.07) [−48.07, 7.54]	−0.06 (0.43) [−0.91, 0.80]
AI-assisted	13.59** (6.65) [0.46, 26.72]	0.66** (0.30) [0.06, 1.25]	0.16* (0.10) [−0.03, 0.35]	11.25 (9.43) [−7.38, 29.88]	9.07 (7.97) [−6.68, 24.83]	5.89 (9.51) [−12.90, 24.69]	6.31 (8.08) [−9.66, 22.27]	19.55*** (6.88) [5.95, 33.14]	0.81*** (0.29) [0.24, 1.38]	16.58** (8.30) [0.18, 32.98]	0.81*** (0.21) [0.39, 1.23]
AI × outside	13.28 (14.79) [−15.95, 42.50]	−0.40 (0.32) [−1.02, 0.22]	−0.04 (0.19) [−0.43, 0.34]	16.47 (18.99) [−21.04, 53.99]	13.05 (12.42) [−11.47, 37.58]	30.12 (19.07) [−7.55, 67.78]	13.40 (12.55) [−11.39, 38.18]	2.10 (12.02) [−21.64, 25.83]	0.04 (0.51) [−0.97, 1.05]	20.64 (16.88) [−12.71, 53.98]	0.10 (0.47) [−0.82, 1.02]
Observations	228	228	228	227	227	227	227	228	228	228	228
Mean outcome	76.75	0.44	0.17	53.74	26.43	43.61	23.35	17.11	2.63	60.96	2.82
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event-by-paper cells and participant home discipline. Vertical has no covariates. Horizontal controls: academic tier, coding years, an indicator for missing coding-experience responses, and GPT familiarity (None/Some, Heavy, Missing). Missing years use event-by-tier medians, with the overall median as fallback; outcomes are not imputed.

Table S11: Event-by-paper fixed effects: current and available-PAP controls (continued)

	Comp. rep. (1)	Minor errors (2)	Major errors (3)	1+ appr. checks (4)	2+ appr. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	Human approp. (8)	Human overall (9)	AI approp. (10)	AI overall (11)
Panel C. Vertical: event-by-paper FE and available controls											
Grad (vs undergrad)	21.86 (15.33) [−8.47, 52.18]	−0.24 (0.41) [−1.05, 0.57]	−0.02 (0.09) [−0.21, 0.16]	10.40 (15.99) [−21.23, 42.03]	8.34 (10.59) [−12.62, 29.30]	−2.33 (16.09) [−34.16, 29.49]	4.68 (10.23) [−15.55, 24.92]	−4.56 (12.51) [−29.30, 20.18]	0.89 (0.56) [−0.21, 1.99]	7.52 (17.60) [−27.30, 42.34]	0.52 (0.53) [−0.53, 1.57]
Researcher (vs undergrad)	22.39 (16.72) [−10.69, 55.46]	−0.32 (0.40) [−1.11, 0.47]	0.20 (0.13) [−0.06, 0.46]	13.25 (19.71) [−25.75, 52.24]	3.67 (11.10) [−18.30, 25.64]	7.04 (18.91) [−30.37, 44.44]	7.90 (10.66) [−13.18, 28.99]	−5.83 (14.13) [−33.79, 22.13]	0.25 (0.64) [−1.02, 1.51]	15.79 (19.38) [−22.54, 54.12]	0.61 (0.60) [−0.59, 1.81]
AI-assisted	6.47 (19.20) [−31.52, 44.46]	0.10 (0.44) [−0.77, 0.97]	−0.12 (0.12) [−0.36, 0.11]	29.84 (19.43) [−8.60, 68.29]	19.41* (11.69) [−3.71, 42.54]	22.85 (18.75) [−14.25, 59.94]	16.06 (11.06) [−5.81, 37.94]	24.14 (15.72) [−6.95, 55.24]	0.83 (0.64) [−0.45, 2.10]	12.07 (18.66) [−24.85, 48.99]	0.62 (0.64) [−0.64, 1.88]
AI × grad	7.77 (20.01) [−31.81, 47.36]	0.82 (0.68) [−0.52, 2.17]	0.31* (0.18) [−0.05, 0.67]	−8.67 (22.26) [−52.71, 35.36]	−7.11 (14.83) [−36.44, 22.22]	−1.25 (21.38) [−43.56, 41.05]	−1.43 (14.20) [−29.52, 26.66]	−6.48 (17.58) [−41.26, 28.30]	−0.14 (0.70) [−1.53, 1.25]	14.38 (21.70) [−28.54, 57.31]	0.15 (0.68) [−1.19, 1.50]
AI × researcher	14.28 (22.37) [−29.97, 58.53]	0.42 (0.52) [−0.60, 1.45]	0.32* (0.18) [−0.03, 0.67]	−18.58 (25.07) [−68.18, 31.03]	−12.03 (16.40) [−44.47, 20.41]	−18.47 (24.10) [−66.15, 29.20]	−11.95 (16.27) [−44.15, 20.24]	−1.73 (19.62) [−40.55, 37.09]	0.40 (0.78) [−1.15, 1.94]	14.76 (24.62) [−33.95, 63.46]	0.46 (0.75) [−1.02, 1.94]
Observations	183	183	183	182	182	182	182	183	183	183	183
Mean outcome	69.40	0.52	0.10	46.15	21.43	37.91	19.23	16.94	2.45	57.92	2.77
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Panel D. Horizontal: event-by-paper FE and available controls											
Outside discipline	−11.55 (12.12) [−35.49, 12.40]	−0.01 (0.26) [−0.52, 0.49]	−0.08 (0.15) [−0.39, 0.22]	−17.60 (14.68) [−46.60, 11.39]	7.20 (10.17) [−12.89, 27.30]	−17.75 (15.27) [−47.92, 12.41]	11.34 (10.82) [−10.04, 32.72]	−5.52 (8.93) [−23.16, 12.12]	−0.06 (0.46) [−0.97, 0.86]	−19.47 (12.78) [−44.71, 5.77]	−0.15 (0.40) [−0.93, 0.64]
AI-assisted	14.10** (6.56) [1.14, 27.05]	0.67** (0.30) [0.07, 1.27]	0.19* (0.10) [0.00, 0.38]	11.15 (9.31) [−7.23, 29.53]	8.41 (7.92) [−7.23, 24.06]	5.76 (9.40) [−12.81, 24.34]	6.01 (7.99) [−9.78, 21.79]	19.34*** (6.77) [5.98, 32.70]	0.84*** (0.28) [0.28, 1.40]	16.34** (8.13) [0.29, 32.39]	0.79*** (0.21) [0.37, 1.20]
AI × outside	17.43 (15.36) [−12.91, 47.76]	−0.42 (0.33) [−1.06, 0.23]	0.01 (0.21) [−0.40, 0.43]	17.00 (18.39) [−19.32, 53.32]	13.84 (12.55) [−10.95, 38.63]	30.43 (18.40) [−5.92, 66.79]	14.34 (12.59) [−10.52, 39.20]	0.25 (11.74) [−22.94, 23.44]	0.03 (0.54) [−1.04, 1.09]	20.52 (15.91) [−10.90, 51.94]	0.11 (0.45) [−0.79, 1.01]
Observations	228	228	228	227	227	227	227	228	228	228	228
Mean outcome	76.75	0.44	0.17	53.74	26.43	43.61	23.35	17.11	2.63	60.96	2.82
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event-by-paper cells only. Both panels include coding years, missing years, and original GPT categories; horizontal also includes tier. The written PAP gives conflicting FE descriptions. Preferred software and the exact event-by-software branch remain unavailable. Missing years use event-by-tier medians, with the overall median as fallback; outcomes are not imputed.

Table S12: Sensitivity to baseline software proficiency controls

	Reproduction (pp)	Minor errors	Major errors	Two approved proposals (pp)	Two qualifying checks (pp)	Appropriate report (pp)	Overall report score
A. Within-field assignments: average AI effect							
Baseline	13.67** (5.68)	0.57*** (0.22)	0.14** (0.07)	15.60*** (5.25)	12.43** (5.12)	19.11*** (5.09)	0.92*** (0.22)
Software-adjusted	14.42** (5.81)	0.51** (0.21)	0.15** (0.06)	12.75** (5.47)	9.28* (5.38)	18.36*** (5.18)	1.00*** (0.23)
Observations (both models)	202	202	202	201	201	202	202
B. Within-field assignments: AI × graduate							
Baseline	27.21 (17.62)	0.65 (0.53)	0.19 (0.14)	−1.15 (12.43)	0.31 (11.62)	−1.65 (14.74)	−0.03 (0.62)
Software-adjusted	28.14 (17.55)	0.56 (0.55)	0.10 (0.14)	0.61 (13.34)	−0.15 (12.08)	1.15 (15.68)	0.14 (0.65)
Observations (both models)	202	202	202	201	201	202	202
C. Within-field assignments: AI × professor/researcher							
Baseline	26.95 (18.54)	0.21 (0.41)	0.28* (0.16)	−6.31 (13.28)	−9.63 (12.80)	−2.63 (16.07)	0.18 (0.62)
Software-adjusted	30.83* (17.89)	0.28 (0.51)	0.19 (0.15)	−13.86 (14.06)	−16.07 (13.92)	−0.24 (16.32)	0.24 (0.62)
Observations (both models)	202	202	202	201	201	202	202
D. Graduate and senior researchers: average AI effect							
Baseline	21.07*** (4.90)	0.58*** (0.18)	0.16** (0.07)	14.19** (5.64)	15.14*** (5.38)	20.07*** (4.60)	0.93*** (0.19)
Software-adjusted	19.17*** (5.00)	0.58*** (0.20)	0.16** (0.07)	10.79* (5.67)	11.45** (5.39)	18.22*** (4.69)	0.92*** (0.21)
Observations (both models)	249	249	249	248	248	249	249
E. Graduate and senior researchers: AI × outside field							
Baseline	8.90 (10.80)	−0.30 (0.28)	−0.02 (0.20)	0.73 (12.51)	13.53 (12.05)	2.65 (9.82)	−0.01 (0.40)
Software-adjusted	6.94 (11.39)	−0.31 (0.30)	−0.03 (0.20)	3.44 (12.21)	18.66 (11.86)	2.92 (10.37)	−0.02 (0.44)
Observations (both models)	249	249	249	248	248	249	249

Notes: Each adjusted model adds categorical baseline proficiency in R, Stata, Python, and Matlab to the corresponding main-table specification. Categories are never used, beginner, intermediate, advanced, and missing; collinear indicators are absorbed where necessary. Missing responses are retained as a category. Baseline and adjusted models use identical observations. Panels A and D report additive AI effects; Panels B and C report differences in AI effects relative to undergraduates; Panel E reports the outside-field interaction. Models retain separate event, article, and participant-discipline fixed effects. Horizontal models also retain rank, coding years, a missing-years indicator, and prior GPT familiarity. Binary effects are percentage points. Parentheses contain robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. This exploratory check controls for baseline proficiency, not preferred software or software used during the task.

Count-model, binary-link, and missing-outcome sensitivity

Table S13 reports average discrete AI effects in error-count units for the PAP count-model branch, alongside OLS on the same retained samples. Five of eight models validate against an independent implementation. The two vertical major-error fits and the PAP-cell horizontal major-error fit have unstable estimates or inference; their inferential entries are withheld. Table S14 retains every model's sample and dispersion diagnostics and the triggered negative-binomial checks.

We attempted 56 binary fixed-effect probit models. Twenty-three pass independent dummy-GLM and numerical checks; 32 exhibit separation, including every current vertical fit. One non-separated reproduction model fails independent numerical validation and is omitted from inference. Separation is checked on the full dummy design using two independent linear-programming solvers. Table S15 reports every attempted model and its sample retention; separated coefficients are not treated as valid estimates. The four count/score outcomes are not re-estimated as binary models.

Table S16 compares average AI probability effects from original-sample OLS, OLS on the probit-retained rows, and probit on those same rows. It also separates the probit product coefficient from the derivative of the AI probability contrast with respect to coding years. The latter retains negative moderation for the four robustness outcomes, while reproduction and human appropriateness remain imprecise. These are functional-form checks, not a solution to incidental-parameter bias or an estimate for excluded registrants.

Table S17 reports sensitivity to the one unavailable robustness submission. OLS linearity permits exact point-estimate bounds by assigning its binary outcomes zero or one. All additive AI effects remain positive throughout these ranges. Referee-report outcomes are now fully observed, so their displayed ranges collapse to a single estimate. These bounds describe outcome availability, not uncertainty from sampling or nonparticipation.

Table S13: Poisson effects in error-count units

A. Within-discipline assignments				
	Current FE Minor	Current FE Major	PAP cells Minor	PAP cells Major
<i>Poisson</i>				
AI gain: undergraduate	0.30 (0.58)	—	0.64 (1.25)	—
AI gain: graduate	1.05** (0.50)	—	1.80 (1.38)	—
AI gain: professor/researcher	0.55*** (0.21)	—	1.29 (0.83)	—
Graduate minus undergraduate	0.75 (0.80)	—	1.16 (1.93)	—
Researcher minus undergraduate	0.25 (0.64)	—	0.65 (1.66)	—
<i>OLS same sample</i>				
AI gain: undergraduate	0.26 (0.40)	—	0.45 (0.87)	—
AI gain: graduate	0.92** (0.44)	—	1.31 (0.82)	—
AI gain: professor/researcher	0.52** (0.25)	—	1.03** (0.44)	—
Graduate minus undergraduate	0.66 (0.57)	—	0.86 (1.18)	—
Researcher minus undergraduate	0.26 (0.46)	—	0.58 (0.97)	—
Poisson observations	172	107	106	65
Poisson inference validated	Yes	No	Yes	No

Notes: AI gains are average predicted changes in error counts when AI access switches from zero to one. Each group is standardized to the same covariate and fixed-effect distribution of the retained sample. Difference rows subtract these group-specific gains; they are not log-link interaction coefficients. OLS is re-estimated with the same specification on the Poisson-retained observations. Parentheses contain robust delta-method standard errors. Stars use unadjusted tests: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Current FE enter event, article, and home discipline separately; PAP-cell specifications and controls are documented in Table S14. Three Poisson fits fail validation; their inference is withheld, and their sample sizes remain visible.

Table S13: Poisson effects in error-count units (continued)

B. Graduate and senior researchers				
	Current FE Minor	Current FE Major	PAP cells Minor	PAP cells Major
<i>Poisson</i>				
AI gain: inside discipline	1.04*** (0.36)	1.87 (1.59)	1.65* (0.96)	—
AI gain: outside discipline	0.34* (0.17)	0.28* (0.15)	0.50* (0.29)	—
Outside minus inside	−0.70* (0.39)	−1.60 (1.58)	−1.15 (1.09)	—
<i>OLS same sample</i>				
AI gain: inside discipline	0.75** (0.29)	0.28** (0.14)	1.11** (0.52)	—
AI gain: outside discipline	0.57*** (0.21)	0.34 (0.25)	1.19* (0.69)	—
Outside minus inside	−0.18 (0.34)	0.06 (0.29)	0.08 (0.78)	—
Poisson observations	199	154	112	82
Poisson inference validated	Yes	Yes	Yes	No

Notes: AI gains are average predicted changes in error counts when AI access switches from zero to one. Each group is standardized to the same covariate and fixed-effect distribution of the retained sample. Difference rows subtract these group-specific gains; they are not log-link interaction coefficients. OLS is re-estimated with the same specification on the Poisson-retained observations. Parentheses contain robust delta-method standard errors. Stars use unadjusted tests: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Current FE enter event, article, and home discipline separately; PAP-cell specifications and controls are documented in Table S14. Three Poisson fits fail validation; their inference is withheld, and their sample sizes remain visible.

Table S14: Count-model diagnostics and triggered negative-binomial checks

Panel A. Poisson dispersion and sample retention

Outcome / model	Eligible N	Poisson N	Residual df	Deviance/df	Pearson/df	NB triggered
Minor errors / Current FE, vertical	203	172	150	1.06	1.68	No
Minor errors / Current FE, horizontal	249	199	170	1.01	1.61	No
Minor errors / PAP cells, vertical	203	106	75	1.70	1.91	Yes
Minor errors / PAP cells, horizontal	249	112	76	1.53	1.79	Yes
Major errors / Current FE, vertical	203	107	87	0.69	1.14	No
Major errors / Current FE, horizontal	249	154	127	0.76	1.51	No
Major errors / PAP cells, vertical	203	65	46	0.46	0.88	No
Major errors / PAP cells, horizontal	249	82	55	1.10	1.58	No

Panel B. Negative-binomial effects in minor-error counts

Vertical comparison	NB	OLS	Horizontal comparison	NB	OLS
AI gain: undergraduate	1.36 (1.37)	0.45 (0.87)	AI gain: inside discipline	1.39* (0.74)	1.11** (0.52)
AI gain: graduate	1.59 (1.17)	1.31 (0.82)	AI gain: outside discipline	0.61* (0.37)	1.19* (0.69)
AI gain: professor/researcher	1.35* (0.78)	1.03** (0.44)	Outside minus inside	−0.78 (0.90)	0.08 (0.78)
Graduate minus undergraduate	0.23 (1.77)	0.86 (1.18)			
Researcher minus undergraduate	−0.01 (1.65)	0.58 (0.97)			
Observations	106	106	Observations	112	112

Notes: The post-PAP numerical NB trigger is deviance/df > 1.5; Pearson dispersion can differ. Poisson removes all-zero FE groups. Current FE are separate event, paper, and home discipline; PAP cells are event-by-paper only. Current vertical omits controls; Horizontal controls: academic tier, coding years, an indicator for missing coding-experience responses, and GPT familiarity (None/Some, Heavy, Missing). PAP models use coding years, missing years, original GPT categories, and tier for horizontal. Both NB fits converge on Poisson-retained rows; Entries are average discrete AI effects in error-count units, standardized to the retained sample as in Table S13, with robust delta-method SEs in parentheses. Difference rows compare these effects. OLS uses the same retained sample. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The PAP-cell horizontal major-error Poisson covariance required repair; its inference is not treated as reliable. These diagnostics do not replace OLS.

Table S15: Probit feasibility and sample retention

	Comp. rep. (1)	1+ appr. checks (2)	2+ appr. checks (3)	1+ qual. checks (4)	2+ qual. checks (5)	Human approp. (6)	AI approp. (7)
Vertical, additive	176/202 S	197/201 S	164/201 S	197/201 S	164/201 S	154/202 S	192/202 S
Vertical, interacted	176/202 S	197/201 S	164/201 S	197/201 S	164/201 S	154/202 S	192/202 S
Horizontal, additive	228/249 F	248/248 S	218/248 S	244/248 S	218/248 S	208/249 F	245/249 F
Horizontal, interacted	228/249 F	248/248 S	218/248 S	244/248 S	218/248 S	208/249 F	245/249 F
Coding years, additive	263/286 F	285/285 F	246/285 F	281/285 F	246/285 F	244/286 F	281/286 F
Coding years, interacted	263/286 F	285/285 F	246/285 F	281/285 F	246/285 F	244/286 F	281/286 F
PAP cells, vertical	114/183 F	134/182 S	99/182 S	140/182 S	86/182 S	84/183 S	139/183 S
PAP cells, horizontal	111/228 N	170/227 S	136/227 S	178/227 S	130/227 S	97/228 F	141/228 F

Notes: Cells report probit N / corresponding OLS N. F = finite and independently validated; S = separation after FE removal. Of 56 fits, 23 are F, 32 are S, and 1 is N. All 14 current vertical fits separate. N marks failed independent numerical validation despite no retained-design separation; no inference is reported for that fit. Probit removes all-zero/all-one FE groups recursively; OLS removes FE singletons under the saved policy. Different Ns therefore reflect estimator retention, not new grades. Separate event, paper, and home-discipline FE apply except for PAP event-by-paper models. Available-PAP controls cannot restore unobserved preferred software. Independent dummy-GLM checks validate all finite estimates; two linear-programming solvers agree on every separation classification. Finite estimates can still face incidental-parameter bias.

Table S16: Probability-scale AI effects and coding-years moderation

	Panel A. Average AI probability effect			Panel B. Moderation per five coding years		
	OLS original (1)	OLS same rows (2)	Probit same rows (3)	OLS product (4)	Probit product (5)	Probit derivative (6)
Comp. rep.	17.08 (4.84) [7.60, 26.56]	17.63 (5.10) [7.63, 27.63]	17.75*** (5.19) [7.58, 27.91]	0.49 (4.72) [−8.80, 9.78]	0.03 (0.21) [−0.37, 0.44]	0.45 (4.70) [−8.77, 9.66]
N	286	263	263	286	263	263
1+ appr. checks	18.46 (5.68) [7.34, 29.58]	18.46 (5.68) [7.34, 29.58]	18.21*** (5.74) [6.96, 29.47]	−10.49** (5.14) [−20.61, −0.37]	−0.33** (0.16) [−0.64, −0.02]	−10.07** (4.54) [−18.97, −1.16]
N	285	285	285	285	285	285
2+ appr. checks	15.21 (4.82) [5.76, 24.66]	17.91 (5.55) [7.02, 28.79]	17.45*** (5.49) [6.69, 28.22]	−9.32** (3.89) [−16.98, −1.66]	−0.48*** (0.17) [−0.82, −0.14]	−10.72** (4.20) [−18.95, −2.49]
N	285	246	246	285	246	246
1+ qual. checks	18.39 (5.61) [7.40, 29.38]	18.67 (5.66) [7.57, 29.77]	17.89*** (5.79) [6.55, 29.24]	−11.23** (4.66) [−20.40, −2.06]	−0.37** (0.15) [−0.66, −0.09]	−10.68*** (4.14) [−18.80, −2.56]
N	285	281	281	285	281	281
2+ qual. checks	16.11 (4.59) [7.11, 25.11]	18.75 (5.28) [8.40, 29.09]	17.85*** (5.27) [7.53, 28.18]	−9.36*** (3.56) [−16.38, −2.35]	−0.55*** (0.18) [−0.90, −0.19]	−11.42*** (4.10) [−19.45, −3.39]
N	285	246	246	285	246	246
Human approp.	18.25 (4.22) [9.98, 26.53]	21.99 (5.05) [12.08, 31.89]	22.92*** (5.02) [13.08, 32.76]	−5.37 (3.56) [−12.38, 1.64]	−0.28 (0.19) [−0.64, 0.09]	−6.15 (4.10) [−14.18, 1.88]
N	286	244	244	286	244	244
AI approp.	22.88 (5.50) [12.10, 33.65]	22.90 (5.53) [12.06, 33.74]	22.12*** (5.69) [10.97, 33.27]	−8.31* (4.93) [−18.02, 1.40]	−0.29* (0.17) [−0.62, 0.05]	−8.01* (4.38) [−16.59, 0.56]
N	286	281	281	286	281	281

Notes: All columns use the pooled interacted coding-years specification. Only outcomes with finite, independently validated probit estimates appear; feasibility for all attempted fits is reported separately. Panel A averages the predicted probability change when AI switches 0 to 1, holding each report's other covariates/FE fixed. OLS original versus same rows isolates sample retention; same-row OLS versus probit isolates functional form. Panel B shows the OLS product, probit latent-index product, and average derivative of the AI probability contrast per five years. All effects are percentage points except column 5 (latent index). SEs and 95% CIs are below estimates; OLS average-effect CIs use normal critical values. Probit uses full dummy-model Robust covariance, including FE uncertainty, and a checked delta-method gradient. Stars are unadjusted (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$); Panel A stars apply only to probit. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Controls: tier, discipline match, and missing years; Years are centered at seven and divided by five. AI-assisted is evaluated at seven coding years.

Table S16: Probability-scale AI effects and coding-years moderation (continued)

	OLS original sample (1)	OLS probit sample (2)	Probit probit sample (3)
Panel C. Current horizontal			
Comp. rep.	21.35 (4.97) [11.61, 31.09]	22.56 (5.24) [12.28, 32.84]	23.32*** (5.48) [12.59, 34.05]
N	249	228	228
Human approp.	20.16 (4.60) [11.15, 29.17]	23.55 (5.47) [12.82, 34.28]	26.02*** (5.50) [15.23, 36.80]
N	249	208	208
AI approp.	23.03 (5.97) [11.33, 34.74]	22.91 (6.00) [11.15, 34.67]	22.95*** (5.78) [11.62, 34.27]
N	249	245	245
Panel D. Available-PAP vertical			
Comp. rep.	15.51 (6.92) [1.96, 29.07]	22.70 (9.52) [4.05, 41.36]	23.55*** (8.73) [6.45, 40.66]
N	183	114	114
Panel E. Available-PAP horizontal			
Human approp.	19.42 (5.84) [7.97, 30.87]	43.16 (11.90) [19.83, 66.49]	43.99*** (10.56) [23.30, 64.68]
N	228	97	97
AI approp.	22.73 (6.72) [9.56, 35.90]	38.02 (10.82) [16.81, 59.23]	35.95*** (9.81) [16.73, 55.17]
N	228	141	141

Notes: Average AI probability changes in percentage points; robust SEs and 95% normal CIs. Same-row OLS and probit use identical retained reports and averaging populations. Full dummy-model Robust covariance includes FE uncertainty. Stars apply only to probit and are unadjusted (**p* < 0.10, ***p* < 0.05, ****p* < 0.01). Current horizontal FE: event, paper, and home discipline separately; available-PAP FE: event-by-paper only. Horizontal controls include tier, coding years, missing years, and GPT familiarity; available-PAP vertical omits the tier control because tier and its interactions are reported. Vertical uses within-discipline assignments; horizontal excludes undergraduates. This compares functional form and retention, not outcomes for excluded registrants.

Table S17: One unavailable robustness submission: coefficient bounds

	1+ appr. checks (1)	2+ appr. checks (2)	1+ qual. checks (3)	2+ qual. checks (4)	Human approp. (5)	Human overall (6)
Panel A. Additive AI effects						
Vertical	21.92	15.60	17.12	12.43	19.11	0.92
Completion range	[21.59, 22.33]	[15.47, 16.21]	[16.85, 17.59]	[12.34, 13.08]	[19.11, 19.11]	[0.92, 0.92]
Unavailable outcomes	1	1	1	1	0	0
Observed / completed N	201 / 202	201 / 202	201 / 202	201 / 202	202 / 202	202 / 202
Horizontal	18.89	14.19	17.38	15.14	20.07	0.93
Completion range	[18.55, 19.46]	[14.09, 15.00]	[17.10, 18.01]	[15.05, 15.97]	[20.07, 20.07]	[0.93, 0.93]
Unavailable outcomes	1	1	1	1	0	0
Observed / completed N	248 / 249	248 / 249	248 / 249	248 / 249	249 / 249	249 / 249
Pooled coding years	18.50	15.25	18.44	16.15	18.26	0.91
Completion range	[18.29, 18.80]	[15.16, 15.67]	[18.25, 18.76]	[16.06, 16.57]	[18.26, 18.26]	[0.91, 0.91]
Unavailable outcomes	1	1	1	1	0	0
Observed / completed N	285 / 286	285 / 286	285 / 286	285 / 286	286 / 286	286 / 286
Panel B. Treatment interactions						
Vertical: AI × grad	3.08	-1.15	-0.30	0.31	-1.65	-0.03
Completion range	[2.88, 3.38]	[-1.22, -0.72]	[-0.44, 0.06]	[0.27, 0.77]	[-1.65, -1.65]	[-0.03, -0.03]
Vertical: AI × researcher	-16.20	-6.31	-25.77	-9.63	-2.63	0.18
Completion range	[-17.49, -14.21]	[-6.81, -3.53]	[-26.71, -23.43]	[-9.94, -6.67]	[-2.63, -2.63]	[0.18, 0.18]
Horizontal: AI × outside	-0.85	0.73	16.89	13.53	2.65	-0.01
Completion range	[-1.37, -0.54]	[-0.02, 0.82]	[16.30, 17.13]	[12.76, 13.60]	[2.65, 2.65]	[-0.01, -0.01]
Pooled: AI × coding years	-10.49	-9.32	-11.23	-9.36	-5.37	-0.24
Completion range	[-10.70, -10.31]	[-9.63, -9.24]	[-11.46, -11.07]	[-9.68, -9.29]	[-5.37, -5.37]	[-0.24, -0.24]

Notes: First lines give observed-data coefficients; brackets give exact point-estimate bounds, not confidence intervals. One AI-assisted participant's robustness submission is unavailable. Assigning each unknown binary robustness outcome 0 or 1 gives the displayed ranges, conditional on the covariates and sample. All submitted referee reports have grades, so their ranges collapse to the observed estimates. Binary coefficients are percentage points; the overall score retains its 0–5 units. Panel B coding years are per five years. All additive bounds remain positive. Fixed effects are event, paper, and participant discipline separately. Vertical models omit covariates; horizontal models control for tier, coding years, missing years, and GPT familiarity. Pooled coding-years models retain tier, discipline match, and additive missing years. No outcome is imputed in the main analysis.

Table S18: Multiplicity checks for component interactions

Outcome	Interaction	Unadjusted p	Holm p	BH q
A. Vertical expertise: 14 interactions				
Reproduction	Graduate	0.124	1.000	0.559
Reproduction	Professor/researcher	0.148	1.000	0.559
Minor errors	Graduate	0.227	1.000	0.636
Minor errors	Professor/researcher	0.603	1.000	0.979
Major errors	Graduate	0.160	1.000	0.559
Major errors	Professor/researcher	0.084	1.000	0.559
Two approved proposals	Graduate	0.926	1.000	0.979
Two approved proposals	Professor/researcher	0.635	1.000	0.979
Two approved + numerical result	Graduate	0.979	1.000	0.979
Two approved + numerical result	Professor/researcher	0.453	1.000	0.979
Appropriate report	Graduate	0.911	1.000	0.979
Appropriate report	Professor/researcher	0.870	1.000	0.979
Overall report score	Graduate	0.955	1.000	0.979
Overall report score	Professor/researcher	0.768	1.000	0.979
B. Horizontal expertise: 7 interactions				
Reproduction	Outside discipline	0.411	1.000	0.959
Minor errors	Outside discipline	0.292	1.000	0.959
Major errors	Outside discipline	0.938	1.000	0.979
Two approved proposals	Outside discipline	0.953	1.000	0.979
Two approved + numerical result	Outside discipline	0.263	1.000	0.959
Appropriate report	Outside discipline	0.787	1.000	0.979
Overall report score	Outside discipline	0.979	1.000	0.979

Notes: Each row tests an AI-by-expertise interaction for one original outcome. Panel A adjusts jointly across both career contrasts and all seven outcomes (14 tests); Panel B adjusts across seven discipline-match interactions. Holm controls familywise error and BH denotes the Benjamini–Hochberg false-discovery adjustment. Adjusting instead across all 21 interactions also yields no rejections at 5%.

S15. Additional outcomes and computational timing

Additional measures unpack the main tasks and show progress during the workday.

Table S19: Coding outcomes: false positives, false negatives, and workflow stages

	False positives (1)	False negatives (2)	Pre-reg. errors (3)	Regression errors (4)	Post-reg. errors (5)
Panel A. Vertical: within-discipline assignments					
Grad (vs undergrad)	0.98 (1.19) [−1.35, 3.31]	0.04 (0.27) [−0.48, 0.57]	−0.03 (0.23) [−0.48, 0.41]	−0.12 (0.08) [−0.26, 0.03]	−0.01 (0.13) [−0.25, 0.24]
Researcher (vs undergrad)	0.26 (0.56) [−0.85, 1.37]	0.12 (0.26) [−0.38, 0.62]	−0.06 (0.22) [−0.50, 0.38]	−0.03 (0.08) [−0.18, 0.12]	−0.04 (0.11) [−0.25, 0.17]
AI-assisted	0.10 (0.68) [−1.23, 1.44]	−0.21 (0.32) [−0.83, 0.41]	−0.10 (0.27) [−0.62, 0.42]	−0.02 (0.08) [−0.18, 0.14]	0.28 (0.18) [−0.08, 0.64]
AI × grad	−0.77 (1.37) [−3.45, 1.92]	−0.59 (0.47) [−1.51, 0.33]	0.59 (0.38) [−0.16, 1.34]	0.05 (0.09) [−0.13, 0.23]	0.20 (0.30) [−0.39, 0.78]
AI × researcher	0.92 (1.13) [−1.30, 3.13]	−0.47 (0.42) [−1.28, 0.35]	0.51* (0.30) [−0.09, 1.11]	0.10 (0.11) [−0.13, 0.32]	−0.11 (0.25) [−0.60, 0.38]
Observations	202	202	202	202	202
Mean outcome	1.28	7.13	0.30	0.05	0.25
Additional controls	—	—	—	—	—
Fixed effects	✓	✓	✓	✓	✓
Panel B. Horizontal: non-undergraduate assignments					
Outside discipline	−0.81 (0.61) [−2.00, 0.39]	0.30 (0.18) [−0.06, 0.66]	−0.25** (0.12) [−0.48, −0.02]	0.00 (0.09) [−0.17, 0.17]	−0.11 (0.12) [−0.34, 0.12]
AI-assisted	0.10 (0.52) [−0.91, 1.11]	−0.76*** (0.22) [−1.19, −0.33]	0.46** (0.19) [0.09, 0.83]	0.03 (0.05) [−0.06, 0.12]	0.36*** (0.13) [0.10, 0.62]
AI × outside	0.17 (0.70) [−1.21, 1.55]	0.38 (0.30) [−0.20, 0.97]	−0.26 (0.22) [−0.69, 0.17]	0.17 (0.15) [−0.13, 0.47]	−0.23 (0.17) [−0.57, 0.12]
Observations	249	249	249	249	249
Mean outcome	1.23	5.99	0.27	0.11	0.21
Additional controls	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Horizontal controls: academic tier, coding years, an indicator for missing coding-experience responses, and GPT familiarity (None/Some, Heavy, Missing). Vertical models have no covariates. False positives are unsupported claims; false negatives are known true-error groups not detected. Stage counts classify true errors by pre-regression, regression, and post-regression work.

Counting unique defects within participant and severity preserves the positive average error-discovery effects and the four positive career-stage error interactions (Table S20). This check collapses repeated records linked to the same defect without changing their classifications.

Table S20: Error discoveries counting records or unique defects

A. Within-discipline assignments				
	Minor Records	Minor Unique defects	Major Records	Major Unique defects
Average AI effect	0.57*** (0.22)	0.51*** (0.18)	0.14** (0.07)	0.14** (0.06)
AI effect: reference group	0.21 (0.35)	0.27 (0.30)	-0.05 (0.10)	-0.05 (0.09)
AI × graduate	0.65 (0.53)	0.42 (0.45)	0.19 (0.14)	0.18 (0.13)
AI × professor/researcher	0.21 (0.41)	0.17 (0.37)	0.28* (0.16)	0.30** (0.15)
B. Graduate and senior researchers				
	Minor Records	Minor Unique defects	Major Records	Major Unique defects
Average AI effect	0.58*** (0.18)	0.50*** (0.15)	0.16** (0.07)	0.14** (0.05)
AI effect: reference group	0.67*** (0.25)	0.59*** (0.21)	0.17* (0.09)	0.17** (0.07)
AI × outside discipline	-0.30 (0.28)	-0.27 (0.24)	-0.02 (0.20)	-0.11 (0.16)

Notes: Records reproduce the main outcome definition. Unique defects count distinct adjudicated defect-group identifiers within participant and severity; classifications are not reassigned. Across the 286 reports, minor counts fall from 119 to 110 and major counts from 42 to 35, affecting five and four participants respectively. Every accepted record maps to an existing defect group. Additive models supply the average-effect row; the remaining rows come from interacted models. The reference is undergraduate in Panel A and within-discipline in Panel B. Samples, covariates, and fixed effects are unchanged from the corresponding main models. Parentheses contain robust standard errors; stars use unadjusted conventional thresholds.

Table S21: Robustness checks: individual expert criteria

	Clear (1)	Feasible (2)	Novel (3)	Valid (4)
Panel A. Vertical: within-discipline assignments				
Grad (vs undergrad)	17.41 (14.90) [−11.80, 46.62]	38.69*** (13.70) [11.83, 65.55]	11.92 (16.44) [−20.31, 44.15]	25.09* (14.96) [−4.23, 54.42]
Researcher (vs undergrad)	24.16 (15.56) [−6.34, 54.66]	42.09*** (13.93) [14.80, 69.39]	20.40 (16.56) [−12.06, 52.86]	25.51 (15.61) [−5.10, 56.11]
AI-assisted	9.17 (16.94) [−24.04, 42.38]	44.92*** (16.93) [11.73, 78.11]	18.03 (18.59) [−18.40, 54.47]	18.48 (17.02) [−14.87, 51.84]
AI × grad	17.13 (19.36) [−20.82, 55.08]	−16.48 (19.15) [−54.02, 21.05]	13.21 (21.22) [−28.39, 54.81]	−0.44 (19.40) [−38.47, 37.59]
AI × researcher	1.77 (20.66) [−38.72, 42.27]	−17.08 (19.54) [−55.37, 21.22]	9.46 (21.63) [−32.94, 51.86]	−6.27 (20.38) [−46.21, 33.67]
Observations	201	201	201	201
Mean outcome	66.17	71.14	63.18	68.66
Additional controls	—	—	—	—
Fixed effects	✓	✓	✓	✓
Panel B. Horizontal: non-undergraduate assignments				
Outside discipline	3.43 (9.36) [−14.92, 21.77]	−7.22 (8.73) [−24.34, 9.89]	−8.22 (9.08) [−26.02, 9.59]	−8.72 (10.30) [−28.91, 11.47]
AI-assisted	17.41** (7.16) [3.38, 31.43]	27.07*** (6.18) [14.94, 39.19]	28.95*** (6.84) [15.53, 42.36]	15.60** (6.81) [2.26, 28.94]
AI × outside	2.62 (12.71) [−22.28, 27.53]	−7.63 (10.93) [−29.05, 13.79]	−20.55* (12.11) [−44.28, 3.18]	7.37 (12.20) [−16.53, 31.28]
Observations	248	248	248	248
Mean outcome	70.97	78.23	69.76	73.39
Additional controls	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Horizontal controls: academic tier, coding years, an indicator for missing coding-experience responses, and GPT familiarity (None/Some, Heavy, Missing). Vertical models have no covariates. Each outcome indicates at least one proposal satisfying the named binary criterion. These are distinct from the joint all-four-criteria approval rule. Unknown expert grades remain missing; the earlier criterion-level export had incorrectly coded these reports as zero. Reports with no submitted checks retain their rubric zero.

Table S22: Referee outcomes: human components and local AI assessments

	Summary accuracy (1)	Literature placement (2)	Weakness diagnosis (3)	Actionable recomm. (4)	Clarity and style (5)	AI approp. (6)	AI overall (7)
Panel A. Vertical: within-discipline assignments							
Grad (vs undergrad)	1.11** (0.50) [0.13, 2.10]	0.06 (0.38) [−0.69, 0.81]	0.77* (0.44) [−0.09, 1.62]	0.76 (0.46) [−0.15, 1.66]	1.03** (0.45) [0.15, 1.92]	12.96 (14.91) [−16.26, 42.17]	0.69 (0.44) [−0.16, 1.55]
Researcher (vs undergrad)	0.85* (0.51) [−0.16, 1.86]	0.29 (0.41) [−0.51, 1.09]	0.83* (0.45) [−0.05, 1.71]	0.93** (0.47) [0.01, 1.86]	1.21** (0.47) [0.28, 2.14]	23.99* (14.35) [−4.14, 52.13]	0.81* (0.44) [−0.06, 1.67]
AI-assisted	1.26** (0.62) [0.05, 2.47]	0.77 (0.48) [−0.17, 1.72]	0.55 (0.47) [−0.37, 1.48]	0.91* (0.50) [−0.06, 1.89]	1.11** (0.54) [0.05, 2.18]	26.18 (16.77) [−6.69, 59.05]	0.97* (0.54) [−0.09, 2.04]
AI × grad	−0.56 (0.72) [−1.97, 0.84]	−0.18 (0.52) [−1.20, 0.85]	0.24 (0.58) [−0.90, 1.38]	0.00 (0.62) [−1.22, 1.22]	−0.30 (0.63) [−1.53, 0.93]	−0.87 (20.15) [−40.37, 38.62]	−0.18 (0.60) [−1.35, 0.99]
AI × researcher	−0.34 (0.74) [−1.78, 1.10]	−0.27 (0.55) [−1.35, 0.81]	0.27 (0.61) [−0.92, 1.47]	0.02 (0.62) [−1.21, 1.24]	−0.55 (0.64) [−1.80, 0.69]	−4.86 (19.39) [−42.86, 33.14]	0.08 (0.62) [−1.13, 1.28]
Observations	202	202	202	202	202	202	202
Mean outcome	2.87	1.14	2.41	2.33	2.93	57.92	2.76
Additional controls	—	—	—	—	—	—	—
Fixed effects	✓	✓	✓	✓	✓	✓	✓
Panel B. Horizontal: non-undergraduate assignments							
Outside discipline	0.12 (0.37) [−0.61, 0.85]	−0.27 (0.26) [−0.77, 0.23]	−0.10 (0.39) [−0.86, 0.66]	−0.23 (0.43) [−1.08, 0.62]	0.16 (0.32) [−0.47, 0.80]	−14.78 (9.46) [−33.31, 3.76]	−0.31 (0.26) [−0.82, 0.19]
AI-assisted	0.81*** (0.25) [0.31, 1.30]	0.58*** (0.18) [0.22, 0.93]	0.76*** (0.26) [0.26, 1.27]	0.87*** (0.26) [0.36, 1.38]	0.69*** (0.23) [0.24, 1.15]	21.16*** (7.20) [7.04, 35.28]	0.94*** (0.19) [0.56, 1.31]
AI × outside	0.00 (0.45) [−0.89, 0.88]	0.57* (0.34) [−0.10, 1.23]	0.08 (0.42) [−0.74, 0.91]	0.13 (0.44) [−0.73, 0.99]	0.29 (0.41) [−0.51, 1.08]	5.62 (11.90) [−17.70, 28.93]	0.02 (0.32) [−0.62, 0.65]
Observations	249	249	249	249	249	249	249
Mean outcome	3.00	1.19	2.63	2.41	3.20	61.85	2.84
Additional controls	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Horizontal controls: academic tier, coding years, an indicator for missing coding-experience responses, and GPT familiarity (None/Some, Heavy, Missing). Vertical models have no covariates. The first five outcomes are human component scores (0–5). AI appropriateness is binary and AI overall is scored 0–5. The local grader used redacted narratives; historical blinding and missing narrative output fields are discussed in Materials and Methods. The AI exercise is not an independent external-model replication.

Table S23: One- and two-check thresholds: additive and interacted models

	1+ appr. checks (1)	1+ appr. checks (2)	2+ appr. checks (3)	2+ appr. checks (4)	1+ qual. checks (5)	1+ qual. checks (6)	2+ qual. checks (7)	2+ qual. checks (8)
Specification	Additive	Interacted	Additive	Interacted	Additive	Interacted	Additive	Interacted
Panel A. Vertical: within-discipline assignments								
Grad (vs undergrad)	17.00* (9.47) [−1.70, 35.70]	15.65 (12.95) [−9.91, 41.21]	16.05** (6.99) [2.26, 29.84]	16.78** (8.44) [0.13, 33.44]	6.55 (9.35) [−11.90, 25.01]	7.19 (13.33) [−19.12, 33.49]	12.51* (6.62) [−0.56, 25.58]	12.52 (7.82) [−2.93, 27.96]
Researcher (vs undergrad)	20.23** (9.90) [0.69, 39.78]	28.39** (13.70) [1.35, 55.43]	20.37*** (6.97) [6.61, 34.13]	23.58*** (8.19) [7.41, 39.75]	10.78 (9.75) [−8.48, 30.03]	23.83* (13.58) [−2.99, 50.65]	18.10*** (6.94) [4.41, 31.79]	22.97*** (8.11) [6.95, 38.98]
AI-assisted	21.92*** (6.52) [9.06, 34.78]	27.06* (15.81) [−4.14, 58.26]	15.60*** (5.25) [5.24, 25.96]	18.60** (9.28) [0.28, 36.91]	17.12** (6.61) [4.07, 30.18]	27.51* (15.54) [−3.16, 58.18]	12.43** (5.12) [2.33, 22.53]	16.13* (8.71) [−1.07, 33.33]
AI × grad	—	3.08 (17.94) [−32.33, 38.48]	—	−1.15 (12.43) [−25.68, 23.39]	—	−0.30 (17.95) [−35.73, 35.13]	—	0.31 (11.62) [−22.62, 23.24]
AI × researcher	—	−16.20 (19.99) [−55.66, 23.25]	—	−6.31 (13.28) [−32.53, 19.90]	—	−25.77 (19.03) [−63.33, 11.80]	—	−9.63 (12.80) [−34.89, 15.63]
Observations	201	201	201	201	201	201	201	201
Mean outcome	47.76	47.76	22.39	22.39	39.80	39.80	19.90	19.90
Additional controls	—	—	—	—	—	—	—	—
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Panel B. Horizontal: non-undergraduate assignments								
Outside discipline	−8.73 (8.13) [−24.74, 7.29]	−8.34 (10.04) [−28.13, 11.44]	4.62 (7.25) [−9.67, 18.91]	4.29 (8.89) [−13.24, 21.81]	−8.59 (8.06) [−24.48, 7.30]	−16.23* (9.65) [−35.25, 2.78]	2.35 (7.34) [−12.12, 16.82]	−3.77 (8.57) [−20.67, 13.13]
AI-assisted	18.89*** (6.38) [6.32, 31.46]	19.14** (7.44) [4.47, 33.82]	14.19** (5.64) [3.06, 25.31]	13.96** (6.46) [1.22, 26.70]	17.38*** (6.30) [4.96, 29.80]	12.26* (7.39) [−2.32, 26.83]	15.14*** (5.38) [4.55, 25.74]	11.04* (6.22) [−1.23, 23.31]
AI × outside	—	−0.85 (13.69) [−27.84, 26.15]	—	0.73 (12.51) [−23.93, 25.39]	—	16.89 (13.65) [−10.01, 43.80]	—	13.53 (12.05) [−10.23, 37.28]
Observations	248	248	248	248	248	248	248	248
Mean outcome	53.23	53.23	27.02	27.02	42.74	42.74	23.39	23.39
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓

Notes: OLS with heteroskedasticity-robust SEs in parentheses and 95% CIs below. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary coefficients and means are in percentage points; counts and 0–5 scores retain their units. Tests are unadjusted unless stated. Fixed effects: event, assigned paper, and participant home discipline, entered separately. Horizontal controls: academic tier, coding years, an indicator for missing coding-experience responses, and GPT familiarity (None/Some, Heavy, Missing). Vertical models omit covariates. Approval requires all four criteria; qualification also requires a recorded numeric result, not independently verified execution. Missing years use event-by-tier medians, with the overall median as fallback; outcomes are not imputed.

Timing outcomes

Error eligibility is defined at the paper level. A participant enters the any-error sample if the assigned paper has at least one accepted minor or major defect in the pooled adjudication record. The major-error sample requires at least one accepted major defect. All participants assigned an eligible paper can enter, including those who found nothing. The opportunity set is determined from adjudicated discoveries after the assignment, rather than from a fixed catalogue known before treatment. This selection limits causal interpretation of the error curves.

Reproduction retains 282 of 286 participants. It excludes one missing event time, two code-running durations, and one ambiguous clock entry. For any error, 28 lack a paper-level opportunity and one lacks a usable event time, leaving 257. For major errors, 109 lack a paper-level opportunity; all remaining 177 have usable timing. Participants with no discovery remain at 420 minutes.

Figure 1 compares milestone progress by AI access; Figure S2 separates within- and outside-discipline assignments. Each panel reports milestone completions and eligible participants. Appendix Table A1 reports the principal timing estimates. Table S25 removes the restriction to papers with an observed error opportunity. It retains participants assigned papers without an accepted discovery, assigning 420 minutes to those non-discoveries. The timing effects remain negative under this broader comparison.

Table S24: Timing checks excluding one paper or event at a time.

Milestone	Sample	Excluded	Fits	Effect range (min)	p range	$p < 0.05$
Reproduction	Within discipline	Paper	20	[−65.2, −33.0]	[0.001, 0.072]	19/20
Reproduction	Within discipline	Event	5	[−70.1, −31.0]	[< 0.001, 0.138]	4/5
Reproduction	Graduate/senior	Paper	20	[−94.0, −66.8]	[< 0.001, < 0.001]	20/20
Reproduction	Graduate/senior	Event	5	[−89.7, −60.4]	[< 0.001, < 0.001]	5/5
First true error	Within discipline	Paper	16	[−71.3, −48.1]	[< 0.001, 0.007]	16/16
First true error	Within discipline	Event	5	[−68.8, −48.6]	[< 0.001, 0.008]	5/5
First true error	Graduate/senior	Paper	16	[−74.3, −58.0]	[< 0.001, 0.001]	16/16
First true error	Graduate/senior	Event	5	[−88.3, −51.4]	[< 0.001, 0.008]	5/5
First true major error	Within discipline	Paper	12	[−46.9, −27.2]	[0.017, 0.103]	6/12
First true major error	Within discipline	Event	5	[−43.3, −22.9]	[0.050, 0.168]	0/5
First true major error	Graduate/senior	Paper	12	[−49.2, −35.3]	[0.026, 0.115]	8/12
First true major error	Graduate/senior	Event	5	[−47.3, −30.1]	[0.029, 0.168]	2/5

Notes: Each fit excludes one paper or event from the adjusted models in Table A1. Ranges cover the saved estimates and two-sided robust tests. All estimates remain negative; statistical support varies, especially for major errors. The final column counts fits with $p < 0.05$. Paper counts differ because the three milestones have different eligible samples. These checks preserve the original controls, fixed effects, and 420-minute horizon.

Table S25: Error timing without selecting papers by observed discoveries

Milestone / comparison	Opportunity sample	All papers	<i>N</i> selected	<i>N</i> all
First true error				
Unadjusted	−47.12*** (15.45)	−42.74*** (14.19)	257	285
Within discipline	−57.89*** (16.38)	−54.62*** (15.58)	189	202
Graduate and senior	−63.73*** (17.15)	−58.89*** (15.53)	224	248
First major error				
Unadjusted	−29.95** (13.92)	−19.82** (8.94)	177	286
Within discipline	−33.17** (16.05)	−21.80** (10.15)	125	202
Graduate and senior	−39.63** (18.70)	−22.70** (10.51)	154	249

Notes: Effects are minutes of restricted time to discovery. Opportunity samples retain papers with an accepted error of the relevant severity in the pooled record. All-paper samples additionally retain participants assigned papers with no accepted discovery, assigning 420 minutes to those non-discoveries. Other timing exclusions remain unchanged. This removes the paper-opportunity restriction; it does not establish that no undiscovered errors exist. Adjusted models follow Appendix Table A1. Negative effects combine earlier discovery with greater discovery probability. Parentheses contain robust standard errors; stars use unadjusted conventional thresholds.

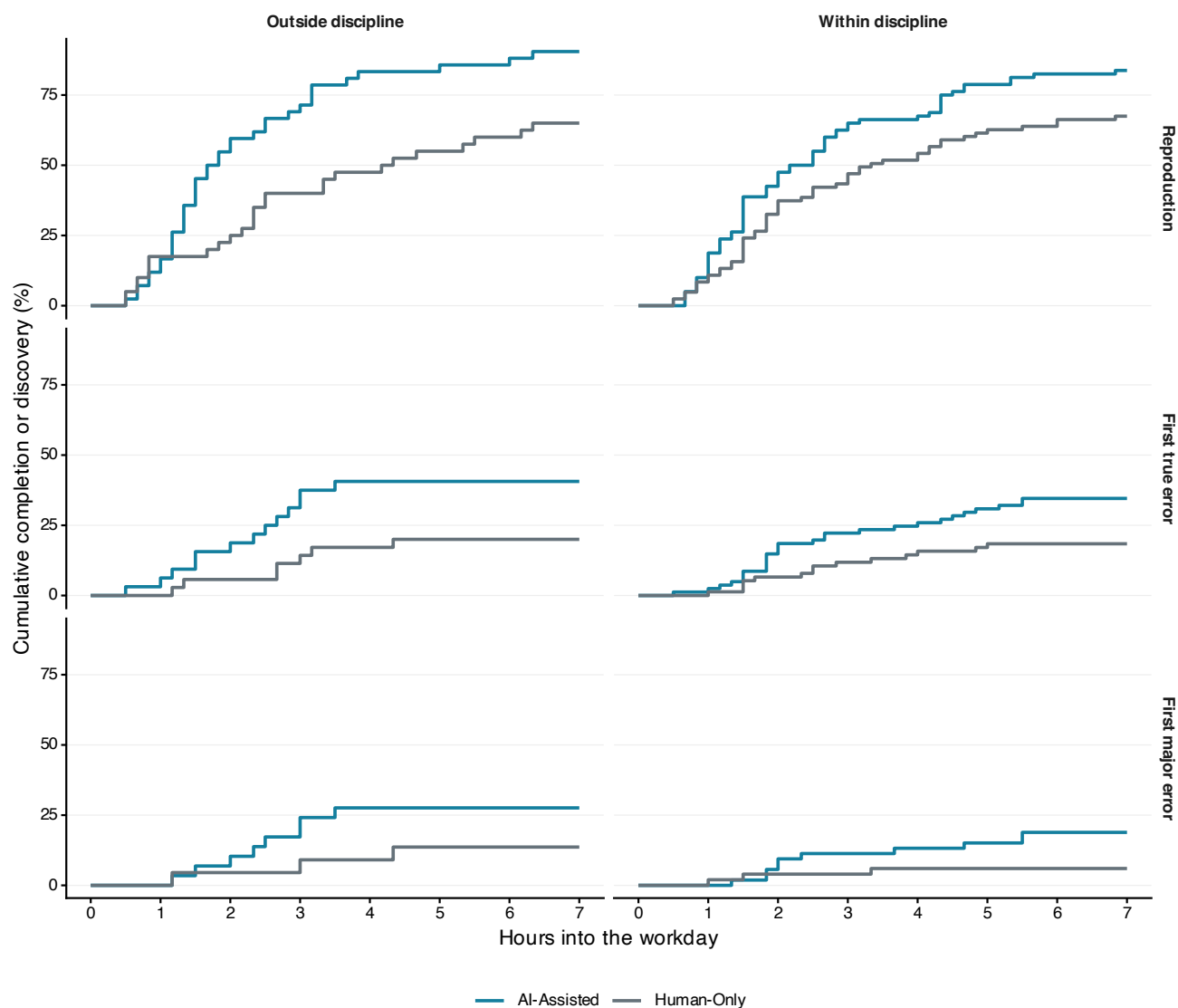


Figure S2: Computational progress within and outside participants' fields. The three rows track reproduction, the first true coding error of either severity, and the first true major coding error. Columns distinguish studies within and outside the participant's discipline. Curves use eligible graduate and senior participants. Each panel shows the events and sample sizes for AI-assisted and human-only participants. Error samples require a paper with an accepted defect of the relevant severity in the pooled adjudication record, plus usable timing records. Non-events remain in the sample through 420 minutes. These are descriptive comparisons within the indicated samples.

SI6. Participant debriefs and reported AI use

Participant accounts describe how assistance entered the workflow. The survey adds descriptions of tool use and perceived effects.

Participant debriefs: corpus and analytic approach

Four semi-structured small-group debriefs recorded at Ottawa on October 16, 2025 describe participants' workflows. Introductions identify groups of three, two, three, and four participants, totaling twelve; recordings contain approximately 129 minutes. Two discussions involve human-only participants and two involve AI-assisted participants. The latter include one outside-discipline group and one mixed group with three within- and one outside-discipline assignment. Composition is inferred from recorded descriptions rather than a reconciled recruitment roster. These are small-group debriefs, not representative samples of the experimental cells.

We adapted the codebook-based workflow in Brodeur et al. [19], retaining theory-seeded categories, inductive refinements, timestamped evidence, and comparison of convergent and divergent accounts. The complete machine transcripts were read and reviewed with AI assistance. The archived ledger contains 68 selected meaning units, including one ambiguous passage excluded from substantive interpretation and one moderator-only context passage. Each record includes source times, codes, epistemic status, a paraphrase, and a qualification. The original coding and subsequent decisions remain separately visible. Unlike the earlier paper's five-analyst procedure, this is a single AI-assisted review, not independent human consensus.

Coding distinguishes event accounts from prior practice, expectations, recommendations, moderator prompts, and ambiguous speech. Questions and qualifications remain attached to their responses; moderator agreement is not additional participant evidence. Negative cases include code that needed little assistance, unimplemented AI suggestions, divergent confidence, and useful versus excessive package recommendations. The recordings contain participant introductions, but the transcripts do not reliably identify the speaker of every later contribution. We therefore do not link individual passages to grades or coding years. The analysis was informed by the quantitative results and does not claim blinded coding, saturation, prevalence, or causal mediation. Paraphrased accounts accompany relevant results; Table A3 retains the synthesis.

The PAP planned four groups of six participants, each lasting sixty minutes. The four available recordings instead contain twelve participants in total and last approximately 23 to 44 minutes each. Participant counts come from introductions in the recordings. One participant mentioned still writing their report when called to the discussion.

Appendix Table A3 summarizes the recorded sessions and contrasting accounts.

How participants described using AI in the two studies

We compared accounts from the current debriefs with retrospective discussions of the earlier experiment [19]. One earlier participant described receiving generic robustness suggestions and then developing the checks themselves. A current participant described choosing checks and asking AI to write and revise the code. Another current participant described an AI suggestion they did not implement. These examples illustrate different uses of assistance, rather than measuring how those uses changed over time.

An earlier participant also expected to perform better if the exercise were repeated, attributing this to both improved models and greater personal experience using AI. This was their expectation, not the result of a repeated experiment. The comparison uses discussions about the 2024 events, rather than accounts of the additional exercise held in 2025.

Exit-survey coverage

Three AI-assisted participants did not have matched exit-survey responses. Their task outcomes remain in the quantitative analysis. Survey summaries use available responses.

Table S26: AI-arm survey: intensity, perceived impact, and tools**Panel A. Distributions by academic tier**

Measure / tier	N	Mean	SD	P10	P25	Median	P75	P90
Intensity / All	143	80.5	19.9	52.4	75.0	85.0	95.0	100.0
Impact / All	143	71.0	35.9	23.4	57.5	85.0	100.0	100.0
Intensity / Undergrad	21	79.4	19.8	55.0	70.0	81.0	98.0	100.0
Impact / Undergrad	21	65.9	49.5	-12.0	62.0	85.0	100.0	100.0
Intensity / Grad	58	83.3	15.4	62.1	80.0	85.0	95.0	100.0
Impact / Grad	58	72.7	34.4	26.4	61.0	85.0	100.0	100.0
Intensity / Researcher	64	78.2	23.3	50.6	70.0	84.5	95.2	100.0
Impact / Researcher	64	71.1	32.4	22.3	46.0	81.0	100.0	100.0

Panel B. Response ranges

Measure	Range	Respondents	Share
Intensity	0	0	0.0%
Intensity	1-25	5	3.5%
Intensity	26-50	7	4.9%
Intensity	51-75	31	21.7%
Intensity	76-90	49	34.3%
Intensity	91-100	51	35.7%
Impact	-100 to -50	2	1.4%
Impact	-49 to -1	5	3.5%
Impact	0	1	0.7%
Impact	1-50	25	17.5%
Impact	51-99	64	44.8%
Impact	100	46	32.2%

Panel C. Tools used during the event

Tool	Respondents	Share
ChatGPT	116	81.1%
Claude	29	20.3%
Claude Code	28	19.6%
OpenAI Codex	25	17.5%
Gemini	8	5.6%
Other AI tool	7	4.9%
Notebook LM	5	3.5%
GitHub Copilot	3	2.1%
Perplexity AI	3	2.1%
DeepSeek	3	2.1%

Notes: Denominators are 143 linked AI-arm respondents among 146 AI-assisted reports. Intensity ranges 0–100 and perceived impact ranges -100–100; both have 143 responses. Quantiles use empirical interpolation. Tools permit multiple selections. These post-treatment self-reports are descriptive, not objective productivity or exogenous causal moderators.

Table S27: AI assistance improves the reported experience of analytical work

	Human-only	AI-assisted	Adjusted difference	<i>p</i>	<i>N</i>
A. Enjoyment					
Event overall	77.7	85.9	10.3 (2.6)	< 0.001	243
Coding	59.9	79.4	21.9 (3.2)	< 0.001	243
Proposing checks	52.0	72.1	20.0 (3.8)	< 0.001	243
Writing the report	59.9	69.8	9.6 (3.4)	0.005	243
B. Expected performance					
Reproduction	65.8	81.5	16.7 (3.4)	< 0.001	243
Finding errors	47.3	61.2	17.4 (3.9)	< 0.001	238
Proposing checks	42.4	68.9	26.8 (3.5)	< 0.001	243
Writing the report	57.3	72.5	15.6 (3.0)	< 0.001	243

Notes: Arm means and adjusted differences use available responses from Toronto, Cambridge, UCL, and KCL. Ottawa has no responses to these items. Each row uses 123 AI-assisted and 120 human-only responses, except expected error-finding performance (119 and 119). Adjusted differences include event, assigned-paper, and participant-discipline fixed effects; robust standard errors appear in parentheses. All eight adjusted differences remain significant at 1% after Holm correction across these exploratory outcomes. Expected performance is self-assessed, not an expert rating. Perceived effects of access are asked separately: AI-assisted respondents report an average benefit of 71.0 ($N = 143$), while human-only respondents report an average effect of denied access of -38.9 ($N = 140$), on the -100 to 100 scale. The differently worded questions are not a common-outcome treatment contrast.

SI7. Analysis plan, protocol record, and grading instructions

This section records the analysis plan, implemented procedures, and remaining documentation checks.

Editorial record. Study-specific ethics and consent documentation, historical grading procedures, participant flow, funding details, competing interests, author approval, and the final public release remain subject to the verification described in the revision notes. Statements from the earlier team study are not substituted for this experiment's records. The participant debrief material is exploratory and retains its original consent and source-review limitations.

Registration and protocol changes

The [analysis plan](#) and [amendment](#) are available on GitHub. The saved deposit audit concerns their original public OSF record. That project was created on October 9, 2025. The plan's first public file version was uploaded on October 16, 2025, just before the start of the Ottawa event. The amendment is dated April 16, 2026; its first public upload was April 29, just before the Toronto event. These historical dates describe the original deposit, not the later GitHub copy.

The amendment added the discipline Management, password-protected advance package distribution, five-category secondary expertise analyses, and expanded undergraduate eligibility; it retained the three-tier primary design. Intake records use several undergraduate eligibility labels, including second-year-and-above and final-year-only categories. The undergraduate group included final-year students and students about to complete, or who had just completed, their penultimate year. These intake labels do not consistently distinguish the two stages.

Measurement and grading audit

The final expert-review snapshot was taken on September 17, 2026. The retained analysis sample contains expert grades for all 275 submitted referee reports. Eleven participants submitted no referee text and receive rubric zeros, giving 286 observed referee outcomes. The direct appropriateness field contains "Scored" rather than a binary judgment. We construct appropriateness from the five component scores. A rubric zero for non-submission is not an expert-assigned grade.

The retained sample contains 442 proposed checks, all reviewed. Of these, 321 contain a digit in the result or coefficient field. No numeric-positive record triggered a restricted screen for explicit nonexecution or year-only entries, but that screen cannot verify execution or correctness. One participant's retained submission contains only a referee-report sheet. Their robustness outcomes remain unknown because the underlying submission is unavailable, not because an expert grade is pending. [Table S17](#) reports endpoint sensitivity for that observation; all submitted referee reports now have grades.

The current review cards omit explicit participant and treatment fields but display the paper title, discipline, journal, and PDF link. They therefore do not meet the PAP's full paper-discipline-blinding description. Archived code also displayed treatment, but code alone does not establish which interface was deployed for each review. The database retains one review per check or report, without independent replicate ratings or a reconciliation history. Historical blinding and consensus cannot be certified from that archive. The complete archive contains 277 saved AI grade objects with appropriateness, overall, and notes, but none retains separate justifications or report-section objects. Section [SI7](#) reproduces the requested instructions; it does not recover missing outputs.

Assignment linkage and baseline availability

Saved assignment sheets contain 487 participant-event records and 474 unique emails, without duplicate emails within an event. Exact within-event matching through canonical, intake, and exit email aliases links 274 of 286 analyzed reports. One linked treatment entry differs from the communicated AI-assisted instructions documented in Section [SI1](#). The 12 unmatched reports are not established departures from randomization, and the 213 saved assignment records without retained analysis reports are not a verified attrition denominator. Eleven Toronto discipline entries labeled Other/Unsupported were recoded to study fields using author-provided corrections recorded on May 19, 2026. The correction ledger records the revised categories, without independent source verification for each recode. Three linked records lack paper titles. The flow and grading counts appear together in [Table A2](#).

The archived randomization script first randomizes AI access within event and five position categories, balancing arm counts to within one participant per category. It then randomly selects outside-field assignments among non-undergraduates, targeting 30% of the event roster, and draws a paper from the eligible field pool. Its catalogue covers three disciplines. The analysis plan instead describes three broad academic strata. These records document intended procedures and an archived implementation, but do not establish which script version and adjustments generated every final assignment across all five events and four fields.

Sixteen analyzed reports lack coding years: eleven have blank matched intake answers, three rely on manual

registration records, and two lack intake matches. The missing-years indicator is additive; complete-case results retain the original model controls. Intake proficiency ratings for R, Stata, Python, and Matlab do not identify preferred software. Software used during the task is post-assignment, so it is not substituted for the unavailable baseline preference or event-by-software fixed effects. Table S12 instead adds the four baseline proficiency ratings as categorical controls.

Table S28: PAP coverage and implementation

This table separates changes in implementation from analyses reported in different locations.

Planned analysis or procedure	Implementation	Reason or explanation	Location
Outcome-specific estimators	OLS is primary. Alternative nonlinear estimates and failed attempts remain in the appendix.	Some nonlinear fits suffered separation or unstable estimates and covariance.	Tables S13–S16
Event-by-article fixed effects	Main models enter event, article, and home discipline separately. Cell-FE estimates are retained.	Singleton cells reduce sample retention. For reproduction, vertical N falls from 202 to 183; horizontal N falls from 249 to 228.	Table S11
Baseline controls and software	Vertical models omit additional covariates. Alternatives include baseline covariates and software proficiency.	Some collected variables describe post-assignment behavior. Intake records proficiency, not preferred software; the exact event-by-preferred-software branch remains unavailable.	Tables S9 , S12
Field-assignment comparison	Horizontal models include graduate and senior researchers.	Only these groups were randomized across field assignments; the PAP assigns undergraduates inside their discipline.	Methods; Section S11
Non-coding analyses: PAP Tables 9 and 10	Proposal and human-report outcomes appear in the main tables; thresholds, components, and AI assessments appear in supplementary tables.	These analyses were reorganized across exhibits rather than omitted.	Main Tables 1–2 ; Tables S23 , S22

Table S28: PAP coverage and implementation (continued)

This table separates changes in implementation from analyses reported in different locations.

Planned analysis or procedure	Implementation	Reason or explanation	Location
Human report appropriateness	Appropriateness requires at least three on every component.	The expert interface collected component scores rather than the planned direct binary judgment.	Section SI1
Error severity and numerical robustness	Major errors reflect potential impact; numerical robustness outcomes screen result fields.	Retained records do not uniformly document a rerun demonstrating impact or independently verify robustness execution.	Sections SI1 , SI7
Expertise and missing outcomes	Planned expertise analyses remain. Pooled coding-experience choices and the prior-AI cutoff are post hoc. Missing robustness outcomes are bounded.	Supplementary coding-experience implementation was selected after examining results. One robustness submission remains unavailable; final expert grades are included.	Tables S3–S6 , SI7
Grading procedures	Human and AI grading procedures are documented separately.	Historical human blinding/consensus is unverified; AI narrative fields were not retained. Repeated independent human ratings are unavailable.	Sections SI1 , SI7
Dated protocol commitments	PAP/amendment links and original public deposit dates are documented.	Public uploads occurred on the Ottawa/Toronto event dates. Earlier commitment, screening, access timing, and historical implementation require dated evidence. Discipline recodes retain an author-correction ledger without independent verification.	Section SI7
Focus groups and release	Four recovered recordings supply qualitative context. Deidentified data and R code are assembled in the replication package.	Qualitative evidence is exploratory; consent and source-review limitations remain documented. Local package assembly does not establish public release.	Sections SI6–SI7

AI-grading instructions and implementation

The first block reproduces the PAP prompt. The second reproduces the common instruction archived with the 273 prepared requests used for the original local-grading shards. These are instructions, not evidence that every requested narrative field was retained. Local score validation establishes coverage and score ranges; it does not certify full prompt fidelity. Supplementary missing-queue grades were added separately. No participant text or identifiers are reproduced here.

PAP 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 workbook 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 (sheet 04_Referee_Report), their reproducibility notes and robustness log, and any supporting tables or figures they produced.

General instructions - Work independently—do not infer facts that are absent from the materials. - Keep the narrative within 1,000 words and organize it under the sections listed below; fold any code or data blockers into those sections instead of creating new headings. - Base every judgment on the participant’s reasoning and evidence; parroting the source paper without evaluation should be penalized. - Highlight robustness checks only when they satisfy the rubric (motivation, documented specification change, comparison to the focal estimate, interpretation); otherwise describe the gap. - Use the full 0–5 range for the overall score (no half points). - Provide two- to four-sentence justifications for each scored dimension, citing the relevant portions of the workbook (quote, paraphrase, or reference to a table or figure). - If critical evidence is missing, record that as a weakness, reflect it in the score, and flag it in the notes.

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. Overall assessment (0–5). Provide a holistic rating that reflects the usefulness of the referee narrative to an editor, weighing accuracy, diagnostic insight, actionable guidance, and clarity. This is not a mechanical average.

Anchors for the 0–5 scale (task 2) - 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.

Report structure Produce prose for the following sections. Each section should be a short paragraph that references the participant’s evidence and names any missing components. - Summary — question, setting, methods, and focal result (including whether it was reproduced and any tolerance issues). - Design and Identification — randomization, balance, attrition, and identification threats. - Robustness — up to two checks with rationale, specification or sample changes, and how the results compare with the focal estimate; note explicitly if required elements are absent. - Ethical and Transparency — preregistration, data access, and consent or IRB references, if any. - Overall Assessment — recommendation to editors/authors synthesizing strengths and weaknesses.

Output format Return JSON with the following fields: appropriateness (“Appropriate” or “Not appropriate”), overall, justifications, report_sections, and notes. The justifications object must contain keys for appropriateness and overall. The report_sections object must contain keys summary, design_identification, robustness, ethical_transparency, and overall_assessment. Include a single string in notes describing any missing information, suspected hallucinations, or uncertainties.

```
{
  "appropriateness": "Appropriate",
  "overall": 4,
  "justifications": {
    "appropriateness": "...",
    "overall": "..."
  },
  "report_sections": {
    "summary": "...",
    "design_identification": "...",
    "robustness": "...",
    "ethical_transparency": "...",
    "overall_assessment": "..."
  },
  "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 or section. - 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.

Archived local-grader instruction

You are AI Referee, an expert evaluator for the AI Replication Games.

Evaluate the supplied submission only. The submission is a replication of a published empirical result. The participant had the original paper, a replication package, and a seven-hour work window. Do not infer facts absent from the supplied materials and do not use external knowledge. Do not mention or infer participant identity, study arm, expertise, event, or task discipline.

Judge whether the participant correctly characterizes whether the reproduced evidence supports the original claim, flags material discrepancies, and grounds the conclusion in submitted code or outputs. Return Appropriate only if all of those conditions hold; otherwise return Not appropriate.

Give a holistic overall integer score from 0 to 5, not a mechanical average. Use the full range: 0 missing/wholly incorrect/incoherent; 1 substantially inaccurate and misleading; 2 partly correct but incomplete or materially inaccurate; 3 adequate and correct at core but limited; 4 strong, accurate, supported, and actionable with minor omissions; 5 exceptional, precise, comprehensive, and especially insightful. Weigh accuracy, diagnostic insight, actionable guidance, and clarity.

Write two to four sentences in each justification. The five report sections must be short paragraphs. Highlight a robustness check only when it has a motivation, documented specification or sample change, comparison to the focal estimate, and interpretation. If a required element is missing, identify the gap instead of inventing it. In notes, record missing information, suspected hallucinations, or uncertainty. Return only the requested JSON object.

The request schema required appropriateness, overall, justifications, five report sections, and notes. All 277 saved grade objects retain appropriateness, overall, and notes; none retains separate justifications or report_sections objects. Reproducing the prompt does not recover these missing output fields. The prompt instructs the grader not to infer treatment or expertise; minimal redaction of submitted narratives does not independently establish perfect blinding. The local exercise remains distinct from an externally replicated grading study.

SI8. Author group and affiliations

The full author list appears on the title page. Affiliation numbers below correspond to the superscripts attached to each name.

Author affiliations

¹University of Ottawa. ²Institute for Replication. ³Chemnitz University of Technology. ⁴University of Toronto. ⁵Charles University. ⁶NHH Norwegian School of Economics. ⁷University of Cambridge. ⁸University of Warwick. ⁹National Chengchi University. ¹⁰University of Oslo. ¹¹Université de Montréal. ¹²Tilburg University. ¹³Temple University. ¹⁴University of Memphis. ¹⁵Vrije Universiteit Amsterdam. ¹⁶Vienna University of Economics and Business. ¹⁷University College London. ¹⁸King's College London. ¹⁹Birkbeck, University of London. ²⁰Lancaster University. ²¹National Taiwan University. ²²Tilburg School of Economics and management. ²³Macquarie University. ²⁴The University of Sydney. ²⁵Massachusetts Institute of Technology. ²⁶Technical University of Munich. ²⁷Independent researcher. ²⁸University of Ghana. ²⁹University of Mannheim, Business School, MISES. ³⁰University of Calgary. ³¹University of Warwick and CEPR. ³²University of Southampton. ³³Carleton University, Economics. ³⁴Erasmus University. ³⁵Department of Clinical Neurosciences, University of Cambridge. ³⁶Queen's University. ³⁷RWI - Leibniz-Institute for Economic Research & Hertie School. ³⁸University of Guelph. ³⁹Maastricht University & University of Chicago. ⁴⁰Tampere University. ⁴¹Hasselt University. ⁴²Faculty of Environmental Sciences, Czech University of Life Sciences Prague. ⁴³Magna Graecia University of Catanzaro. ⁴⁴Hong Kong Baptist University. ⁴⁵Imperial College London. ⁴⁶University of Technology Sydney. ⁴⁷University of Economics Ho Chi Minh city. ⁴⁸Wilfrid Laurier University. ⁴⁹NIDI/University of Groningen. ⁵⁰University of Roehampton. ⁵¹Binghamton University. ⁵²McGill University. ⁵³La Fondation pour les études et recherches sur le développement international (Ferdi). ⁵⁴Stockholm University. ⁵⁵Bank of Canada. ⁵⁶University of Leicester. ⁵⁷Michael Okpara University of Agriculture, Umudike. ⁵⁸China Europe International Business School. ⁵⁹Department of Business Management, Masaryk University, Brno, Czech Republic. ⁶⁰Heriot-Watt University. ⁶¹European University Institute. ⁶²UAI Business School. ⁶³MRC Cognition and Brain Sciences Unit, Cambridge. ⁶⁴University College Dublin. ⁶⁵Universidad Adolfo Ibañez. ⁶⁶University of Zurich. ⁶⁷RWI - Leibniz-Institute for Economic Research. ⁶⁸SUNY Albany. ⁶⁹Leiden University. ⁷⁰University of Alabama. ⁷¹Linnaeus University and Stockholm School of Economics. ⁷²University of Kassel. ⁷³Collegio Carlo Alberto. ⁷⁴Department of Methodology and Statistics, Tilburg University. ⁷⁵University of Edinburgh. ⁷⁶Cornell University. ⁷⁷Psychology and Language Sciences, University College London. ⁷⁸Rensselaer Polytechnic Institute. ⁷⁹Hasselt University, University of Liège. ⁸⁰University of Birmingham. ⁸¹University of Göttingen. ⁸²University of Mannheim. ⁸³Linnaeus University, Department of Economics and Statistics. ⁸⁴University of Warsaw. ⁸⁵The University of Manchester. ⁸⁶Institute of Economic Studies, Charles University. ⁸⁷University of Oxford. ⁸⁸MRC Cognition and Brain Sciences Unit. ⁸⁹University of Basel and University Hospital Basel. ⁹⁰University of Kassel and Institute for Employment Research (IAB). ⁹¹Warsaw University of Life Sciences. ⁹²Corvinus University of Budapest. ⁹³SVA System Vertrieb Alexander GmbH. ⁹⁴University at Albany - State University of New York. ⁹⁵Sapienza University of Rome. ⁹⁶University of Florida. ⁹⁷University of Kentucky. ⁹⁸University of Konstanz. ⁹⁹McMaster University. ¹⁰⁰University of Vienna. ¹⁰¹University of Minnesota. ¹⁰²University of Halle. ¹⁰³Universidad Carlos III de Madrid. ¹⁰⁴Krea University. ¹⁰⁵University of Graz. ¹⁰⁶The Psychometrics Centre, University of Cambridge. ¹⁰⁷RWI / Uni Potsdam. ¹⁰⁸University of Alberta. ¹⁰⁹Technical University of Munich (TUM). ¹¹⁰University of Sheffield. ¹¹¹KU Leuven. ¹¹²Coventry University. ¹¹³Centre for Environmental Sciences, Hasselt University. ¹¹⁴London School of Economics. ¹¹⁵Cardinal Stephen Wyszyński University, Warsaw. ¹¹⁶Humboldt University zu Berlin. ¹¹⁷Southern Methodist University. ¹¹⁸Erasmus University Rotterdam. ¹¹⁹Lake Forest College. ¹²⁰Aix-Marseille School of Economics. ¹²¹Robert Koch Institute. ¹²²Johns Hopkins University. ¹²³Zeppelin University. ¹²⁴University of the Balearic Islands. ¹²⁵University of Milano-Bicocca. ¹²⁶Manchester Metropolitan University. ¹²⁷University of Nottingham. ¹²⁸Wake Forest University. ¹²⁹University of Lausanne. ¹³⁰City St George's, University of London. ¹³¹Netherlands Interdisciplinary Demographic Institute. ¹³²Università degli Studi dell'Insubria. ¹³³HEC Liege, University of Liege. ¹³⁴Department of Economics, University of Calgary. ¹³⁵Clemson University. ¹³⁶Carleton University. ¹³⁷Carnegie Mellon University in Qatar. ¹³⁸Oxford Internet Institute, University of Oxford. ¹³⁹Culture Based Development (CBD) Paradigm. ¹⁴⁰Masaryk University. ¹⁴¹University of California, Berkeley. ¹⁴²ESSCA School of Management. ¹⁴³Mainz University (Germany). ¹⁴⁴Stockholm Business School, Stockholm University. ¹⁴⁵Collegium of World Economy, SGH Warsaw School of Economics. ¹⁴⁶Stanford University. ¹⁴⁷Victoria University of Wellington. ¹⁴⁸University of Lausanne and Swiss Finance Institute.

Author contributions

Proposed group-based assignments. JPap, DVa, DMi, GABa, BBar, LFa, LFi, ABr: designed research, performed research, analyzed data, and wrote the paper.

RAI, VABu, BBak, PDa, UDu, EGa, SGe, AGKa, AGu, THa, YSHu, AJa, DLe, AMa, XMa, JSi, LWa, BYa, MYa, JYo: performed research and analyzed data through expert assessment of participant outputs.

PCCh, DSh, JTYWa: performed research through event organization.

Tab, BAi, FAK, SAK, MALi, SAL, JAL, MAlt, DAn, ZALa, RAs, ABac, FBBi, ABan, TBa, SMBa, MBe, LPBe, KBi, MBla, JBl, MBlu, JBr, RBr, EBr, MBr, SBr, MBu, NCal, ACa, GGCa, NCar, Ech, Sch, YYTCh, ZCh, Ach, NCi, FCl, JCo, NCo, DMCo, MCGa, VVCr, CCz, TDa, RdICr, AdVa, KBDe, BAdE, LDe, NDh, HDi, MDi, VDi, TDo, ADs, EDu, TDu, MEI, NCEw, KFa, MFa, KFe, GFe, MFi, JFi, EFO, RFO, JFr, YHJFu, SGa, ALGa, DGe, LDGe, CGo, JGo, SGr, HGr, ASGu, KGu, FHa, MHa, JHa, OHa, CHa, AHa, CSHa, MHe, SCFHe, DEHe, AHl, SHHe, GHo, WHO, AHu, HHua, YHu, ZHu, JHu, HHul, ZIl, CGIs, NIt, JJa, EJa, Bji, HJo, SJo, VKa, Ekan, KKa, EKas, EKe, DKi, RKi, VKl, JKo, SKo, SKu, NKu, PKu, KKy, SLa, PHLa, RMCLa, ALe, FLe, CLi, SLi, ZLi, JTLi, PLi, SLiu, YLi, AELMi, ELo, KLo, CLu, PLu, MLu, LMa, MMA, CMa, JMa, TMa, AMc, TMc, FMc, RMc, BMe, SMen, SMer, DMe, MGMe, SMi, MMo, LMu, KKMmu, NMu, SNa, ANe, KNg, LHNg, SKNg, TDNg, TNg, VNg, BONo, MINz, DOHa, VOk, TOI, AOOn, AOOr, OOz, MPa, CAPa, APa, MNPe, LPh, APCo, UPh, EPI, FPr, DPr, APr, MPr, PPU, QQi, AQU, MQne, HRa, LRam, LRaz, JRTa, VRei, VRes, NRi, RRo, NRu, OSak, OSal, CSa, USA, GSa, MSc, YSc, RSe, RSeV, TSh, PSi, NSi, LSi, NSp, MSt, GAST, SSt, XSu, SDSw, NSy, KTa, YSTa, ETa, GTa, WTa, MTa, CTe, DTh, TTi, ETa, ETr, HPLTr, ATs, ATu, STy, JUD, OUI, TMTVo, SAVo, MWa, RWa, AWa, JLWa, ZWa, JWa, MWebb, CSWe, MWebe, NWe, CWe, PWi, MWri, VYWu, YXi, KXi, CXi, HXu, PYa, XYa, WYi, JYYo, LYu, TZa, AZa, FZe, HZh, RZh, ZZh, AJZh, JZo: performed research through the experimental assignments.