Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science

Abel Brodeur[1,2*], David Valenta[1], Alexandru Marcoci[3,4], Juan P. Aparicio[1,2]

Derek Mikola[1,2], Bruno Barbarioli[1,2], Rohan Alexander[5], Lachlan Deer[6]

Tom Stafford[7], Lars Vilhuber[8], Gunther Bensch[9], et al.[10]

[1]Department of Economics, University of Ottawa, Ottawa & K1N 9A7, Canada.

[2]Institute for Replication, University of Ottawa, Ottawa & K1N 9A7, Canada.

[3]School of Politics and International Relations, University of Nottingham, Nottingham NG7 2RD, UK.

[4]Centre for the Study of Existential Risk, University of Cambridge, Cambridge CB2 1SB, UK.

[5]Faculty of Information and Department of Statistical Sciences, University of Toronto, Toronto, M5S 3G6, Canada.

[6]Department of Marketing, Tilburg University, Tilburg, 5037AB, The Netherlands.

[7]School of Psychology, University of Sheffield, Sheffield, UK.

[8]Department of Economics, ILR School, Cornell University, Ithaca, NY, 14853, United States.

[9]RWI – Leibniz Institute for Economic Research, Essen, 45128, Germany.

[10]See next page for full author list. [*]Corresponding author. Email: abrodeur@uottawa.ca

Large Language Models (LLMs) such as ChatGPT are transforming how scientists conduct and validate research. LLMs are thus seen as promising tools to improve scientific reproducibility. We experimentally test how collaboration between researchers and LLM assistants influences the reproduction of quantitative social science findings. We assigned 288 researchers to 103 teams working

in three groups: human-only, AI-assisted, and AI-led. In the AI-led group, the LLM conducted reproducibility checks with minimal human oversight. Human-only and AI-assisted teams reproduced published results at comparable rates, and both outperformed AI-led teams. Human-only teams also identified more major errors than AI-assisted and AI-led teams. Finally, both human-only and AI-assisted teams outperformed AI-led approaches in both proposing and implementing robustness checks. In an exploratory analysis, we observe that the gap in most outcomes between AI-led and the other two groups began to narrow by the final event of 2024. Despite rapid model advances, expert human judgment currently remains indispensable for reliable empirical verification.

Abel Brodeur (University of Ottawa; Institute for Replication), David Valenta (University of Ottawa), Alexandru Marcoci (University of Nottingham, University of Cambridge), Juan P. Aparicio (University of Ottawa; Institute for Replication), Derek Mikola (University of Ottawa; Institute for Replication), Bruno Barbarioli (University of Ottawa; Institute for Replication), Rohan Alexander (University of Toronto), Lachlan Deer (Tilburg University), Tom Stafford (Sheffield University), Lars Vilhuber (Cornell University), Gunther Bensch (RWI - Leibniz Institute for Economic Research), Fabio Motoki (University of Texas Rio Grande Valley; University of East Anglia), Mohamed Abdelhady (Carleton University), Yousra Abdelmoula (Statistics Canada), Ghina Abdul Baki (University of Ottawa), Tomás Aguirre (Centre for the Governance of AI), Sriraj Aiyer (University of Oxford), Shumi Akhtar (The University of Sydney), Farida Akhtar (Macquarie University), Melle R. Albada (Vienna University of Economics and Business), Micah Altman (MIT), David Angenendt (Technical University of Munich), Zahra Arjmandi Lari (Independent researcher), Jorge Armando De León Tejada (Universidad del Rosario), David Rodriguez Arana (Universidad del Rosario), Igor Asanov (International Center for Higher Education Research and Faculty of Economics, University of Kassel), Anastasiya-Mariya Asanov Noha (University of Kassel, INCHER), Rebecca Ashong (University of Ghana), Tobias Auer (London School of Economics), Francisco J. Bahamonde-Birke (Tilburg University), Bradley J. Baker (Temple University), Söhnke M. Bartram (University of Warwick and CEPR), Dongqi Bao (University of Zurich), Lucija Batinovic (Linköping University), Tommaso Batistoni (University of Oxford), Monica Beeder (University of Southampton), Louis-Philippe Beland (Carleton University), Carsten Gero Bienz (Norwegian School of Economics), Christ Billy Aryanto (Faculty of Psychology, Atma Jaya Catholic University of Indonesia), Cylcia Bolibaugh (University of York), Carl Bonander (University of Gothenburg), Ramiro Bravo (The University of Manchester), Egor Bronnikov (Maastricht University; George Mason University), Stephan Bruns (Hasselt University), Nino Buliskeria (Nazarbayev University), Sara Caicedo-Silva (Universidad de los Andes), Andrea Calef (University College London, School of Management), Gustavo A. Castillo Alvarez (Universidad de Los Andes), Solomon Caulker (United Methodist University Sierra Leone), Simonas Cepenas (ISM University of Management and Economics), Arthur Chatton (Université Laval), Zirou Chen (University of Toronto), Ngozi Chioma Ewurum (Michael Okpara University of Agriculture, Umudike, Nigeria), Anda-Bianca Ciocîrlan (University of Sheffield), Felix J. Clouth (Tilburg University), Jason Collins (University of Tech-

nology Sydney), Nikolai Cook (Wilfrid Laurier University), Cesar Cornejo (The London School of Hygiene & Tropical Medicine), João Craveiro (University of Sheffield), Jonathan Créchet (University of Ottawa), Jing Cui (University of Ottawa), Niveditha Chalil Vayalabron (School of Earth and Planetary Science, National Institute of Science Education and Research, India), Christian Czymara (Netherlands Interdisciplinary Demographic Institute), Carlos Daniel Bermúdez Jaramillo (Universidad del Rosario), Hannes Datta (Tilburg University), Lien Denoo (Tilburg University), Arshia Dhaliwal (Carleton University), Nency Dhameja (Binghamton University), Elodie Djemai (Université Paris-Dauphine), Erwan Dujeancourt (Stockholm University), Uğurcan Dündar (Vienna University of Economics and Business), Thibaut Duprey (Bank of Canada), Yasmine Eissa (The American University in Cairo), Youssef El Fassi (HEC Lausanne), Ismail El Fassi (University of St. Gallen), Keaton Ellis (UC Berkeley), Ali Elminejad (Nazarbayev University), Mahmoud Elsherif (University of Birmingham and Leicester), Aysil Emirmahmutoglu (NHH Norwegian School of Economics), Giulian Etingin-Frati (University of Zurich), Emeka Eze (Michael Okpara University of Agriculture), Jan Fabian Dollbaum (University College Dublin), Jan Feld (Victoria University of Wellington), Andres Felipe Rengifo Jaramillo (Business School; Universidad de los Andes), Guidon Fenig (University of Ottawa), Victoria Fernandes (Bank of Canada), Lenka Fiala (University of Ottawa; Institute for Replication; Tilburg University), Lukas Fink (FU Berlin), Mojtaba Firouzjaeiangalougah (Masaryk University), Sara Fish (Harvard University), Jack Fitzgerald (Vrije Universiteit Amsterdam), Rachel Forshaw (Heriot-Watt University), Alexandre Fortier-Chouinard (Université Laval), Louis Fréget (CEPREMAP), Joris Frese (European University Institute), Jacopo Gabani (World Bank; Centre for Health Economics, University of York), Sebastian Gallegos (UAI Business School), Max C. Gamill (University of Sheffield), Attila Gáspár (HUN-REN Centre for Economic and Regional Studies), Romain Gauriot (Deakin University), Evelina Gavrilova (NHH Norwegian School of Economics), Diogo Geraldes (University College Dublin), Giulio Giacomo Cantone (University of Sussex), Grant Gibson (McMaster University & CRDCN), Dirk Goldschmitt (University of Sheffield), Amélie Gourdon-Kanhukamwe (King's College London), Andrea Gregor de Varda (University of Milano-Bicocca), Idaliya Grigoryeva (UC San Diego), Alexi Gugushvili (University of Oslo), Aaron H.A. Fletcher (University of Sheffield), Florian Habermann (University of Lausanne), Márton Hablicsek (Leiden University), Joanne Haddad (Université Libre de Bruxelles), Jonathan D. Hall (University of Alabama), Olle Hammar (Linnaeus Univer-

sity and Institute for Futures Studies), Malek Hassouneh (University of Toronto), Carina I. Hausladen (ETH Zürich), Sophie C. F. Hendrikse (Tilburg University), Matthew Hepplewhite (University of Oxford), Anson T. Y. Ho (Toronto Metropolitan University), Senan Hogan-Hennessy (Cornell University), Elliot Howley (University of Nottingham), Gaoyang Huang (Swiss Federal Institute of Technology in Zurich), Héloïse Hulstaert (Hasselt University; University of Liège), Zlatomira G. Ilchovska (University of York; University of Birmingham), Niklas Jakobsson (Karlstad University), Joakim Jansson (Linnaeus University; Research Institute of Industrial Economics), Ewa Jarosz (University of Warsaw), Hossein Jebeli (Bank of Canada), Yanchen Jiang (Harvard University), Hiba Junaid (Bart's Life Sciences, Bart's Health NHS Trust; Queen Mary University of London), Rohan Kalluraya (Cornell University), Sunny Karim (Carleton University), Edmund Kelly (University of Oxford), Eva Kimel (University of York), Sorravich Kingsuwankul (Vrije Universiteit Amsterdam), Valentin Klotzbücher (University of Freiburg), Daniel Krähmer (University of Munich), Pijus Krūminas (ISM University of Management and Economics), Nicholas Kruus (Schelling Research), Essi Kujansuu (University of Innsbruck; University of Turku), Christoph F. Kurz (Ludwig-Maximilians-Universität Munich) Stephan Küster (Freie Universität Berlin), Blake Lee-Whiting (University of Toronto), Felix Lewandowski (University of Nottingham), Tongzhe Li (University of Guelph), Ruoxi Li (Yale University), Dan Liu (Australian National University), Jiacheng Liu (Purdue University), Helix Lo (University of Tokyo), Katharina Loter (Tilburg University), Felipe Macedo Dias (Cornell University), Christopher R. Madan (University of Nottingham), Nicolas Mäder (University of San Diego), Marco Mandas (University of Cagliari), Cesar Mantilla (Pontificia Universidad Javeriana), Jan Marcus (FU Berlin), Diego Marino Fages (Durham University), Xavier Martin (Tilburg University), Ryan McWay (University of Minnesota), Daniel Medina-Gaspar (Universidad EAFIT), Sisi Meng (Cornell University), Lingyu Meng (University of Sheffield), Simon Merz (University of Halle), Alex P. Miller (University of Southern California), Thibault Mirabel (Equalis Capital), Dibya Deepta Mishra (Rice University), Sumit Mishra (Krea University), Belay W. Moges (Dilla University), Morteza Mohandes Mojarrad (Tilburg university), Myra Mohnen (University of Ottawa), Louis-Philippe Morin (University of Ottawa), Lucija Muehlenbachs (University of Calgary), Gastón Mullin (Tilburg University), Andreea Musulan (University of Montreal), Sara Muzzì (University of Milano Bicocca), James A. C. Myers (University of Sheffield), Florian Neubauer (RWI - Leibniz Institute for Economic Research), Tuan

Nguyen (Hasselt University), Ali Niazi (University of Calgary), Ardyn Nordstrom (Carleton University), Bartłomiej Nowak (Cardinal Stefan Wyszynski University), Daneal O'Habib (Bank of Canada), Tim Ölkers (University of Göttingen), Justin Ong (University of Sheffield), Valeria Orozco Castiblanco (IESE, Universidad de Navarra), Ömer Özak (SMU), Ali I. Ozkes (SKEMA Business School, GREDEG, Université Côte d'Azur), Mikael Paaso (Erasmus University Rotterdam), Shubham Pandey (Universität Osnabrück), Varvara Papazoglou (University of Sheffield), Romeo Penheiro (University of Houston), Linh Pham (Lake Forest College), Ulrike Phieler (Vienna University of Economics and Business), Peter Pütz (Bielefeld University), Quan Qi (University at Albany, SUNY), Jingyi Qiu (University of Michigan), David A. Reinstein (The Unjournal), Juuso Repo (INVEST Flagship Research Center, University of Turku), Nicolas Rudolf (University of Lausanne), Shree Saha (Cornell University), Orkun Saka (City St George's, University of London), Chiara Saponaro (University of Milano-Bicocca), Georg Sator (University of Nottingham), Martijn Schoenmakers (Tilburg University), Raffaello Seri (InsIDE Lab, DiEco, Università degli Studi dell'Insubria), Meet Shah (Toronto Metropolitan University), Paul Sibille (University of Liege), Christoph Siemroth (University of Essex), Vladimir Skavysh (Bank of Canada), Ben Slater (University of Cambridge), Wenting Song (University of California, Davis), Stefan Staubli (University of Calgary), Tobias Steindl (University of Regensburg), Nomwendé Steven Waongo (University of Ottawa), Paul Stott (University of Manchester), Stephenson Strobel (McMaster University), Roshini Sudhaharan (Tilburg University), Pu Sun (Dongbei University of Finance and Economics), Scott D. Swain (Clemson University), Oleksandr Talavera (University of Birmingham), Hanz M. Tantiangco (University of Sheffield), Georgy Tarasenko (Cornell University), Boyd Tarlinton (Department of Primary Industries, Queensland), Mariam Tarraf (Carleton University), Ken Teoh (International Monetary Fund), Rémi Thériault (Université du Québec à Montréal), Bethan Thompson (SRUC), Tonghui Tian (Carleton University), Wenjie Tian (University of Ottawa), Manuel Tobias Rein (Tilburg University), Emmanuel Tolani (University of Bonn), Nicolai Topstad Borgen (University of Oslo), Solveig Topstad Borgen (University of Oslo), Javier Torralba (Tilburg University), Carolina Velez-Ospina (World Bank), Man Wai Mak (Carleton University), Lukas Wallrich (Birkbeck, University of London), Zeyang Wang (Vanderbilt University), Leah Ward (University of Manchester), Matthew D. Webb (Carleton University), Duncan Webb (Princeton University), Bryan S. Weber (College of Staten Island, CUNY), Christoph Weber (ESSCA School of Management),

Wei-Chien Weng (National Taiwan University), Christian Westheide (University of Vienna), Tom Wilkinson (University of Sheffield), Kwong-Yu Wong (National University of Singapore), Marcin Wroński (Collegium of World Economy, SGH Warsaw School of Economics), Zhuangchen Wu (University of Birmingham), Qixia Wu (University of Ottawa), Victor Y. Wu (Stanford University), Bohan Xiao (University of Ottawa), Feihong Xu (Northwestern University), Cong Xu (National Chengchi University; Aalto University), Pranav Yadav (Tilburg University), Yu Yang Chou (University College London), Luther Yap (Princeton University), Myra Yazbeck (University of Ottawa), Bo Yao (Lancaster University), Zuzanna Zagrodzka (University of Sheffield), Tahreen Zahra (Carleton University), Mirela Zaneva (University of Oxford), Xiaomeng Zhang (Nanjing Audit University), Ziwei Zhao (University of Lausanne; Swiss Finance Institute), Han Zhong (University of Toronto), Aras Zirgulis (ISM University of Management and Economics), Jiacheng Zou (Columbia University), Floris Zoutman (NHH Norwegian School of Economics), Christelle Zozoungbo (Penn State University).

1 Introduction

Reproducibility is a cornerstone of robust quantitative empirical research, where complex methodologies and data handling techniques are common (*1–8*). Despite advancements in reproducibility protocols (*9*), concerns persist regarding the accuracy and reliability of published findings (*10–17*). Unclear reporting, peer review challenges, and methodological advances requiring expertise when evaluating quantitative studies, all contribute to the current reproducibility and replication crises in the behavioral and social sciences. This study investigates how artificial intelligence (AI) tools, such as Large Language Models (LLMs), could support researchers, data editors, and scientific journals in computationally reproducing research. We focus on three modes of AI and human interaction: human-only teams, human teams with AI assistance (the "AI-assisted" approach), and teams that provided only limited oversight while AI carried out reproducibility checks (the "AI-led" approach). We use ChatGPT because it processes different file formats effectively for reproduction and is used most frequently by researchers. (*18*).

This paper tests how effectively AI reproduces scientific articles and works in complex cases where coding errors or methodological inconsistencies arise. We employ a randomized controlled trial design involving three treatment arms. We contribute to a large literature documenting the benefits and limitations of human-AI integration, as well as full automation (*19–33*). This is crucial for science because current methods for performing computational reproducibility and robustness checks are expensive, time consuming, and require advanced technical skills (*34, 35*). We also contribute to a growing body of literature documenting the potential pitfalls of integrating human and artificial intelligence, such as over-reliance and expertise erosion (*36, 37*). This research also provides some comparative productivity measures in highly specialized intellectual tasks. Our research is quite different from current studies, which focus on customer support agents and low-skill occupations (*27, 38*).

We focus on three primary outcomes across the treatment groups: (1) computational reproducibility success rates, (2) error detection capabilities, and (3) robustness check quality. Understanding these outcomes contributes to a broader understanding of AI, and offers insights into the optimal balance of human and AI involvement in research reproduction tasks.

2 Procedures

The first ten coauthors organized seven AI replication games between February and November 2024. All remaining coauthors and a few of the organizers participated in one of those games. The participants were a mix of graduate students, postdoctoral fellows, professors, and researchers from non-academic organizations with a doctoral degree. Table S4 provides details on team composition. Randomization was carried out in two steps for each of the seven events. In step one, coauthors were randomly assigned to a team of three to evaluate the reproducibility of a quantitative social science article. The randomization in step one was conditional on the software preferences reported by participants (Stata or R) *and* the mode of participation (in person or virtual). In step two, each team was randomly assigned to one of three treatment arms: human-only, AI-assisted, or AI-led.

Each team was assigned a study from leading social science journals (i.e., economics, political science, or behavioral science/psychology). Each event included two studies with known coding errors (one in Stata and one in R) that had been identified by the lead authors in a prior study but were not publicly disclosed at the time of the AI replication game. Teams and local organizers had no information about the study they would be reproducing until the day of the event. In total, 12 studies were used.

The resources for the event were given to the teams at 09:00 local time on the day of the event. We shared with them: the journal article and online appendix as PDFs, the original authors' replication package, and screenshots of the exhibit to reproduce from the article (see Supplementary Materials). The screenshots were implemented after the pilot event to assist AI-led teams, as they were not allowed to read the PDF files. Teams had seven hours to complete three tasks: (i) computationally reproduce a few pre-determined results, (ii) detect coding errors, and (iii) suggest and implement up to two robustness checks. The three tasks were independent from each other (e.g., teams did not need to fix coding errors to computationally reproduce results). Teams could leave before the end of the event if they believed they had completed their tasks. Upon completion, teams were asked to email the lead authors a (templated) time log documenting whether they completed computational reproducibility, with a list of all coding errors uncovered, and two robustness checks. All AI-assisted and AI-led teams used ChatGPT during the event, and had to provide their AI conversation history (i.e., a transcript of all prompts and responses exchanged with ChatGPT).

Access to a paid subscription of ChatGPT (powered initially by GPT-4 and subsequently by other models) was provided to all members in the AI-assisted and AI-led teams. AI-assisted and AI-led teams took part in a mandatory one-hour training on the usage of ChatGPT. Participants viewed the training live or later *via* recording. The AI training was optional for human-only teams. Additional details on our AI training and models available and their capabilities can be found in the Supplementary Materials Section 4.4.

Human-only teams were not allowed to use ChatGPT or any other AI tool. The AI-assisted groups were allowed to use ChatGPT without limitation (but no other AI tool). AI-led teams had to perform their three tasks only through ChatGPT. For example, AI-led teams were not allowed to read the article or look at the data and code. They had to feed the article to ChatGPT along with an image of the table(s)/figure(s) to be reproduced. They were asked to first use ChatGPT's Python interpreter module to conduct the analysis. However, they were allowed to run analysis code locally (in R or Stata) when ChatGPT failed to run the analysis itself. When running code locally, the teams were not allowed to use any other code except code provided by ChatGPT, though the teams could adjust file paths and their environment without the assistance of ChatGPT. During the pre-games AI training, participants were shown examples of how to upload the article and replication files to ChatGPT and how to use the Python interpreter module. We relied on the integrity of the AI-led teams to *not* look at the studies, codes, or files. That is, we asked them to pass everything through ChatGPT; we did not give specific guidance on how teams should operate. Teammates could work independently or jointly throughout the event.

In summary, we have 103 teams in Study I: 33 human-only teams (92 researchers), 35 AI-assisted teams (93 researchers), and 35 AI-led teams (103 researchers). Table S4 shows the treatment arms are balanced across observables.

2.1 Three Tasks

Participants had three objective tasks with measurable outcomes. First, teams were asked to computationally reproduce a few selected results in the study assigned to them. The numerical results were selected by AB based on their relative importance to the main claims of the article. Computational reproducibility involves using the same data as the original authors and running their code.

In the templated log, teams recorded the time taken to computationally reproduce the numerical result. Notably, AB, JA, and DM were able to computationally reproduce all results before the event, requiring only minimal adjustments (e.g., updating file paths). We have two different dependent variables for computational reproducibility: one outcome as a binary (completed computational reproducibility versus did not complete), and one that is time (in minutes) from the start of the event to when teams completed a computational reproduction. A computational reproduction is defined as the successful execution of the original authors' codes and the production of numerical results in line with those in the article.

Second, we compare how effective different team types were in finding coding errors or data irregularities. For simplicity, we refer to these as "errors." We categorize errors as major or minor based on whether they could, in theory, have an impact on the claims tested. For instance, a coding error or data irregularity that impacts the dependent or independent variables is considered a major error, as it could have an impact on the estimation results. In contrast, missing packages/paths or versioning issues are considered a minor error. These coding errors are typically easily fixed by the reproducers and do not impact the validity of the claims made by the original authors.

Third, we asked each team to propose and perform two robustness checks. A robustness check is defined in our study as an additional statistical computation. We instructed that these robustness checks should not repeat ones already mentioned in the study or its supplementary materials, that they should be feasible, and that heterogeneity analysis (e.g., comparing female and male respondents) was not considered a robustness check.

Defining what makes a robustness check "good" or "bad" is not straightforward. We define four binary criteria for evaluating the quality of robustness checks: (i) clarity of purpose and execution; (ii) feasibility; (iii) novelty (i.e., not previously done by the original authors); and (iv) relevance to the validity of the empirical strategy. Items (i) through (iii) are basic necessary conditions. Item (iv) requires that the purpose of the robustness test is to provide evidence regarding the credibility of the empirical strategy (*39–41*). All four criteria must be met for a robustness check to be considered "good." Exceptionally, running corrected code in an attempt to correct major errors in the original paper, is coded as a "good" robustness check, regardless of whether it complies with the previous criteria.

We measure differences by team type in proposing and implementing robustness tests using

four measures. The first two are whether teams *proposed* one or two "good" robustness checks. The third and fourth dependent variables are whether the reproducers *implemented* one or two "good" robustness checks, respectively.

3 Results

Our analyses were preregistered after the pilot event in Toronto. We list deviations from our preregistration in the Supplementary Materials Section and note throughout whether the analysis is exploratory.

3.1 Computational Reproducibility

Our main finding is that computational reproducibility rates varied substantially across the groups. Most human-only (94%; 31/33) and AI-assisted (91% /32/35) teams could computationally reproduce the results, while only 37% (13/35) of AI-led teams were able to do so (see Table **??**). Table 2 shows the ordinary least squares (OLS) estimates of our main regression model (see Table S5 for logit and poisson regressions and Table **??** for coefficient estimates concerning the control variables). We find that human-only teams are about 59 percentage points more likely than AI-led teams to successfully computationally reproduce the results ($p < 0.000$). In contrast, there is no statistically significant difference between human-only and AI-assisted teams ($p = 0.771$).

We next investigate how the distribution of time-to-computational reproduction varies across groups. Figure S8 illustrates Kaplan-Meier curves showing, by treatment arm, how long teams took to reproduce their paper. The proportion of teams that reproduce their paper does not reach 100% after seven hours in any treatment arm because all treatment arms contain some teams who could not reproduce their paper. This is especially noticeable for AI-led teams. We find that human-only and AI-assisted teams are significantly faster than AI-led teams. There is no statistically significant difference between human and AI-assisted teams.

In an exploratory analysis, we investigate whether AI-assisted and AI-led teams improved over time. In our setting, improvements could be due to new ChatGPT versions and increased researchers' skills over time. In Figure S2, we show the difference in computational reproducibility rates for each event. Visually, AI-led teams did not improve over time when compared to human-only teams dur-

ing the first couple of events in 2024. We observe that the reproducibility rate gap between human-only and AI-led teams was over 50 percentage points for most events in 2024. In contrast, this gap had narrowed by the final event of 2024.

3.2 Coding Errors or Data Irregularities

We have two dependent variables concerning coding error detection: counts of major and minor errors detected. We find that human-only teams identified on average 1.42 minor and 1.36 major errors. In contrast, AI-assisted and AI-led teams uncovered on average 0.94 and 0.51 minor errors and 0.63 and 0.23 major errors, respectively (Table S3). Table 2 provides OLS estimates indicating that compared to both AI-assisted and AI-led teams, human-only teams uncovered more major errors ($p = 0.013$ and $p < 0.000$, respectively) and minor errors ($p = 0.075$ and $p < 0.000$, respectively). We also find that AI-assisted teams uncovered significantly more minor ($p = 0.022$) and major ($p = 0.017$) errors than AI-led teams. The Supplementary Materials Section provides examples of errors and a discussion. In an exploratory analysis, we show no differences for teams working with Stata in comparison to R (Table S8).

Figure S9 illustrates Kaplan-Meier curves showing how long teams took to find a first minor error (top panel) and a first major error (bottom panel). We find that the speed at which AI-assisted teams uncover a first error is not statistically significantly different from that of human-only teams, and that AI-led teams are statistically significantly slower than human-only teams.

Our findings suggest that human-only teams were more effective at detecting both major and minor errors compared to AI-led teams, highlighting a challenge in AI-led teams' ability to autonomously navigate and interpret complex code and detect data irregularities. We also find that human-only teams performed significantly better than AI-assisted teams on error detection, particularly in identifying errors with potentially significant implications (i.e., major errors). The difference between human-only and AI-assisted teams could have several explanations, including over-reliance on AI (*32*). We explore potential mechanisms in our qualitative analysis (Section 3.5). We also provide non-causal evidence in an exploratory analysis in Table S9 that AI-assisted teams with more AI experience uncovered more errors, suggesting that AI training and practice may be critical for realizing the full benefits of AI assistance.

In an exploratory analysis, we investigate if the performance of AI-led teams in detecting errors improved over time. Figures S4 and S6 provide evidence of an improvement relative to human-only groups, especially for major errors. This could be attributed to factors such as access to new ChatGPT versions or increased researchers' skills over time.

3.3 Proposed Robustness Checks

We find a clear, consistent performance hierarchy across both experiments: human-only and AI-assisted teams outshine AI-led teams. We find that every human-only (33/33) and AI-assisted (35/35) teams proposed at least one good robustness check, whereas only 29 of 35 AI-led teams did so. Table 2 provides OLS estimates and show that the difference between AI-led groups and the other two groups is statistically significant ($p = 0.017$ and $p = 0.023$, respectively). We find that 29 of 33 human-only and 30 of 35 AI-assisted teams suggested two good checks, compared with just 22 of 35 AI-led teams (Table 2, $p0.022$ and $p = 0.035$).

When we look at actually implementing those checks, AI-led teams were almost 32 percentage points less likely than the other two groups to carry out a "good" robustness check ($p = 0.002$ and $p = 0.003$), and six AI-led teams supplied no robustness checks at all. These six teams' checks were judged as "bad" mostly because of a lack of clarity and duplicating analyses already run by the original authors.

Our results indicate that AI-led teams, while able to produce robustness checks with some level of quality, faced more challenges in aligning with the criteria. These difficulties may stem from limited human guidance in interpreting the empirical strategy and in assessing the feasibility of the checks.

3.4 Additional Analyses for AI-Assisted Teams

Table S10 presents an exploratory correlational analysis examining the relationship between AI usage (measured by total prompts) and performance in AI-assisted teams. See Figure S7 for descriptive statistics on AI usage for AI-assisted teams. For this analysis, we divided teams into lower or higher AI-usage groups using a median split based on the total number of prompts they employed.

The findings indicate that AI-assisted teams with lower AI usage were less likely to achieve

computational reproduction of the original results. However, these same teams identified a greater number of both minor and major errors. They also completed the computational reproduction task more quickly and required less time to detect their first minor and major errors. Of note, our sample is small and none of the differences are statistically significant at conventional levels. These results relate to a literature studying over-reliance on AI support (*36, 37, 42–44*).

3.5 Focus Groups

Between 18 April and 30 April 2025, we conducted six one-hour focus groups ($n = 25$) involving AI-led ($n = 8$), AI-assisted ($n = 11$) and human-only ($n = 6$) participants. NEED TO MENTION CLEAR ISSUE HERE DOCUMENTED BY R2 and R3. Being upfront that a methodlogical issue is that participants know results. Thematic analysis revealed clear patterns: AI-assisted participants reported that AI assistance sped things up, while AI-led participants reported the opposite. Human-only participants believed they were the most effective at detecting major errors, with AI-led participants trailing. These patterns appear to hinge on how teams set a "delegation threshold." AI-assisted teams strategically outsourced micro-tasks (e.g., boilerplate code, file location) while retaining conceptual control, whereas AI-led groups ceded entire analytic stages to ChatGPT and struggled when automation failed.

Data illuminated the practical consequences of these choices. Initial optimism about LLMs quickly gave way to "prompt fatigue," overconfidence, and mounting frustration, especially among AI-led participants who faced hallucinated paths, truncated context windows, and prolonged debugging loops. Human expertise remained necessary for detecting subtle errors and for arbitrating disagreements between AI output and reality. Nevertheless, when used judiciously, LLMs accelerated routine work, suggested robustness checks, and broadened analytical ambition for less-experienced coders. Therefore, our focus-group findings imply that effective LLM prompting is becoming a specialized research skill and that near-term gains will come from augmenting—not replacing—human judgment.

4 Discussion

Computational reproducibility, error detection, and robustness checks are essential components of empirical research validation, but are resource-intensive tasks. Our comparative analysis of human-only, AI-assisted, and AI-led teams sheds light on how AI may be integrated into the expensive reproduction process, accelerating it and improving its overall reliability.

Although recent advancements in LLMs have opened possibilities for AI integration in research (*45, 46*), our results temper expectations for immediate, widespread AI autonomy in reproducibility. We show that while AI-driven reproduction has the potential to save time and money for a subset of studies, a human component remains important in ensuring successful computational reproduction for most studies. Furthermore, given that AI-assisted or AI-led team performance does not exceed human-only teams', but requires additional costs like a paid subscription, AI in computational reproducibility does not promise cost savings yet.

Consequently, for most studies, AI's optimal role in reproducibility may therefore still be as a collaborator rather than an autonomous executor. For instance, LLM systems could conduct a first pass to help identify coding errors (*47*) and propose potential fixes (*48, 49*). Then, on a subsequent step, humans would still play the pivotal role of performing a more in-depth evaluation.

4.1 Summary of Findings

AI-led teams faced notable challenges compared to both AI-assisted and human-only teams. Only 37% of AI-led teams were able to successfully complete computational reproducibility, highlighting a substantial gap in the capacity of AI in 2024 to autonomously guide researchers through complex quantitative analyses. Similarly, in error detection, AI-led teams documented significantly fewer major and minor errors than either AI-assisted or human-only teams. These findings underscore the importance of still integrating human expertise until the gap is closed. As LLMs continue to evolve, sustained benchmarking against humans will be crucial to ensure that future AI-led efforts close and potentially surpass the existing performance gap.

4.2 Limitations

One limitation is our sole focus on OpenAI's ChatGPT, meaning that we cannot generalize to all current AI models. Furthermore, the limited timeframe of seven hours for study teams to complete their reproductions may not adequately reflect the conditions under which reproducibility efforts are conducted depending on the field of science. Similarly, researchers would not normally know that the articles they are reproducing contain errors. Finally, we relied on a small number of research papers illustrating a limited range of social science methodologies and techniques, which makes it difficult to generalize our findings to work with AI systems across different social sciences.

4.3 Implications for Human-AI Collaboration in Research

Our findings support the notion that, while AI tools hold promise for aiding in reproducibility tasks, the state of technology as of late 2024 is not yet advanced enough for full autonomy in complex empirical workflows. Human expertise remains critical to navigate challenges and provide interpretative guidance for reproducibility and error detection. The AI-assisted model—where humans work alongside AI tools—did not emerge as a winner over humans-only teams in overall outcomes, but consistently outperformed AI-led teams.

In scenarios where computational reproducibility, error detection, and robustness checks require in-depth understanding, domain knowledge, and flexible problem-solving, human involvement currently adds value. The ability to contextualize, interpret, and implement complex quantitative research remains a human strength, underscoring the current limitations of AI in fully autonomous reproduction efforts.

4.4 Outlook

Looking ahead, future advancements in models optimized through reinforcement learning to solve reasoning problems using chain of thought (CoT) could address the limitations we reported, possibly improving the model's ability to reproduce complex quantitative research through iterative, reasoning-driven processes. As LLMs evolve to incorporate better contextual understanding and reasoning, their role in reproducibility tasks could shift from support to a more central position, especially in less complex, structured reproduction settings. Future iterations of AI tools may in-

corporate improvements in interpreting code and data irregularities, detecting nuanced errors, and generating plausible robustness checks with minimal human input. Such advancements could enhance AI's ability to autonomously execute reproducibility tasks, reducing the reliance on human oversight for routine or straightforward reproducibility challenges.

Future research should consider the potential for training models specifically in social science and quantitative research contexts. Current LLMs are trained on vast datasets but may lack specificity in understanding the unique demands of empirical social science research. AI systems tailored for social science reproduction could potentially improve reproducibility outcomes, reducing the barriers AI currently faces in autonomously handling the nuances of quantitative research. Additionally, incorporating continuous feedback and learning mechanisms could allow AI-assisted and AI-led teams to improve performance over time, as AI learns from each reproduction task and adapts based on human feedback.

In exploratory follow-up work, we began analyzing ChatGPT transcripts from the AI-led and AI-assisted teams to identify prompting styles and failure patterns (see ChatGPT transcripts in Supplementary Materials). This transcript-based analysis offers promising avenues for future research on LLM–human collaboration and may inform the design of more structured prompt agents and reproducibility support tools. The results obtained through these experiments are also valuable to other researchers and for our future endeavors. For instance, we are currently working on an open-source LLM replicator, which will use the best prompts obtained through the games in a few-shot learning approach, improving future models' performance as it has been shown to be an effective task-specific enhancement technique (*50*). Furthermore, it is possible to fine-tune open-source models with our dataset of curated reproductions in order to better align the responses with the desired outcomes.

Table 1: Comparison of Human, AI-Assisted, and AI-Led Metrics

Variable	Human-Only	AI-Assisted	AI-Led	Human-Only vs AI-Assisted	Human-Only vs AI-Led	AI-Assisted vs AI-Led
Reproduction	0.939 (0.242)	0.914 (0.284)	0.371 (0.490)	0.025 [0.697]	0.568 [<0.001]	0.543 [<0.001]
Minutes to reproduction	82.0 (39.8)	93.3 (85.4)	179.7 (68.4)	-11.3 [0.505]	-97.7 [<0.001]	-86.4 [0.002]
Number of minor errors	1.424 (1.696)	0.943 (1.454)	0.514 (0.919)	0.481 [0.213]	0.910 [0.007]	0.429 [0.145]
Minutes to first minor error	100.7 (77.1)	81.9 (44.6)	161.0 (103.3)	18.7 [0.381]	-60.3 [0.071]	-79.1 [0.010]
Number of major errors	1.364 (1.496)	0.629 (0.942)	0.229 (0.490)	0.735 [0.017]	1.135 [<0.001]	0.400 [0.029]
Minutes to first major error	153.2 (86.1)	138.4 (55.9)	196.0 (97.7)	14.8 [0.577]	-42.8 [0.284]	-57.6 [0.099]
At least one good robustness check	1.000 (0.000)	1.000 (0.000)	0.829 (0.382)	0.000 [-]	0.171 [0.012]	0.171 [0.010]
At least two good robustness checks	0.879 (0.331)	0.857 (0.355)	0.629 (0.490)	0.022 [0.796]	0.250 [0.017]	0.229 [0.029]
Ran at least one good robustness check	0.939 (0.242)	0.943 (0.236)	0.571 (0.502)	-0.003 [0.953]	0.368 [<0.001]	0.371 [<0.001]
Ran at least two good robustness checks	0.788 (0.415)	0.800 (0.406)	0.457 (0.505)	-0.012 [0.903]	0.331 [0.005]	0.343 [0.003]

Note: Columns 2–4 present means and standard errors in parentheses for individual groups (Human-only, AI-Assisted, and AI-Led); columns 5–7 present differences in means and p-values in brackets for group comparisons (Human-Only vs AI-Assisted, Human-Only vs AI-Led, and AI-Assisted vs AI-Led).

Table 2: Causal relationship between treatment groups and reproducibility outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Reproduction	Minor errors	Major errors	One good robustness	Two good robustness	Ran one robustness	Ran two robustness
AI-Assisted	-0.018	-0.487*	-0.646**	-0.009	-0.014	-0.032	-0.009
	(0.063)	(0.270)	(0.254)	(0.027)	(0.103)	(0.061)	(0.113)
	[-0.144; 0.107]	[-1.025; 0.051]	[-1.153; -0.139]	[-0.063; 0.046]	[-0.220; 0.191]	[-0.155; 0.090]	[-0.233; 0.216]
AI-Led	-0.593***	-1.050***	-1.136***	-0.167**	-0.250**	-0.323***	-0.290**
	(0.090)	(0.258)	(0.235)	(0.068)	(0.107)	(0.098)	(0.126)
	[-0.773; -0.413]	[-1.565; -0.536]	[-1.604; -0.667]	[-0.302; -0.031]	[-0.463; -0.037]	[-0.518; -0.127]	[-0.540; -0.040]
Controls	✓	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.738	0.951	0.728	0.942	0.786	0.816	0.680
p-val (AI-Assisted = AI-Led)	0.000	0.019	0.015	0.021	0.032	0.003	0.017
Obs.	103	103	103	103	103	103	103

Note: Standard errors in parentheses; confidence intervals in brackets. Human-only group omitted.

Controls: number of teammates; game-by-software fixed effects; maximum and minimum position skill fixed effects; attendance fixed effects.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

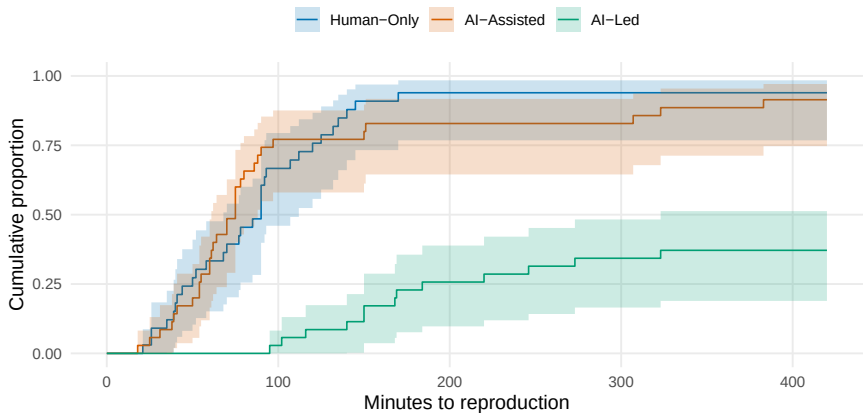

Figure 1: Kaplan-Meier curves, showing the proportion of teams who computationally reproduced the paper by time *t* along with curve bands

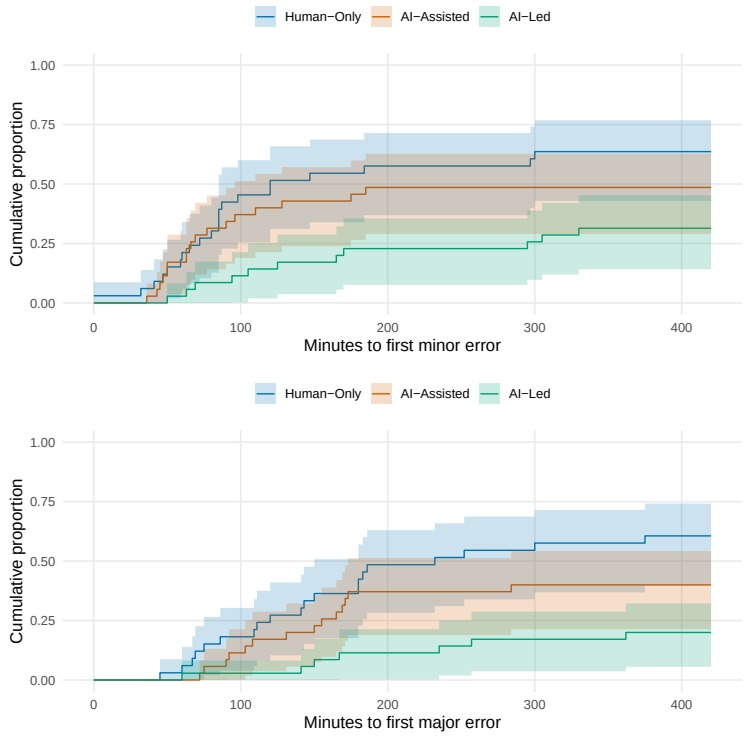

Figure 2: Kaplan-Meier curves, showing the proportion of teams who found their first coding error by time *t* along with curve bands

References and Notes

1. A. Brodeur, *et al.*, Promoting reproducibility and replicability in political science. *Research & Politics* **11** (1), 20531680241233439 (2024).

2. D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, V. Stodden, Reproducible research in computational harmonic analysis. *Computing in Science & Engineering* **11** (1), 8–18 (2008).

3. M. Fišar, *et al.*, Reproducibility in Management Science. *Management Science* **70** (3), 1343–1356 (2024).

4. P. Gertler, S. Galiani, M. Romero, How to Make Replication the Norm. *Nature* **554** (7693), 417–9 (2018).

5. S. N. Goodman, D. Fanelli, J. P. Ioannidis, What does research reproducibility mean? *Science Translational Medicine* **8** (341), 341ps12–341ps12 (2016).

6. M. Milkowski, W. M. Hensel, M. Hohol, Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience* **45** (3), 163–172 (2018).

7. National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science* (National Academies Press) (2019), doi:10.17226/25303, `https://www.nap.edu/catalog/25303`.

8. C. Pérignon, K. Gadouche, C. Hurlin, R. Silberman, E. Debonnel, Certify reproducibility with confidential data. *Science* **365** (6449), 127–128 (2019).

9. L. Vilhuber, Report by the AEA Data Editor. *AEA Papers and Proceedings* **112**, 813–23 (2022), doi:10.1257/pandp.112.813.

10. A. Brodeur, *et al.*, Mass Reproducibility and Replicability: A New Hope (2024), Institute for Replication Discussion Paper 107.

11. A. C. Chang, P. Li, Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not". *Critical Finance Review* **11** (1), 185–206 (2022).

12. S. Crüwell, *et al.*, What's in a badge? A computational reproducibility investigation of the open data badge policy in one issue of Psychological Science. *Psychological Science* **34** (4), 512–522 (2023).

13. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science* **349** (6251), aac4716 (2015).

14. P. Obels, D. Lakens, N. A. Coles, J. Gottfried, S. A. Green, Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science* **3** (2), 229–237 (2020).

15. C. Pérignon, *et al.*, Computational reproducibility in finance: Evidence from 1,000 tests. *The Review of Financial Studies* **37** (11), 3558–3593 (2024).

16. V. Stodden, J. Seiler, Z. Ma, An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences* **115** (11), 2584–2589 (2018).

17. B. Wood, R. Müller, A. Brown, Push Button Replication: Is Impact Evaluation Evidence for International Development Verifiable? (2018), `https://osf.io/n7a4d/`, OSF Preprints.

18. A. Hrycyshyn, H. Eassom, *ExplanAItions: An AI Study*, Tech. rep., John Wiley & Sons, Hoboken, NJ (2025), `https://www.wiley.com/content/dam/wiley-dotcom/en/b2c/content-fragments/explanaitions-ai-report/pdfs/Wiley_ExplanAItions_AI_Study_February_2025.pdf`.

19. G. Bansal, *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* (2021), pp. 1–16.

20. G. Bansal, B. Nushi, E. Kamar, E. Horvitz, D. S. Weld, Is the most accurate ai the best teammate? Optimizing AI for teamwork, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35 (2021), pp. 11405–11414.

21. E. Bondi, *et al.*, Role of human-AI interaction in selective prediction, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36 (2022), pp. 5286–5294.

22. Á. A. Cabrera, A. Perer, J. I. Hong, Improving human-AI collaboration with descriptions of AI behavior. *Proceedings of the ACM on Human-Computer Interaction* **7** (CSCW1), 1–21 (2023).

23. E. Goh, *et al.*, Influence of a Large Language Model on Diagnostic Reasoning: A Randomized Clinical Vignette Study. *medRxiv* (2024).

24. B. Koepnick, *et al.*, De novo protein design by citizen scientists. *Nature* **570** (7761), 390–394 (2019).

25. H. Liu, V. Lai, C. Tan, Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* **5** (CSCW2), 1–45 (2021).

26. H. Mozannar, *et al.*, Effective human-AI teams via learned natural language rules and onboarding. *Advances in Neural Information Processing Systems* **36** (2024).

27. S. Noy, W. Zhang, Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381** (6654), 187–192 (2023).

28. C. Reverberi, *et al.*, Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports* **12** (1), 14952 (2022).

29. M. Schemmer, P. Hemmer, M. Nitsche, N. Kühl, M. Vössing, A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making, in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (2022), pp. 617–626.

30. M. H. Tessler, *et al.*, AI can help humans find common ground in democratic deliberation. *Science* **386** (6719), eadq2852 (2024).

31. M. Vaccaro, J. Waldo, The effects of mixing machine learning and human judgment. *Communications of the ACM* **62** (11), 104–110 (2019).

32. M. Vaccaro, A. Almaatouq, T. Malone, When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* pp. 1–11 (2024).

33. B. Wilder, E. Horvitz, E. Kamar, Learning to complement humans, in *Proceedings of the 29th International Joint Conference on Artificial Intelligence* (2020), pp. 1526–1533.

34. Cherry Bekaert LLP, Report of Independent Auditor (2022), doi:10.1257/aer.112.6.2083, `https://pubs.aeaweb.org/doi/10.1257/aer.112.6.2083`.

35. J.-E. Colliard, C. Hurlin, C. Perignon, The Economics of Computational Reproducibility (2022), SSRN: https://ssrn.com/abstract=3418896.

36. Z. Buçinca, M. B. Malaya, K. 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).

37. L. J. Skitka, K. L. Mosier, M. Burdick, Does automation bias decision-making? *International Journal of Human-Computer Studies* **51** (5), 991–1006 (1999).

38. E. Brynjolfsson, D. Li, L. Raymond, Generative AI at work. *Quarterly Journal of Economics* **140**, 889–942 (2025).

39. M. Del Giudice, S. W. Gangestad, A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science* **4** (1), 2515245920954925 (2021).

40. X. Lu, H. White, Robustness checks and robustness tests in applied economics. *Journal of Econometrics* **178**, 194–206 (2014).

41. M. B. Nuijten, Assessing and improving robustness of psychological research findings in four steps, in *Avoiding questionable research practices in applied psychology* (Springer), pp. 379–400 (2022).

42. V. Lai, H. Liu, C. Tan, "Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–13.

43. Y. Zhang, Q. V. Liao, R. K. Bellamy, Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), pp. 295–305.

44. H. Vasconcelos, *et al.*, Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* **7** (CSCW1), 1–38 (2023).

45. Y. K. Dwivedi, *et al.*, Opinion Paper:"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* **71**, 102642 (2023).

46. B. D. Lund, *et al.*, ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology* **74** (5), 570–581 (2023).

47. N. Wadhwa, *et al.*, Core: Resolving code quality issues using llms. *Proceedings of the ACM on Software Engineering* **1** (FSE), 789–811 (2024).

48. Y. Zhang, Detecting code comment inconsistencies using llm and program analysis, in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering* (2024), pp. 683–685.

49. D. Nam, A. Macvean, V. Hellendoorn, B. Vasilescu, B. Myers, Using an llm to help with code understanding, in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (2024), pp. 1–13.

50. T. B. Brown, *et al.*, Language Models are Few-Shot Learners (2020), `https://arxiv.org/abs/2005.14165`.

51. Open AI, File uploads FAQ (2024), `https://help.openai.com/en/articles/8555545-file-uploads-faq/` [Accessed: November 28, 2024].

52. Open AI, Learning to Reason with LLMs (2024), `https://openai.com/index/learning-to-reason-with-llms/` [Accessed: November 18, 2024].

53. V. Braun, V. Clarke, Using thematic analysis in psychology. *Qualitative research in psychology* **3** (2), 77–101 (2006).

Acknowledgments

We would like to thank Gabriel Zimmerman for research assistance.

Funding: This research and AI replications games were funded by Open Philanthropy project "Benchmarking LLM agents on real-world tasks: Reproducibility" and the Alfred P. Sloan Foundation Foundation grant G-2023-22326. We also benefited from funding to host games from the Universities of Toronto, Ottawa, Cornell and Tilburg. Mahmoud Elsherif acknowledges funding from Leverhulme Early Career Research Fellowship-ECF-2022-761. Shumi Akhtar acknowledges funding DP200102935 awarded by the Australian Research Council Grant.

Author contributions: ABr, DVa, AMa, JPAp, DMi, BBa, RAl, GSa: conception of Study I. ABr, DVa, AMa, JPAp, DMi, BB, RAl, FAk, NTBo, CCo, LFi, JFi, JFr, DRe, GGi, STo, LPBe, MME, AMu, NMa, SMe, PSu, RSe, VSk, GTa, LYa, BWe: conception of the revision and Study II. JPAp, ABr, GBe, CGBi, ZAr, IAs, TAu, CBo, RBr, SBr, CBo, ACh, ADh, EDu, YEi, JFi, OHa, AHo, GHu, HHu, EKe, VKl, NKr, JLi, RLi, KLo, AMANo, SMe, SMi, SMu, FNe, TNg, UPh, MRe, PSi, SSt, BTa, MTa, OTa, DVa, CWe, VYWu, ZWu: analysis and interpretation of data. ABr, DVa, AMa, JPAp, DMi, BBa wrote the original draft, lead revision, and take responsibility for the content – while SAi, IAs, MAlt, FAk, SAk, BJBa, LBa, GBe, MBe, CGBi, CBo, YBo, RBr, SBr, ACh, JCo NCo, FCl, LDe, EDj, TDu, AEl, IElFa, GFe, JFe, LFi, LFr, JFi, JFDo, AGá, JGa, SGa, GGi, AGKa, DGo, IGr, EGZo, FHa, JHa, MHa, SHe, AHo, GHu, ZIl, JJa, EKe, EKi, SKi, NKr, EKu, SKu, BLWh, DLi, JLi, KLo, CMa, RMcW, XMa, SMe, BMo, FMo, MMo, LPMo, LMu, AMANo, FNe, AOz, OOz, SPa, UPh, PPu, QQi, NRu, OSa, DRe, JRe, MTRe, RSe, PSi, SSt, TSt, PSu, GTa, RTh, TTi, LWa, CWe, MWe, MWr, WCWe, VYWu, BYa, LYa, MZa, HZh, XZh, AZi, ZZa, FZo contributed to editing and review through commenting on drafts of the paper. DVa: AI training. AGKa, JPAp, RAl, FBBi, RFo, ANo: focus group moderators and conception. CBAr, DAn, FCl, NCo, NDh, AFCh, GEFr, JFr, DGe, YJi, PKr, SKu, JMa, NMa, MME, CSa, OSa, WSo, TSt,

GTa, AdVa, LYa: focus group participation. DVa: Event organization: ABr, RAl, LDe, TSa, LVi. All coauthors except ABr, DVa, AMa, DMi, BBa, TSa: data acquisition through AI replication games.

Competing interests: The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada. AM is a UKRI Policy Fellow seconded to the Department for Science, Innovation and Technology. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department for Science, Innovation and Technology or the UK Government.

Data and materials availability: We make our (i) AI training materials and recording, (ii) data and codes, (iii) pre-analysis plan and (iv) template form available here: `https://osf.io/sz2g8/`. We declare no restrictions on sharing or re-use.

Supplementary materials

Materials and Methods
Figures S1 to S9
Tables S1 to S14

Supplementary Materials for Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science

Abel Brodeur (University of Ottawa; Institute for Replication), David Valenta (University of Ottawa), Alexandru Marcoci (University of Nottingham, University of Cambridge), Juan P. Aparicio (University of Ottawa; Institute for Replication), Derek Mikola (University of Ottawa; Institute for Replication), Bruno Barbarioli (University of Ottawa; Institute for Replication), Rohan Alexander (University of Toronto), Lachlan Deer (Tilburg University), Tom Stafford (Sheffield University), Lars Vilhuber (Cornell University), Gunther Bensch (RWI - Leibniz Institute for Economic Research), Fabio Motoki (University of Texas Rio Grande Valley; University of East Anglia), Mohamed Abdelhady (Carleton University), Yousra Abdelmoula (Carleton University), Ghina Abdul Baki (University of Ottawa), Tomás Aguirre (Centre for the Governance of AI), Sriraj Aiyer (University of Oxford), Shumi Akhtar (The University of Sydney), Farida Akhtar (Macquarie University), Melle R. Albada (Vienna University of Economics and Business), Micah Altman (MIT), David Angenendt (Technical University of Munich), Zahra Arjmandi Lari (Independent researcher), Jorge Armando De León Tejada (Universidad del Rosario), David Rodriguez Arana (Universidad del Rosario), Igor Asanov (International Center for Higher Education Research and Faculty of Economics, University of Kassel), Anastasiya-Mariya Asanov Noha (University of Kassel, INCHER), Rebecca Ashong (University of Ghana), Tobias Auer (London School of Economics), Francisco J. Bahamonde-Birke (Tilburg University), Bradley J. Baker (Temple University), Söhnke M. Bartram (University of Warwick and CEPR), Dongqi Bao (University of Zurich), Lucija Batinovic (Linköping University), Tommaso Batistoni (University of Oxford), Monica Beeder (University of Southampton), Louis-Philippe Beland (Carleton University), Carsten Gero Bienz (Norwegian School of Economics), Christ Billy Aryanto (Faculty of Psychology, Atma Jaya Catholic University of Indonesia), Cylcia Bolibaugh (University of York), Carl Bonander (University of Gothenburg), Ramiro Bravo (The University of Manchester), Egor Bronnikov (Maastricht University; George Mason University), Stephan Bruns (Hasselt University), Nino Buliskeria (Nazarbayev University), Sara

Caicedo-Silva (Universidad de los Andes), Andrea Calef (University College London, School of Management), Gustavo A. Castillo Alvarez (Universidad de Los Andes), Solomon Caulker (United Methodist University Sierra Leone), Simonas Cepenas (ISM University of Management and Economics), Arthur Chatton (Université Laval), Zirou Chen (University of Toronto), Ngozi Chioma Ewurum (Michael Okpara University of Agriculture, Umudike, Nigeria), Anda-Bianca Ciocîrlan (University of Sheffield), Felix J. Clouth (Tilburg University), Jason Collins (University of Technology Sydney), Nikolai Cook (Wilfrid Laurier University), Cesar Cornejo (The London School of Hygiene & Tropical Medicine), João Craveiro (University of Sheffield), Jonathan Créchet (University of Ottawa), Jing Cui (University of Ottawa), Niveditha Chalil Vayalabron (School of Earth and Planetary Science, National Institute of Science Education and Research, India), Christian Czymara (Netherlands Interdisciplinary Demographic Institute), Carlos Daniel Bermúdez Jaramillo (Universidad del Rosario), Hannes Datta (Tilburg University), Lien Denoo (Tilburg University), Arshia Dhaliwal (Carleton University), Nency Dhameja (Binghamton University), Elodie Djemai (Université Paris-Dauphine), Erwan Dujeancourt (Stockholm University), Uğurcan Dündar (Vienna University of Economics and Business), Thibaut Duprey (Bank of Canada), Yasmine Eissa (The American University in Cairo), Youssef El Fassi (HEC Lausanne), Ismail El Fassi (University of St. Gallen), Keaton Ellis (UC Berkeley), Ali Elminejad (Nazarbayev University), Mahmoud Elsherif (University of Birmingham and Leicester), Aysil Emirmahmutoglu (NHH Norwegian School of Economics), Giulian Etingin-Frati (University of Zurich), Emeka Eze (Michael Okpara University of Agriculture), Jan Fabian Dollbaum (University College Dublin), Jan Feld (Victoria University of Wellington), Andres Felipe Rengifo Jaramillo (Business School; Universidad de los Andes), Guidon Fenig (University of Ottawa), Victoria Fernandes (Bank of Canada), Lenka Fiala (University of Ottawa; Institute for Replication; Tilburg University), Lukas Fink (FU Berlin), Mojtaba Firouzjaeiangalougah (Masaryk University), Sara Fish (Harvard University), Jack Fitzgerald (Vrije Universiteit Amsterdam), Rachel Forshaw (Heriot-Watt University), Alexandre Fortier-Chouinard (Université Laval), Louis Fréget (CEPREMAP), Joris Frese (European University Institute), Jacopo Gabani (World Bank; Centre for Health Economics, University of York), Sebastian Gallegos (UAI Business School), Max C. Gamill (University of Sheffield), Attila Gáspár (HUN-REN Centre for Economic and Regional Studies), Romain Gauriot (Deakin University), Evelina Gavrilova (NHH Norwegian School of Economics), Diogo Geraldes (University College Dublin), Giulio Gia-

como Cantone (University of Sussex), Grant Gibson (McMaster University & CRDCN), Dirk Goldschmitt (University of Sheffield), Amélie Gourdon-Kanhukamwe (King's College London), Andrea Gregor de Varda (University of Milano-Bicocca), Idaliya Grigoryeva (UC San Diego), Alexi Gugushvili (University of Oslo), Aaron H.A. Fletcher (University of Sheffield), Florian Habermann (University of Lausanne), Márton Hablicsek (Leiden University), Joanne Haddad (Université Libre de Bruxelles), Jonathan D. Hall (University of Alabama), Olle Hammar (Linnaeus University and Institute for Futures Studies), Malek Hassouneh (University of Toronto), Carina I. Hausladen (ETH Zürich), Sophie C. F. Hendrikse (Tilburg University), Matthew Hepplewhite (University of Oxford), Anson T. Y. Ho (Toronto Metropolitan University), Senan Hogan-Hennessy (Cornell University), Elliot Howley (University of Nottingham), Gaoyang Huang (Swiss Federal Institute of Technology in Zurich), Héloïse Hulstaert (Hasselt University; University of Liège), Zlatomira G. Ilchovska (University of York; University of Birmingham), Paola Jaimes Santamaria (Center for Economic and Policy Research), Niklas Jakobsson (Karlstad University), Joakim Jansson (Linnaeus University; Research Institute of Industrial Economics), Ewa Jarosz (University of Warsaw), Hossein Jebeli (Bank of Canada), Yanchen Jiang (Harvard University), Hiba Junaid (Bart's Life Sciences, Bart's Health NHS Trust; Queen Mary University of London), Rohan Kalluraya (Cornell University), Sunny Karim (Carleton University), Edmund Kelly (University of Oxford), Eva Kimel (University of York), Sorravich Kingsuwankul (Vrije Universiteit Amsterdam), Valentin Klotzbücher (University of Freiburg), Daniel Krähmer (University of Munich), Pijus Krūminas (ISM University of Management and Economics), Nicholas Kruus (Schelling Research), Essi Kujansuu (University of Innsbruck), Christoph F. Kurz (Ludwig-Maximilians-Universität Munich) Stephan Küster (Freie Universität Berlin), Blake Lee-Whiting (University of Toronto), Felix Lewandowski (University of Nottingham), Tongzhe Li (University of Guelph), Ruoxi Li (Yale University), Dan Liu (Australian National University), Jiacheng Liu (Purdue University), Helix Lo (University of Tokyo), Katharina Loter (Tilburg University), Felipe Macedo Dias (Cornell University), Christopher R. Madan (University of Nottingham), Nicolas Mäder (University of San Diego), Marco Mandas (University of Cagliari), Cesar Mantilla (Pontificia Universidad Javeriana), Jan Marcus (FU Berlin), Diego Marino Fages (Durham University), Xavier Martin (Tilburg University), Ryan McWay (University of Minnesota), Daniel Medina-Gaspar (Universidad EAFIT), Sisi Meng (Cornell University), Lingyu Meng (University of Sheffield), Simon Merz (University of Halle), Alex P. Miller (Univer-

sity of Southern California), Thibault Mirabel (Equalis Capital), Dibya Deepta Mishra (Rice University), Sumit Mishra (Krea University), Belay W. Moges (Dilla University), Morteza Mohandes Mojarrad (Tilburg university), Myra Mohnen (University of Ottawa), Louis-Philippe Morin (University of Ottawa), Lucija Muehlenbachs (University of Calgary), Gastón Mullin (Tilburg University), Andreea Musulan (University of Montreal), Sara Muzzì (Department of Statistics and Quantitative Methods and Department of Medicine and Surgery, University of Milan-Bicocca), James A. C. Myers (University of Sheffield), Florian Neubauer (RWI - Leibniz Institute for Economic Research), Tuan Nguyen (Hasselt University), Ali Niazi (University of Calgary), Ardyn Nordstrom (Carleton University), Bartłomiej Nowak (Cardinal Stefan Wyszynski University), Daneal O'Habib (Bank of Canada), Tim Ölkers (University of Göttingen), Justin Ong (University of Sheffield), Valeria Orozco Castiblanco (IESE, Universidad de Navarra), Ömer Özak (SMU), Ali I. Ozkes (SKEMA Business School, GREDEG, Université Côte d'Azur), Mikael Paaso (Erasmus University Rotterdam), Shubham Pandey (Universität Osnabrück), Varvara Papazoglou (University of Sheffield), Romeo Penheiro (University of Houston), Linh Pham (Lake Forest College), Ulrike Phieler (Vienna University of Economics and Business), Peter Pütz (Bielefeld University), Quan Qi (University at Albany, SUNY), Jingyi Qiu (University of Michigan), David A. Reinstein (The Unjournal), Juuso Repo (INVEST Flagship Research Center, University of Turku), Nicolas Rudolf (University of Lausanne), Shree Saha (Cornell University), Orkun Saka (City St George's, University of London), Chiara Saponaro (University of Milano-Bicocca), Georg Sator (University of Nottingham), Martijn Schoenmakers (Tilburg University), Raffaello Seri (InsIDE Lab, DiEco, Università degli Studi dell'Insubria), Meet Shah (Toronto Metropolitan University), Paul Sibille (University of Liege), Christoph Siemroth (University of Essex), Vladimir Skavysh (Bank of Canada), Ben Slater (University of Cambridge), Wenting Song (University of California, Davis), Stefan Staubli (University of Calgary), Tobias Steindl (University of Regensburg), Nomwendé Steven Waongo (University of Ottawa), Paul Stott (University of Manchester), Stephenson Strobel (McMaster University), Roshini Sudhaharan (Tilburg University), Pu Sun (Dongbei University of Finance and Economics), Scott D. Swain (Clemson University), Oleksandr Talavera (University of Birmingham), Hanz M. Tantiangco (University of Sheffield), Georgy Tarasenko (Cornell University), Boyd Tarlinton (Department of Primary Industries, Queensland), Mariam Tarraf (Carleton University), Ken Teoh (International Monetary Fund), Rémi Thériault (Université du Québec à Montréal), Bethan Thompson

(SRUC), Tonghui Tian (Carleton University), Wenjie Tian (University of Ottawa), Manuel Tobias Rein (Tilburg University), Emmanuel Tolani (University of Bonn), Nicolai Topstad Borgen (University of Oslo), Solveig Topstad Borgen (University of Oslo), Javier Torralba (Tilburg University), Carolina Velez-Ospina (World Bank), Man Wai Mak (Carleton University), Lukas Wallrich (Birkbeck, University of London), Zeyang Wang (Vanderbilt University), Leah Ward (University of Manchester), Matthew D. Webb (Carleton University), Duncan Webb (Princeton University), Bryan S. Weber (College of Staten Island, CUNY), Christoph Weber (ESSCA School of Management), Wei-Chien Weng (National Taiwan University), Christian Westheide (University of Vienna), Tom Wilkinson (University of Sheffield), Kwong-Yu Wong (National University of Singapore), Marcin Wroński (Collegium of World Economy, SGH Warsaw School of Economics), Zhuangchen Wu (University of Birmingham), Qixia Wu (University of Ottawa), Victor Y. Wu (Stanford University), Bohan Xiao (University of Ottawa), Feihong Xu (Northwestern University), Cong Xu (National Chengchi University; Aalto University), Pranav Yadav (Tilburg University), Yu Yang Chou (University College London), Luther Yap (Princeton University), Myra Yazbeck (University of Ottawa), Bo Yao (Lancaster University), Zuzanna Zagrodzka (University of Sheffield), Tahreen Zahra (Carleton University), Mirela Zaneva (University of Oxford), Xiaomeng Zhang (Nanjing Audit University), Ziwei Zhao (University of Lausanne; Swiss Finance Institute), Han Zhong (University of Toronto), Aras Zirgulis (ISM University of Management and Economics), Jiacheng Zou (Columbia University), Floris Zoutman (NHH Norwegian School of Economics), Christelle Zozoungbo (Penn State University).

This PDF file includes:

Materials and Methods

A version-tagged copy of the code and data is permanently archived at `https://github.com/I4Replication/AI-Games`. All our materials are available here: `https://osf.io/sz2g8/`.

Pre-Registration

Our preanalysis plan was preregistered on the Open Science Framework (OSF) on May 2nd, 2024: `https://osf.io/sz2g8/`. The preregistration was done after our pilot event at the University of Toronto.

Of note, the preanalysis plan refers to AI-assisted teams as cyborg teams and AI-led teams as machine teams. AI-led teams are also referred to as 'machines with limited human assistance.'

We consider the 2025 games as not preregistered. The 2025 virtual games were conceived and added after receiving reviews from four reviewers at *Nature*.

We note only the following deviations from the preanalysis plan for the 2024 AI games:

- The preanalysis plan mentions the Ottawa, Sheffield, Cornell, Cambridge, and Tilburg games. We ended up not organizing games at the University of Cambridge and replaced those games with the Bogota games. We also mentioned in the preanalysis plan that we were hoping to have at least one more event in 2024/2025. We added a 2024 fully virtual game session on November 22nd, 2024 (prior to submitting to Nature).

- The preanalysis plan mentions three dependent variables for the robustness checks. We added a fourth dependent variable: the replicators managed to implement their two robustness checks. The other three dependent variables are all preregistered.

- We included additional secondary analyses in the manuscript: (i) OLS regressions testing differences for R and Stata teams, (ii) heterogeneity analysis by number of prompts, (iii) heterogeneity analysis by experience.

- We added qualitative analyses during the peer review process at Nature.

- We rely on Kaplan-Meier curves following a suggestion from a reviewer at Nature.

Research Questions

Here are the primary research questions that were preregistered:

1. Do AI-led teams computationally reproduce more results than AI-assisted and human-only teams?

2. Are AI-led teams faster to computationally reproduce results than AI-assisted and human-only teams?

3. Do AI-led teams detect more major and minor coding errors or data irregularities than AI-assisted and human teams?

4. Are AI-led teams faster at detecting major and minor coding errors or data irregularities than AI-assisted and human teams?

5. Do AI-led teams propose better robustness checks than AI-assisted and human-only teams?

6. Are AI-led teams more capable of implementing robustness checks than AI-assisted and human-only teams?

7. Do AI-assisted teams computationally reproduce more results than human-only teams?

8. Are AI-assisted teams faster to computationally reproduce results than human-only teams?

9. Do AI-assisted teams detect more major and minor coding errors or data irregularities than human-only teams?

10. Are AI-assisted teams faster at detecting major and minor coding errors or data irregularities than human-only teams?

11. Do AI-assisted teams propose better robustness checks than human-only teams? 12. Are AI-assisted teams more capable of implementing robustness checks than human-only teams?

We also explored the following exploratory (preregistered) research questions:

13. Are AI-led teams improving their performance over time at computationally reproducing results, detecting coding errors or data irregularities, and providing good robustness checks?"

14. Are AI-assisted teams improving their performance over time at computationally reproducing results, detecting coding errors or data irregularities, and providing good robustness checks?

We also tackle an exploratory research question that was not preregistered in the article:

15. Do AI-assisted teams overrely or underrely on AI?

AI Replication Games Advertisement

The Institute for Replication advertised the AI replication games through social media (Bluesky and X) and emails. Events were also promoted on the Institute's webpage (`https://i4replication.org/games.html`). Only graduate students, postdoctoral fellows, professors and researchers from non-academic organizations with a PhD could register. All participants were promised coauthorship to this paper.

The typical social media posts included the following information:

"This is a one-day event that brings researchers together to collaborate on reproducing quantitative results published in high-ranking social science journals. You will have the opportunity to network with fellow researchers and develop your coding and AI skills.

Open to all researchers: faculty, post-docs, and graduate students. Knowledge of Python or R or Stata is essential. Participants will be randomly assigned to one of three teams: Machine with restricted human assistance, Cyborg or Human.

All participants will get coauthorship on a meta-research journal paper which combines the work of all teams.

Register here: Link to Registration Form."

Participants Exclusion

We did not accept registration from participants with no knowledge of Stata nor R. We also excluded from participating a very small number of researchers with no knowledge of R and who did not have a Stata license.

As noted in the main text, a few organizers participated in one of the games. They did not know about the papers to be reproduced at their respective event.

Randomization

As mentioned in the main text, randomization was carried out in two steps for each of the seven events. In step one, coauthors were randomly assigned to a team of three, conditional on the software

preferences reported by participants (Stata or R) and the mode of participation (in person or virtual). In step two, each team was randomly assigned to one of three treatment arms.

Each team was assigned a study from leading social science journals (i.e., economics, political science, or behavioral science/psychology). All team members within a team were not necessarily from the same field. Studies were assigned based on knowledge of R and Stata. In practice, most psychologists are more comfortable with R, so they were assigned a behavioral or political science study, whereas economists are more comfortable with Stata and were thus much more likely to be assigned an economics study.

Documents to be Filled During the Games

During the event, each team filled out an Excel sheet documenting their outcomes. See the Excel document here: `https://osf.io/sz2g8/`. The document "Template Time Stamp" includes 3 sheets to be filled by each team. The first sheet is for computational reproducibility. Teams need to fill out the time that they have computationally reproduced the exhibit. The second sheet documents coding errors detected. Teams need to add a row for each coding error and data irregularity and enter the time they have detected them. In the last sheet, teams need to provide a description of their two robustness checks and provide estimates if they could implement those. Researchers in the AI-assisted and AI-led groups are also asked to share their prompts/conversations at the end of the event.

Selection of Papers Used During Events

For each event, two studies published in leading social science journals are selected by AB. The studies are published in a journal with a data and code availability policy. One study is coded in Stata; the other is coded in R. The studies have all been reproduced by the Institute for Replication before the AI replication games. The Institute for Replication runs about two "regular" replication games, in contrast to these AI events, each month. At every such event, teams of researchers try to reproduce results from peer-reviewed publications. They then prepare reports of their findings, which are subsequently shared with the original authors and made public on average six months following an event. Importantly, this means the Institute for Replication had over 20 published studies with known reproduction results *but have not yet been made publicly available* to choose from at

any point in 2024. We could not take studies with publicly known coding issues since ChatGPT may be able to "know" coding errors or data discrepancies without "finding" them. This list of papers with known reproduction results that had not yet been made public is the corpus we sampled from for each of the AI replication games.

The sampling cannot be random or blind for a few reasons. First, the variation in reproduction packages (sometimes called *replication* packages in the social sciences) is too large. In both scenarios, where folders reproduce studies perfectly or folders that cannot be deciphered at all, would yield no variation in at least one of our outcomes. (No coding errors exist in the former; we can't evaluate the correctness of the code in the latter.) Second, studies need to rely on publicly available data and codes, or the exercise is futile. Third, we need to match the software abilities of participants to each study. Within this corpus, we selected studies known to have coding errors or data irregularities. All teams were told that they needed to uncover coding errors or data irregularities.

Some studies were used for two events. The following studies were selected:

Pilot: Toronto Replication Games (with virtual researchers in Europe):

1-X. Labandeira et al., "Major Reforms in Electricity Pricing: Evidence from a Quasi-Experiment", The Economic Journal, (2022), vol.132(May): 1517–1541, DOI: 10.1093/ej/ueab076. Replication package: `https://zenodo.org/records/5423782`.

2-P. Christensen and C. Timmins, "Sorting or Steering: The Effects of Housing Discrimination on Neighborhood Choice", Journal of Political Economy, (2022), vol.130(August): 2110–2163, DOI: 10.1086/720140. Replication package: `https://github.com/peterchristensen/Sorting-or-Steering`.

Materials: Documents shared on Dropbox with participants. We did not provide the screenshots for this pilot event.

Ottawa Replication Games:

1-X. Labandeira et al., "Major Reforms in Electricity Pricing: Evidence from a Quasi-Experiment", The Economic Journal, (2022), vol.132(May): 1517–1541, DOI: 10.1093/ej/ueab076. Replication package: `https://zenodo.org/records/5423782`.

2-M. Wolfowicz et al., "Arrests and Convictions but Not Sentence Length Deter Terrorism in 28 European Union Member States" Nature Human Behaviour, vol.7: (2023), 1878–1889, DOI: 10.1038/s41562-023-01695-6. Replication package: `https://doi.org/10.5281/zenodo.8196717`.

Materials: https://osf.io/5v2km/

Sheffield Replication Games:

1-P. Atanasov et al., "Taste-Based Gender Favouritism In High-Stake Decisions: Evidence from the Price is Right", The Economic Journal, (2023), vol.134(February): 856-883, DOI: 10.1093/ej/uead087. Replication package: https://doi.org/10.5281/zenodo.8372384.

2-R. Bajo-Buenestado and M. A. Borrella-Mas, "The Heterogeneous Tax Pass-Through Under Different Vertical Relationships", The Economic Journal, (2022), vol.132(July): 1684–1708, DOI: 10.1093/ej/ueac007. Replication package: https://doi.org/10.5281/zenodo.5824590.

Materials: https://osf.io/z48ax/

Cornell Replication Games:

1-S. Hill and M. E. Roberts, "Acquiescence Bias Inflates Estimates of Conspiratorial Beliefs and Political Misperceptions", Political Analysis, (2023) vol.31: 575–590, DOI: 10.1017/pan.2022.2. Replication package: https://doi.org/10.7910/DVN/TVJCTX.

2-R. Bajo-Buenestado and M. A. Borrella-Mas, "The Heterogeneous Tax Pass-Through Under Different Vertical Relationships", The Economic Journal, (2022), vol.132(July): 1684–1708, DOI: 10.1093/ej/ueac007. Replication package: https://doi.org/10.5281/zenodo.5824590.

Materials: https://osf.io/ncje7/

Bogota Replication Games:

1-S. Hill and M. E. Roberts, "Acquiescence Bias Inflates Estimates of Conspiratorial Beliefs and Political Misperceptions", Political Analysis, (2023) vol.31: 575–590, DOI: 10.1017/pan.2022.2. Replication package: https://doi.org/10.7910/DVN/TVJCTX.

2-M. Comola and S. Prina, "The Interplay Among Savings Accounts and Network-Based Financial Arrangements: Evidence from a Field Experiment", The Economic Journal, (2023), vol.133(January): 516–535, DOI: 10.1093/ej/ueac053. Replication package: https://doi.org/10.5281/zenodo.6985683.

Materials: https://osf.io/hx67q/

Tilburg Replication Games:

1-N. Lee, "Do Policy Makers Listen to Experts? Evidence from a National Survey of Local and State Policy Makers", American Political Science Review, (2022), vol.116(2): 677-688, DOI: 10.1017/S0003055421000800. Replication package: https://doi.org/10.7910/DVN/S2SNOT.

2-S. B. Holt and K. Vinopal, "Examining Inequality in the Time Cost of Waiting", Nature Human Behaviour, (2023), vol.7: 545–555, DOI: 10.1038/s41562-023-01524-w. Replication package: https://github.com/stevebholt/waiting-time.

Materials: https://osf.io/dqw5y/

Virtual Replication Games: Europe (Part I)

1-P. Ager et al., "How the Other Half Died: Immigration and Mortality in U.S. Cities", The Review of Economic Studies, (2024), vol.91(1): 1–44, DOI: 10.1093/restud/rdad035. Replication package: https://dx.doi.org/10.5281/zenodo.7506459.

2-S. Herskowitz, "Gambling, Saving, and Lumpy Liquidity Needs", American Economic Journal: Applied Economics, (2021), vol.13(1): 72–104, DOI: 10.1257/app.20180177. Replication package: https://www.openicpsr.org/openicpsr/project/115162/version/V1/view.

Materials: https://osf.io/tcn7k/

Virtual Replication Games: North America (Part II)

3-S. Herskowitz, "Gambling, Saving, and Lumpy Liquidity Needs", American Economic Journal: Applied Economics, (2021), vol.13(1): 72–104, DOI: 10.1257/app.20180177. Replication package: https://www.openicpsr.org/openicpsr/project/115162/version/V1/view.

4-A. G. de Zavala et al., "Mindful-Gratitude Practice Reduces Prejudice at High Levels of Collective Narcissism", Psychological Science, (2024), vol.35(2): 137-149, DOI: 10.1177/09567976231220902. Replication package: https://osf.io/t7kxa/?view_only=39c692dbf3034e1593b07906cf3e635a.

Materials: https://osf.io/67925/

2025 Virtual Replication Games: Europe and North America (Study II)

1-A. Goldenberg et al., "Homophily and Acrophily as Drivers of Political Segregation", Nature Human Behaviour, (2023) vol.7, 219–230, DOI: 10.1038/s41562-022-01474-9. Replication package: Data/code for study I is available here: https://osf.io/nz4dk/. Data/code for studies 2–4 is available here: https://osf.io/649fq/. Data/code for the agent-based model is available here: https://osf.io/ad7vh/.

2-A. Kennard, "Who Controls the Past: Far-Sighted Bargaining in International Regimes", American Journal of Political Science, (2023), vol.67(3), 553-568, DOI: 10.1111/ajps.12747. Replication package: https://doi.org/10.7910/DVN/5M10KG..

3-R. M. Gonzalez, "Cell Phone Access and Election Fraud: Evidence from a Spatial Regression

Discontinuity Design in Afghanistan", American Economic Journal: Applied Economics (2021), vol.13(2), 1–51, DOI: 10.1257/app.20190443. Replication package: https://www.openicpsr.org/openicpsr/project/118467/version/V1/view.

4-M. Mohanan et al., "Different Strokes for Different Folks? Experimental Evidence on the Effectiveness of Input and Output Incentive Contracts for Health Care Providers with Varying Skills", American Economic Journal: Applied Economics, (2021), vol.13(4), 34–69, DOI: 10.1257/app.20190220. Replication package: https://www.openicpsr.org/openicpsr/project/119741/version/V1/view.

5-A. Banerjee et al., "E-governance, Accountability, and Leakage in Public Programs: Experimental Evidence from a Financial Management Reform in India", American Economic Journal: Applied Economics, (2020), vol.12(4), 39–72, DOI: 10.1257/app.20180302. Replication package: https://www.aeaweb.org/journals/dataset?id=10.1257/app.20180302.

6-N. Ajzenman, "The Power of Example: Corruption Spurs Corruption", American Economic Journal: Applied Economics, (2021), vol.13(2), 230–57, DOI: 10.1257/app.20180612. Replication package: https://www.openicpsr.org/openicpsr/project/118971/version/V1/view.

Materials: https://osf.io/3g958/

We also included a note in the pdf of the article mentioning something along the lines of: "First, computationally reproduce Table X, columns Y and Z. Second, detect coding errors/data irregularities. Last, propose and try to implement two robustness check."

AI Training

Researchers took part in a one-hour-long training on the usage of ChatGPT. This training was mandatory for the researchers in AI-assisted and AI-led groups.

Recordings and materials are publicly available here: https://osf.io/sz2g8/. The training included the following topics:

1. **Introduction, Overview of ChatGPT, and Access**

    - Introduction to the capabilities of ChatGPT and its applications in reproducing scientific studies, coding, and data analysis.

- Instructions on accessing ChatGPT, creating an account, and accessing the Institute for Replication workspace/team subscription.
- Explanation of subscription tiers, model capabilities, limitations on message usage, and privacy settings.

2. **Interaction with ChatGPT**

   - Techniques for optimizing prompts and ChatGPT's responses, such as providing contextual information.
   - Strategies to manage randomness in outputs or when the model gets "stuck," such as opening new chats and regenerating answers.

3. **Sharing Chats**

   - Information on how to generate shareable links to sessions and manage privacy, including restrictions on who can access shared chats.
   - Explanation on how to save chats as a webpage when the chat cannot be shared as a link (e.g., when the chat includes images).

4. **Coding Assistance**

   - Explanation of how ChatGPT can assist with coding, including practical examples such as writing code for converting data formats (e.g., R's .rds to Stata's .dta) and debugging code.

5. **File and Image Upload**

   - Introduction to ChatGPT's ability to process uploaded files.
   - Overview of supported file types (e.g., PDFs, Word documents, CSVs, Excel files) and limitations regarding file size.
   - Example of uploading an academic article to inquire about research questions, identification strategy, and robustness checks.
   - Explanation of the potential benefits of uploading an image of a results table/figure instead of only the PDF file.

- Example of uploading an image of a results table from a study and inquiring about it.
- The image upload was not mentioned or demonstrated during the Toronto event.

6. **Conducting Data Analysis Using ChatGPT**

   - Introduction to using ChatGPT's Data Analysis Module for executing Python code and performing data analysis.
   - Example of uploading a replication package of an article and replicating regression analyses using the Python module.
   - AI-led teams were instructed to first attempt to run the authors' codes/scripts using the data analysis capabilities of ChatGPT. If this analysis failed, teams were instructed to run the code in their local environment by following instructions provided by ChatGPT, as introduced in the Coding Assistance example.

7. **ChatGPT API**

   - Explanation of the ChatGPT API for automating repetitive tasks and integrating AI capabilities into code.
   - Example code shown for connecting to the ChatGPT API in R.

8. **Customizing ChatGPT**

   - Information on setting up personalized models with custom instructions for specific needs.
   - Mention of ChatGPT's memory feature that retains information across sessions, and how information that should not be retained can be deleted. The ChatGPT memory feature was not mentioned during the Toronto training session.

9. **Explanation of Differences Among ChatGPT Models**

   - Differences between ChatGPT 4 and 4o were first discussed during the Sheffield training event.
   - Introduction of GPT-o1-preview and GPT-o1-mini models was first provided during the Bogota event.

- Capabilities of ChatGPT 4o with canvas were introduced during the last event.

ChatGPT Models

Researchers in the AI-assisted and AI-led groups were provided with access to ChatGPT Team. Table S1 presents an overview of the ChatGPT models available to these researchers during each event. Table S2 provides details about the capabilities of these models. Throughout all events, researchers had access to the main flagship model, GPT-4, and/or GPT-4o. These models were capable of processing files, equipped with a Python environment for interpreting code and conducting data analysis, and had internet access.

The file upload was limited to a maximum of 512MB per file, and further limited to two million tokens for text files, approximately 50MB for CSV files and spreadsheets, and 20MB per image for images. A user file size is capped at 10GB and organization at 100GB. (*51*) However, the practical limitations based on the Python environment's capabilities were likely lower.

Only researchers in the Bogota, Tilburg, and virtual-only events had access to the GPT-o1-preview and GPT-o1-mini models. These models were trained using reinforcement learning to perform complex reasoning and, unlike the 4/4o models, can produce an internal chain of thought before responding to users. (*52*)

Usage limits for certain models were applied by OpenAI. During the Toronto and Ottawa events, these limits were explicitly stated, with the Team subscription limit set at 100 messages per three hours per user. Researchers were instructed to collaborate with their teammates if the limit was reached or use the unlimited GPT-3.5 model. For the remaining events, usage limits for the GPT-4/4o models were no longer explicitly mentioned by OpenAI but were likely higher. The GPT-o1-preview model was limited to 50 queries per week, while GPT-o1-mini was limited to 50 queries per day.

ChatGPT Prompts

We conducted an exploratory analysis of ChatGPT transcripts. This revealed meaningful variation in prompting styles: in iteration, delegation, and over-reliance on AI suggestions. These patterns could form a basis for a future typology of prompting behaviors or failure modes. Preliminary review of AI-led transcript logs suggests that common failure modes included: (i) reliance on default

Games	Date	Training Date	Image*	ChatGPT versions available						
				3.5	4	4o	4o-mini	o1-preview and o1-mini	4o with canvas	4.5, o3, o4-mini, and o4-mini-high
Toronto	Feb 20	Feb 14	No	Yes	Yes					
Ottawa	May 3	Apr 26	Yes	Yes	Yes					
Sheffield	Jun 17	Jun 12	Yes	Yes	Yes	Yes				
Cornell	Aug 12	Jul 31	Yes		Yes	Yes	Yes			
Bogota	Oct 4	Sep 23†	Yes		Yes	Yes	Yes	Yes	Yes‡	
Tilburg	Oct 18	Sep 30	Yes		Yes	Yes	Yes	Yes	Yes‡	
Virtual	Nov 22	Nov 8	Yes		Yes	Yes	Yes	Yes	Yes	
2025 Event	Apr 30	Apr 28	Yes		Yes	Yes	Yes		Yes§	Yes

* Image upload trained as part of the pre-games training and screenshots of relevant results from the studies provided to researchers.
† Training using recording of the Cornell training + o1-preview model slide added to presentation
‡ While GPT-4o with canvas was available for the Bogota and Tilburg events, it was not mentioned during the training.
§ Canvas become generally integrated into the environment for the various models.

Table S1: ChatGPT models available by training

package behavior without validation, (ii) misunderstanding file structure or model specifications, (iii) hallucination of non-existent variables or steps, (iv) limited ability to debug code execution errors, and (v) lack of prompt iteration or self-critique.

As a case study, we investigate the prompts from two teams of three researchers.

In the first team, each user brought value, but in distinct ways. Teammate 1 excelled in questioning the correctness of the figure generation process and raised crucial flags about the interpretability of marginal effects – though they sometimes required nudging to resolve issues independently. Teammate 2, while technically competent, engaged less rigorously with the underlying model logic and accepted ChatGPT's answers too readily, potentially missing subtle but important discrepancies. Teammate 3 was the most thorough and experimental, implementing multiple robustness checks (cross-validation, outlier drops, subgroup splits) and producing interpretable, well-supported diagnostic output. Teammate 3's iterative prompting was both strategic and technically sound, positioning ChatGPT more as a collaborator than a calculator.

Takeaway: To maximize ChatGPT's potential for replication, teams benefit most from strategic prompting (Teammate 3), skeptical review of outputs (Teammate 1), and a clear sense of why

Model	Date Introduced	File Upload	Python Code Interpreter	Web Browsing	Reasoning
GPT-3.5	Before 1st event	No	No	No	No
GPT-4	Before 1st event	Yes	Yes	Yes	No
GPT-4o	May 13, 2024[1]	Yes	Yes	Yes	No
GPT-4o-mini	July 18, 2024[2]	Yes*	Yes*	Yes*	No
o1-preview	September 12, 2024[3]	No	No	No	Yes
o1-mini	September 12, 2024[3]	No	No	No	Yes
GPT-4o with canvas	October 3, 2024[4]	Yes	Yes	No	No
GPT-4.5 (preview)	February 27, 2025[5]	Yes	Yes	Yes	No
o3	April 16, 2025[6]	Yes	Yes	Yes	Yes
o4-mini	April 16, 2025[6]	Yes	Yes	Yes	Yes
o4-mini-high	April 16, 2025[6]	Yes	Yes	Yes	Yes

* While 4o-mini supported these functions at the time of the last training it did not necessarily at the time of introduction.
[1] `https://openai.com/index/hello-gpt-4o/`
[2] `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`
[3] `https://openai.com/index/introducing-openai-o1-preview/`
[4] `https://openai.com/index/introducing-canvas/`
[5] `https://openai.com/index/introducing-gpt-4-5/`
[6] `https://openai.com/index/introducing-o3-and-o4-mini/`

Table S2: ChatGPT capabilities

each analysis step is done. Future participants might learn from Teammate 3's modular prompting, Teammate 1's critical eye, and Teammate 2's clarity of implementation–ideally combining all three.

The second team showed strong consistency in core tasks: all three members successfully reproduced Figure 4 and applied the core logistic model. The first and second teammates shared overlapping concerns around code reliability, variable scaling, and indexing accuracy. However, only the third teammate conducted a wide array of robustness tests (cross-validation, subgroup splits, clustering, outlier removal). The second teammate completed reproduction but did not explore model sensitivity or robustness.

The biggest missed opportunity was a lack of follow-up on proposed robustness ideas (standardization, TF-IDF alternatives, placebo tests), despite their clear feasibility. A more integrated approach would have helped cross-pollinate promising ideas across team members.

Coding Errors and Data Irregularities

All studies to be reproduced had known coding errors that had been identified by the lead authors in a prior study but were not publicly disclosed at the time of the AI replication game. Of note, participants during the AI replication games identified additional errors.

We define coding errors as minor or major depending on whether the coding error could, in theory, have an impact on the claims tested. AB, JA, and DM discussed all errors uncovered during the AI games and classified coding errors as major or minor. Coding errors uncovered range from minor errors, such as reporting the wrong p-value or not specifying in the article the inclusion of a control variable, to major coding errors, such as miscoding the dependent variable or main independent variable and conducting a many-to-many merge instead of a many-to-one merge.

We did not keep track of the number of times that the three of us discussed hard cases (i.e., whether a reported error was not an error, a minor error or a major error), but it was rare. For the 2025 AI games (study II), we kept track of these disagreements. We always agreed on whether a reported error was truly an error (e.g., we always agreed that missing packages/paths or versioning issues were not errors), and all agreed on whether errors were major or minor for 47 out 50 errors uncovered by teams.

In what follows, we provide concrete examples of major coding errors and data irregularities. In the article entitled "Arrests and Convictions but Not Sentence Length Deter Terrorism in 28 European Union Member States," one of the major coding errors is in the coding of the dependent variable. The authors state in the article that the terrorism rate used as their dependent variable is the inverse hyperbolic sine (IHS) of the per capita rate of terrorist attacks. But the code reveals that the dependent variable takes impossible values and is thus not the IHS of the per capita rate of terrorist attacks. For instance, countries with zero terror attacks are assigned strictly positive values, which is impossible. Another major coding error is that some European countries were imputed as experiencing zero terror attacks when the number of terror attacks was missing in the raw data. There is an editor's note for this article at Nature Human Behaviour. The note was released as a result of a Matters Arising submission by one of our reproducers.

In the article "Sorting or Steering: The Effects of Housing Discrimination on Neighborhood Choice," one of the major coding errors involved assigning a value of zero for the variable 'of color'

to both individuals identified as 'white' and as 'other' in the raw data. A major data irregularity is the inclusion of fixed effects for the string variable 'city'. The raw variable is case sensitive and has many spelling mistakes. A comment detailing these errors is revised and resubmitted at the Journal of Political Economy.

In the main article, we document that AI-led groups identified significantly fewer coding errors and data irregularities. AI-led groups likely uncovered fewer major coding errors due to the nuanced and contextual nature of these errors. It is plausible that AI-led groups struggled to identify technically correct but conceptually flawed code errors. These errors, such as a many-to-many merge instead of a many-to-one merge, produce duplicate entries without causing a runtime error. While the code executes without issues, the underlying conceptual mistake leads to incorrect data handling. This type of error is particularly challenging for AI to detect, as it requires an understanding of the conceptual intent behind the code rather than just its syntactic correctness.

More generally, many coding mistakes involve subtle misapplications of statistical transformations, such as assigning incorrect values or mishandling missing data, which often require domain expertise and a deep understanding of the data's structure. AI tools, while efficient at automating tasks, may struggle with interpreting complex logical relationships, ambiguous data definitions, or recognizing implausible outcomes without explicit programming. In contrast, human-led groups are better equipped to identify errors that hinge on contextual reasoning, such as the incorrect coding of dependent variables or misassignments due to case-sensitive inconsistencies in datasets.

Robustness Checks

We propose four different binary measures which we believe qualify a good robustness check: (i) clarity (not vague) regarding purpose and execution; (ii) feasible, (iii) not previously done by the original author(s); and (iv) focuses on the validity of the empirical strategy.

In addition, we classify any corrections to major errors and rerunning the script as a "good" robustness check, although not complying with all the previous criteria. As stated in the preanalysis plan, one of AB, JA, and DM reviewed each robustness check based on the above criteria. In the event that at least one of the above categories is hard to classify, we discussed and classified together. The classification was done by JA across all events to avoid measurement error, and hard cases were discussed with AB and DM. For the 2025 AI games (Study II), AB, JA, and DM all coded the

validity of each robustness check. We agreed on over 91% of robustness checks categorization and discussed cases for which there was initial disagreement.

Clarity (not vague) regarding purpose and execution: It is possible that teams will not adequately describe their robustness check. This could be due to ChatGPT not sufficiently describing what they are doing, or from their own explanation. An example of a vague robustness check would be "adding control variables." In contrast, a clear robustness check would be to precisely document which variable should be added as a control.

Feasible: For teams that are able to implement the robustness check, we categorize the robustness check as feasible. For teams that do not implement a robustness check, the question we ask is whether or not, with more time but the same resources, it could be implemented.

Not previously done by the original author(s): All teams of reproducers–independent of which type of team they are–have access to the original study, online appendix, and the replication packages. All teams must verify that their recommended robustness check was not previously done. AB, JA, and DM verified with each study whether the proposed robustness checks were included in the article or the appendix.

Validity: While robustness checks can serve multiple purposes, we view them as alternative specifications that test the main conclusion(s) of a study. A valid robustness check tests the reliability and stability of the results. Examples of invalid robustness checks include: using bad controls, misspecified models (bad instrument), etc.

Power Calculations

Table S11 reports the results of our ex-post power calculations. Using the realized experimental data, we estimated an OLS model as specified in Table 2 for each outcome variable. We resampled residuals with replacement 10,000 times to simulate new outcomes based on the fitted model, and recorded the number of times the treatment effects on these simulated outcomes were statistically significant at the 5% level. I.e., letting Y be an outcome and X be the design matrix of independent variables, after running regression

$$Y = X\beta + \epsilon$$

and generating residuals

$$\hat{\epsilon} = Y - X\hat{\beta},$$

for each of 10,000 runs k, we resampled the elements of $\hat{\epsilon}$ with replacement to generate resampled residual vector $\hat{\epsilon}_k$. We then generated $\tilde{Y}_k = Y + \hat{\epsilon}_k$ and assessed whether estimates of the treatment effects of the AI-assisted and AI-led arms on $\tilde{Y}_k$ are statistically significantly different from zero at a 5% significance level.

Higher shares of statistically significant treatment effect estimates arising from this resampling process indicate a higher ex post power to detect the observed effect. For example, column 3 of panel A indicates that we achieve 83.8% power to detect the effect of AI-assisted treatment, and 99.9% power to detect the effect of AI-led treatment for the outcome "major errors". For the AI-assisted treatment, the power for outcomes "minor errors" and "major errors" is about 60.3% and 83.8% (respectively) for study I and 51.3% and 79.9% (respectively) when Study I and Study II are combined. For the effect of AI-led treatment, the ex-post power ranges between 78.5% and 100% for Study I and is comparable for Study II.

Focus Groups: Corpus and Analytic Approach

We applied reflexive thematic analysis in the sense of Braun and Clarke (2006) (*53*). Before coding began, we built a shared codebook seeded with theory-driven categories linked to our research questions (e.g., "task delegation," "error detection") and left space for inductive labels to emerge. Five analysts then read each transcript in full and independently tagged text segments with the evolving codebook. The coders iteratively reconciled discrepancies, refined code definitions, and updated the codebook; all individual codes remained visible in a common spreadsheet so that divergent interpretations stayed traceable. Finally, we overlaid the coding layers, examined convergent and divergent patterns across treatment arms, and iteratively collapsed related codes into higher-order themes. This reflexive yet transparent workflow preserved the nuance of individual readings while yielding a coherent thematic structure—crucial for triangulating the qualitative insights with the quantitative performance metrics reported below.

Focus Groups: Additional Secondary Results

Teams began with high hopes for AI: "we were both quite optimistic about ChatGPT and what it can do" (FG3-A-P3, 29 Apr 2025). Hopes faded quickly: "I dampened down that enthusiasm because I saw that … the work was getting done better without it" (FG4-A-P2, 30 Apr 2025) and "something that should be push-button goes from a task that should take minutes into a task that takes hours" (FG1-L-P3, 18 Apr 2025).

Many AI-assisted teams treated ChatGPT as a tool of last resort. One recalled "we started with the assumption of going through the code by ourselves" (FG3-A-P4), another "only copy-pasted the error message … if we were desperate" (FG2-A-P4), and a third "just ask quite a lot of questions honestly … only when we found an error" (FG2-A-P1). AI-led teams, by contrast, attempted fully automated runs, hoping to "put as little input as possible from myself" (FG3-L-P3). When archives overflowed the context window they switched—too late—to piecemeal prompting. Thus, speed hinged on where teams drew the *delegation threshold*: fast AI-assisted teams outsourced micro-tasks yet kept strategic control; AI-led teams ceded entire analytic stages and struggled to reclaim them.

Human expertise remained essential. One AI-led participant reflected, "Let's approach the analysis as researchers … build it up from the bottom up" (FG3-L-P3). Yet they still felt hopeless at error-hunting: "We could have done this for ten more hours—we would not find the error. No way." (FG1-L-P3). Others blamed AI's blind spots: "[ChatGPT] is very convenient for summarising stuff but then there's this five percent it drops which is crucial" (FG1-A-P5) and "some issues are so particular—say institutional details relevant for an identification strategy" (FG3-H-P2).

Across AI arms, a pragmatic repertoire emerged: specify software version, feed minimal code snippets, and request pseudocode before executable commands. Still, prompting was effortful: "Eventually I just ran out of ideas on how else to prompt it" (FG6-L-P3). Over-confidence compounded matters; once ChatGPT raised no alarms, a teammate admitted "our second trial by ourselves was quite light … we didn't try very hard, to be honest" (FG3-A-P5). Some even deferred responsibility: "If the replication is wrong … it's GPT-4's problem" (FG5-A-P1). The resulting pipelines were fast but shallow—mirroring AI-assisted teams' lower major-error counts.

AI-led transcripts brim with exasperation: "ChatGPT was like an undergrad who hadn't used

econometrics … unapologetically arrogant when it's wrong" (FG6-L-P3). Another recalled "hours of nothing … talking to ChatGPT, trying to find an error and not getting anywhere" (FG1-L-P4). Repetitive copy-paste loops and hallucinated file paths eroded trust: "I gave it the zipped folder and it said it couldn't open it—though it had done so in the previous chat" (FG6-L-P1). AI-assisted frustration was milder—"that is wrong … and then I have to ask it again … it's just *slow*" (FG4-A-P2)—because they reverted to manual coding once cost–benefit turned negative: "The focus shifted to manual coding because the first attempt with ChatGPT failed, so I thought: let's not waste too much time there" (FG4-A-P2).

AI reshaped collaboration. AI-assisted members often kept separate ChatGPT tabs: "both … separately working one-on-one with ChatGPT, which was a little bit sad" (FG3-A-P6). Coordination costs mounted: "we were circling back discussing "Is this a fair prompt?"" (FG3-L-P3). Duplicated effort surfaced: "by the time ChatGPT produced the table, the team had already figured it out" (FG4-A-P2). Disagreement between AI and humans triggered arbitration: "ChatGPT flagged something; I tried to replicate it and couldn't" (FG4-L-P1). AI-led teams felt "the human-human connection was lower than in a standard team" (FG4-L-P3). Human-only groups relied on dense pair-programming; reading papers together was "something only humans can understand" (FG2-H-P2), a practice that surfaced subtle errors invisible to models.

AI-assisted teams accelerated routine steps yet—owing to prompt fatigue and overconfidence—missed errors that diligent human-only teams caught. AI-led teams, mired in technical bottlenecks and diminished peer oversight, achieved neither speed nor depth. As one AI-led participant put it: "We could reproduce that number super-fast … and then it was *hours of nothing*" (FG6-L-P4). Conversely, an AI-assisted analyst conceded: "ChatGPT just sped things up … we were much faster. Yeah, I don't think we were better, we were just faster" (FG4-A-P3).

Despite frustrations, participants acknowledged AI's benefits when used as a complement. AI suggested robustness checks: "some really interesting suggestions we didn't think of ourselves" (FG3-L-P3). It helped locate code snippets: "it very quickly directed us to the relevant files" (FG3-L-P3). Less-experienced coders valued boilerplate help: "GPT-4 can definitely code better than me" (FG5-A-P1). Human-arm participants admitted the absence of AI nudged them toward "low-hanging fruit," implying AI can broaden analytical ambition.

Participants also praised rapid LLM progress but warned that fully automated replication re-

mains distant: “Right now it just seems too big of a task” (FG4-L-P1).

Focus Groups: Implications for future human–AI workflows

Effective LLM use is now a specialized skill on par with fluency in `Stata` or `R`. Tool designers should prioritize larger context windows, transparent code provenance, and archive-level ingestion to cut the friction that crippled AI-led teams. Pedagogically, prompt engineering and AI-assisted debugging must join the *reproduction toolkit*. Most importantly, our findings caution against premature automation: April-2025 LLMs excel as accelerators for routine sub-tasks but falter as autonomous reproducers. Collective, deliberative human judgment currently remains indispensable for detecting the subtle conceptual and coding errors that determine whether published science withstands scrutiny.

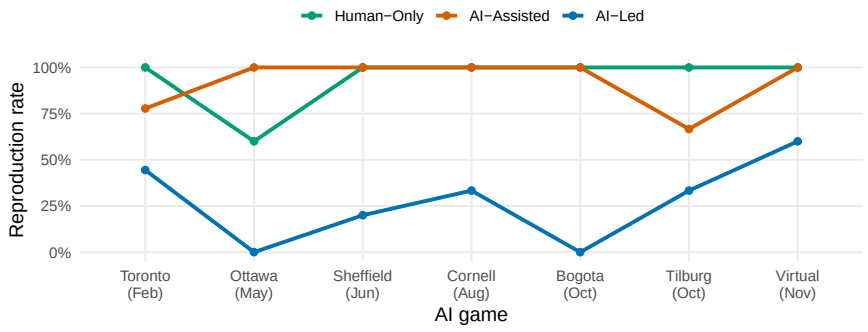

Figure S1: Computational reproducibility rates across events and treatment groups

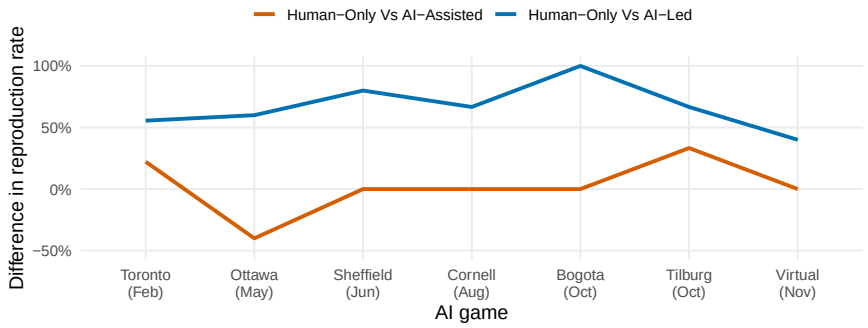

Figure S2: Difference in average computational reproducibility rate by groups across AI replication game

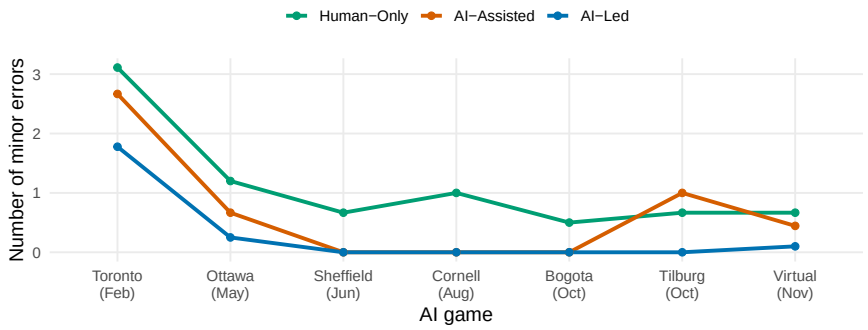

Figure S3: Number of minor errors detected across events and treatment groups

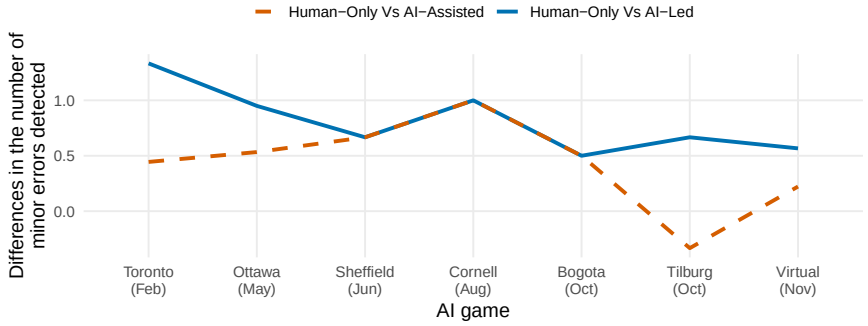

Figure S4: Difference in average minor errors by groups across AI replication game

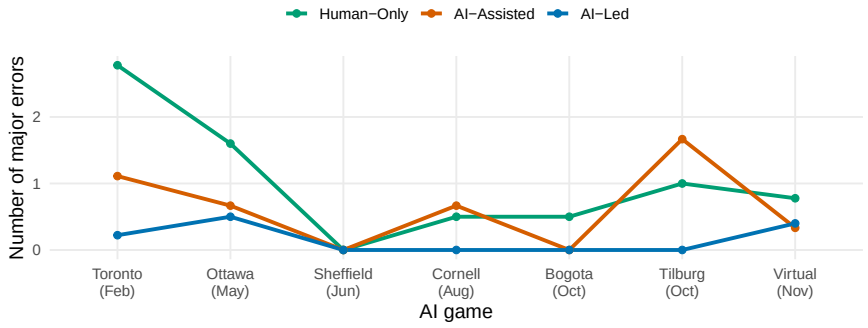

Figure S5: Number of major errors detected across events and treatment groups

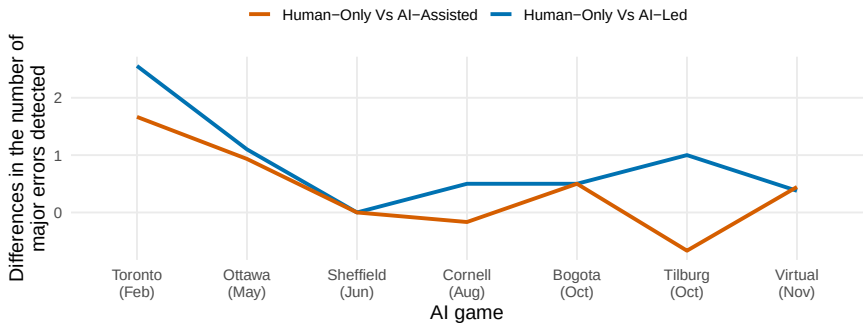

Figure S6: Difference in average major errors by groups across AI replication game

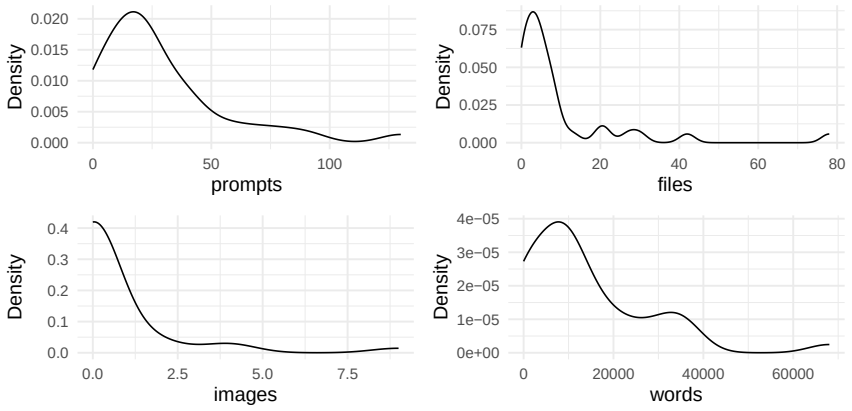

Figure S7: Distribution of number of prompts, files, images, and words used between AI-Assisted teams and ChatGPT

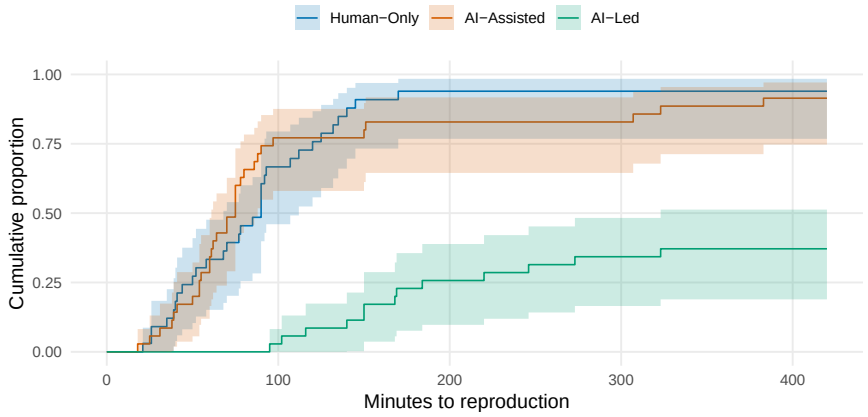

Figure S8: Kaplan-Meier curves, showing the proportion of teams who computationally reproduced the paper by time *t* across Study I

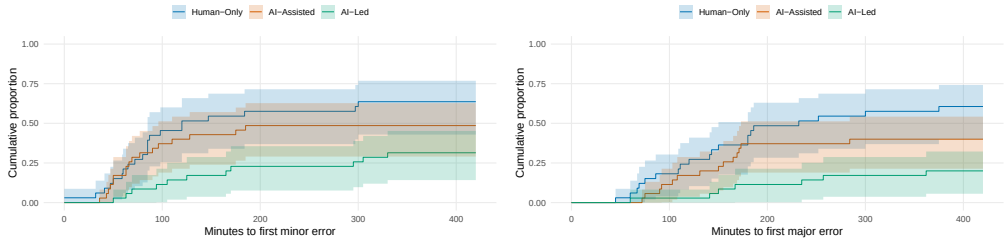

Figure S9: Kaplan-Meier curves, showing the proportion of teams who found their first coding error by time *t* across Study I

Table S3: Comparison of Human, AI-Assisted, and AI-Led Metrics

Variable	Human-Only	AI-Assisted	AI-Led	Human-Only vs AI-Assisted	Human-Only vs AI-Led	AI-Assisted vs AI-Led
Reproduction	0.939 (0.242)	0.914 (0.284)	0.371 (0.490)	0.025 [0.697]	0.568 [<0.001]	0.543 [<0.001]
Minutes to reproduction	82.0 (39.8)	93.3 (85.4)	179.7 (68.4)	-11.3 [0.505]	-97.7 [<0.001]	-86.4 [0.002]
Number of minor errors	1.424 (1.696)	0.943 (1.454)	0.514 (0.919)	0.481 [0.213]	0.910 [0.007]	0.429 [0.145]
Minutes to first minor error	100.7 (77.1)	81.9 (44.6)	161.0 (103.3)	18.7 [0.381]	-60.3 [0.071]	-79.1 [0.010]
Number of major errors	1.364 (1.496)	0.629 (0.942)	0.229 (0.490)	0.735 [0.017]	1.135 [<0.001]	0.400 [0.029]
Minutes to first major error	153.2 (86.1)	138.4 (55.9)	196.0 (97.7)	14.8 [0.577]	-42.8 [0.284]	-57.6 [0.099]
At least one good robustness check	1.000 (0.000)	1.000 (0.000)	0.829 (0.382)	0.000 [-]	0.171 [0.012]	0.171 [0.010]
At least two good robustness checks	0.879 (0.331)	0.857 (0.355)	0.629 (0.490)	0.022 [0.796]	0.250 [0.017]	0.229 [0.029]
Ran at least one good robustness check	0.939 (0.242)	0.943 (0.236)	0.571 (0.502)	-0.003 [0.953]	0.368 [<0.001]	0.371 [<0.001]
Ran at least two good robustness checks	0.788 (0.415)	0.800 (0.406)	0.457 (0.505)	-0.012 [0.903]	0.331 [0.005]	0.343 [0.003]

Note: Columns 2–4 present means and standard errors in parentheses for individual groups (Human-only, AI-Assisted, and AI-Led); columns 5–7 present differences in means and p-values in brackets for group comparisons (Human-Only vs AI-Assisted, Human-Only vs AI-Led, and AI-Assisted vs AI-Led).

Table S4: Balance of Team-Level Characteristics by Group

Variable	Human-Only	AI-Assisted	AI-Led	Human-Only vs AI-Assisted	Human-Only vs AI-Led	AI-Assisted vs AI-Led
Number of teammates	2.606 (0.496)	2.429 (0.655)	2.829 (0.568)	0.177 [0.214]	-0.223 [0.091]	-0.400 [0.008]
Minimum academic level: Professor	0.091 (0.292)	0.086 (0.284)	0.086 (0.284)	0.005 [0.941]	0.005 [0.941]	0.000 [1.000]
Minimum academic level: Postdoc	0.030 (0.174)	0.114 (0.323)	0.057 (0.236)	-0.084 [0.190]	-0.027 [0.597]	0.057 [0.400]
Minimum academic level: Researcher	0.152 (0.364)	0.171 (0.382)	0.029 (0.169)	-0.020 [0.827]	0.123 [0.076]	0.143 [0.047]
Minimum academic level: Student	0.727 (0.452)	0.629 (0.490)	0.829 (0.382)	0.099 [0.392]	-0.101 [0.321]	-0.200 [0.061]
Maximum academic level: Professor	0.576 (0.502)	0.514 (0.507)	0.686 (0.471)	0.061 [0.617]	-0.110 [0.355]	-0.171 [0.147]
Maximum academic level: Postdoc	0.152 (0.364)	0.257 (0.443)	0.143 (0.355)	-0.106 [0.289]	0.009 [0.921]	0.114 [0.238]
Maximum academic level: Researcher	0.091 (0.292)	0.057 (0.236)	0.000 (0.000)	0.034 [0.600]	0.091 [0.070]	0.057 [0.156]
Maximum academic level: Student	0.182 (0.392)	0.171 (0.382)	0.171 (0.382)	0.010 [0.912]	0.010 [0.912]	-0.000 [1.000]
Average years of coding experience	9.000 (4.484)	8.267 (3.060)	9.740 (3.365)	0.733 [0.431]	-0.740 [0.442]	-1.474 [0.059]
Min ChatGPT level: Never	0.303 (0.467)	0.143 (0.355)	0.257 (0.443)	0.160 [0.115]	0.046 [0.679]	-0.114 [0.238]
Min ChatGPT level: Beginner	0.485 (0.508)	0.629 (0.490)	0.571 (0.502)	-0.144 [0.239]	-0.087 [0.482]	0.057 [0.632]
Min ChatGPT level: Intermediate	0.152 (0.364)	0.200 (0.406)	0.143 (0.355)	-0.048 [0.607]	0.009 [0.921]	0.057 [0.533]
Min ChatGPT level: Advanced	0.061 (0.242)	0.000 (0.000)	0.029 (0.169)	0.061 [0.144]	0.032 [0.527]	-0.029 [0.321]
Max ChatGPT level: Never	0.000 (0.000)	0.029 (0.169)	0.029 (0.169)	-0.029 [0.335]	-0.029 [0.335]	-0.000 [1.000]
Max ChatGPT level: Beginner	0.152 (0.364)	0.143 (0.355)	0.086 (0.284)	0.009 [0.921]	0.066 [0.408]	0.057 [0.460]
Max ChatGPT level: Intermediate	0.515 (0.508)	0.514 (0.507)	0.629 (0.490)	0.001 [0.994]	-0.113 [0.352]	-0.114 [0.341]
Max ChatGPT level: Advanced	0.333 (0.479)	0.286 (0.458)	0.257 (0.443)	0.048 [0.677]	0.076 [0.498]	0.029 [0.792]

Note: Columns 2–4 present means and standard errors in parentheses for individual groups (Human-only, AI-Assisted, and AI-Led); the difference columns show mean differences and p-values in brackets for the indicated group comparisons.

Table S5: Causal relationship between treatment groups and reproducibility outcomes using Logit and Poisson regressions

	(1) Reproduction	(2) Minor errors	(3) Major errors	(4) Two good robustness	(5) Ran one robustness	(6) Ran two robustness
AI-Assisted	-0.724	-0.454	-0.667	-0.388	-0.959	-0.025
	(1.124)	(0.216)	(0.302)	(1.202)	(1.758)	(0.949)
	[-2.928; 1.479]	[-0.877; -0.031]	[-1.259; -0.075]	[-2.743; 1.968]	[-4.405; 2.487]	[-1.886; 1.835]
AI-Led	-6.478	-1.177	-1.802	-2.480	-3.701	-1.777
	(1.648)	(0.228)	(0.422)	(1.227)	(1.280)	(0.926)
	[-9.709; -3.247]	[-1.625; -0.729]	[-2.630; -0.974]	[-4.885; -0.075]	[-6.209; -1.192]	[-3.591; 0.037]
Model	Logit	Poisson	Poisson	Logit	Logit	Logit
Controls	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.738	0.969	0.750	0.786	0.816	0.680
p-val (AI-Assisted vs. AI-Led)	0.002	0.009	0.014	0.049	0.129	0.024
Obs.	103	98	84	103	103	103

Note: Standard errors in parentheses, confidence intervals in brackets; human-only group omitted. Marginal effects reported for Logit models.

Controls include number of teammates; game-by-software fixed effects; maximum and minimum position skill fixed effects; attendance fixed effects.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S6: Estimates for the control variables

	(1) Reproduction	(2) Minor errors	(3) Major errors	(4) One good robustness	(5) Two good robustness	(6) Ran one robustness	(7) Ran two robustness
Branch: AI-Assisted	-0.018	-0.487*	-0.646**	-0.009	-0.014	-0.032	-0.009
	(0.063)	(0.270)	(0.254)	(0.027)	(0.103)	(0.061)	(0.113)
	[-0.142; 0.105]	[-1.017; 0.043]	[-1.145; -0.147]	[-0.062; 0.045]	[-0.217; 0.188]	[-0.153; 0.088]	[-0.229; 0.212]
Branch: AI-Led	-0.593***	-1.050***	-1.136***	-0.167***	-0.250**	-0.323***	-0.290**
	(0.090)	(0.258)	(0.235)	(0.068)	(0.107)	(0.098)	(0.126)
	[-0.770; -0.416]	[-1.556; -0.544]	[-1.597; -0.674]	[-0.300; -0.033]	[-0.460; -0.040]	[-0.515; -0.130]	[-0.536; -0.044]
Number of teammates	0.052	0.344*	0.224	-0.026	0.091	-0.038	0.059
	(0.069)	(0.208)	(0.196)	(0.050)	(0.083)	(0.077)	(0.094)
	[-0.083; 0.187]	[-0.063; 0.751]	[-0.161; 0.609]	[-0.124; 0.071]	[-0.071; 0.253]	[-0.190; 0.114]	[-0.125; 0.242]
Game: Ottawa	-0.086	-1.248**	-0.243	-0.049	0.098	-0.351*	-0.069
	(0.155)	(0.755)	(0.460)	(0.172)	(0.217)	(0.177)	(0.205)
	[-0.389; 0.217]	[-2.728; 0.233]	[-1.144; 0.659]	[-0.386; 0.288]	[-0.326; 0.523]	[-0.698; -0.003]	[-0.471; 0.334]
Game: Sheffield	-0.180	-1.604*	-0.653	0.159	-0.006	-0.336	-0.010
	(0.259)	(0.644)	(0.464)	(0.136)	(0.400)	(0.310)	(0.360)
	[-0.688; 0.328]	[-2.866; -0.342]	[-1.563; 0.257]	[-0.106; 0.425]	[-0.791; 0.779]	[-0.944; 0.271]	[-0.715; 0.696]
Game: Cornell	0.276	-1.547**	-1.075	0.061	0.317	0.055	0.285
	(0.190)	(0.540)	(0.499)	(0.120)	(0.201)	(0.180)	(0.235)
	[-0.097; 0.649]	[-2.605; -0.488]	[-2.053; -0.097]	[-0.175; 0.296]	[-0.077; 0.711]	[-0.299; 0.409]	[-0.176; 0.746]
Game: Bogota	0.014	-1.074	-1.465*	0.016	0.071	-0.189	-0.170
	(0.175)	(1.000)	(1.021)	(0.136)	(0.350)	(0.178)	(0.354)
	[-0.328; 0.357]	[-3.034; 0.885]	[-3.467; 0.536]	[-0.250; 0.283]	[-0.615; 0.757]	[-0.537; 0.160]	[-0.864; 0.524]
Game: Tilburg	0.231	-2.185***	0.532	0.067	0.398	-0.076	0.230
	(0.173)	(0.697)	(0.720)	(0.112)	(0.168)	(0.175)	(0.184)
	[-0.109; 0.570]	[-3.551; -0.820]	[-0.878; 1.943]	[-0.151; 0.286]	[0.069; 0.728]	[-0.418; 0.266]	[-0.130; 0.591]
Game: Virtual Europe	0.004	-1.976***	-0.818	0.098	0.351*	0.092	0.245
	(0.161)	(0.623)	(0.587)	(0.110)	(0.174)	(0.128)	(0.250)
	[-0.311; 0.319]	[-3.197; -0.756]	[-1.970; 0.333]	[-0.117; 0.313]	[0.010; 0.691]	[-0.158; 0.343]	[-0.246; 0.735]
Game: Virtual North America	0.123	-1.545***	-0.583	0.088	0.327	-0.130	0.092
	(0.180)	(0.469)	(0.458)	(0.112)	(0.163)	(0.179)	(0.205)
	[-0.229; 0.476]	[-2.464; -0.627]	[-1.481; 0.315]	[-0.131; 0.308]	[0.009; 0.646]	[-0.481; 0.222]	[-0.309; 0.494]
Software: R	-0.169	0.915**	0.249	0.006	0.118	-0.014	0.096
	(0.154)	(0.621)	(0.513)	(0.121)	(0.183)	(0.123)	(0.184)
	[-0.471; 0.133]	[-0.303; 2.133]	[-0.756; 1.255]	[-0.232; 0.244]	[-0.240; 0.476]	[-0.256; 0.228]	[-0.265; 0.457]
Maximum academic level: Researcher	0.148	-1.548**	0.336	-0.015	-0.213	0.105	-0.053
	(0.180)	(0.628)	(1.559)	(0.091)	(0.195)	(0.159)	(0.219)
	[-0.205; 0.501]	[-2.780; -0.317]	[-2.720; 3.391]	[-0.194; 0.165]	[-0.595; 0.168]	[-0.207; 0.416]	[-0.483; 0.377]
Maximum academic level: Postdoc	0.110	0.078	0.275	0.032	-0.090	0.314**	0.196
	(0.177)	(0.337)	(0.318)	(0.082)	(0.145)	(0.146)	(0.172)
	[-0.237; 0.457]	[-0.583; 0.739]	[-0.348; 0.898]	[-0.129; 0.193]	[-0.375; 0.194]	[0.027; 0.601]	[-0.141; 0.533]
Maximum academic level: Professor	0.030	0.008	0.340	-0.043	-0.165	0.107	-0.008
	(0.140)	(0.245)	(0.262)	(0.090)	(0.128)	(0.145)	(0.147)
	[-0.244; 0.305]	[-0.472; 0.487]	[-0.174; 0.854]	[-0.220; 0.134]	[-0.415; 0.085]	[-0.177; 0.391]	[-0.296; 0.279]
Minimum academic level: Researcher	-0.140	-0.126	0.285	-0.012	0.184	0.078	0.269
	(0.091)	(0.345)	(0.519)	(0.066)	(0.116)	(0.108)	(0.137)
	[-0.319; 0.039]	[-0.802; 0.549]	[-0.732; 1.302]	[-0.140; 0.117]	[-0.044; 0.411]	[-0.134; 0.289]	[0.000; 0.537]
Minimum academic level: Postdoc	-0.080	0.089	-0.150	-0.127	-0.082	0.033	0.005
	(0.214)	(0.851)	(0.435)	(0.134)	(0.183)	(0.123)	(0.195)
	[-0.500; 0.339]	[-1.578; 1.756]	[-1.001; 0.702]	[-0.390; 0.136]	[-0.441; 0.278]	[-0.209; 0.275]	[-0.377; 0.388]
Minimum academic level: Professor	-0.094	0.697	0.535	0.001	-0.081	0.006	-0.076
	(0.150)	(0.431)	(0.414)	(0.051)	(0.187)	(0.141)	(0.223)
	[-0.388; 0.199]	[-0.147; 1.542]	[-0.277; 1.347]	[-0.099; 0.101]	[-0.446; 0.285]	[-0.270; 0.282]	[-0.513; 0.362]
Attendance: In-Person	-0.124	0.327	0.205	-0.012	-0.072	0.227**	0.196
	(0.120)	(0.369)	(0.245)	(0.085)	(0.121)	(0.120)	(0.127)
	[-0.358; 0.111]	[-0.397; 1.051]	[-0.274; 0.685]	[-0.178; 0.153]	[-0.309; 0.165]	[-0.009; 0.463]	[-0.053; 0.446]
Game: Ottawa × Software: R	-0.281	-1.622**	-0.306	-0.094	-0.066	-0.216	-0.352
	(0.285)	(0.741)	(0.647)	(0.250)	(0.307)	(0.324)	(0.353)
	[-0.840; 0.278]	[-3.075; -0.170]	[-1.573; 0.962]	[-0.585; 0.397]	[-0.668; 0.535]	[-0.851; 0.419]	[-1.044; 0.341]
Game: Sheffield × Software: R	0.322	-1.355	-0.751	-0.276	-0.260	0.178	-0.321
	(0.316)	(0.632)	(0.621)	(0.200)	(0.458)	(0.341)	(0.430)
	[-0.297; 0.942]	[-2.594; -0.116]	[-1.968; 0.465]	[-0.668; 0.116]	[-1.157; 0.637]	[-0.490; 0.846]	[-1.164; 0.522]
Game: Cornell × Software: R	-0.265	-1.309	0.424	-0.012	-0.115	-0.449	-0.557
	(0.244)	(0.614)	(0.742)	(0.151)	(0.239)	(0.227)	(0.303)
	[-0.743; 0.212]	[-2.513; -0.105]	[-1.031; 1.879]	[-0.308; 0.284]	[-0.584; 0.354]	[-0.894; -0.005]	[-1.150; 0.037]
Game: Bogota × Software: R	0.112	-1.867*	0.727	0.010	0.103	-0.054	0.044
	(0.319)	(1.100)	(1.032)	(0.165)	(0.384)	(0.259)	(0.420)
	[-0.514; 0.737]	[-4.023; 0.289]	[-1.296; 2.751]	[-0.314; 0.334]	[-0.649; 0.855]	[-0.562; 0.454]	[-0.778; 0.867]
Game: Tilburg × Software: R	-0.364	-0.235	-1.553*	-0.026	-0.697**	-0.250	-0.915**
	(0.252)	(0.802)	(0.898)	(0.134)	(0.294)	(0.310)	(0.270)
	[-0.858; 0.129]	[-1.807; 1.337]	[-3.312; 0.207]	[-0.290; 0.237]	[-1.274; -0.119]	[-0.857; 0.357]	[-1.445; -0.386]
Game: Virtual Europe × Software: R	0.234	-0.837	-0.597	-0.024	-0.424	-0.111	-0.287
	(0.222)	(0.774)	(0.745)	(0.140)	(0.287)	(0.189)	(0.341)
	[-0.201; 0.669]	[-2.355; 0.681]	[-2.057; 0.864]	[-0.298; 0.250]	[-0.986; 0.138]	[-0.480; 0.259]	[-0.955; 0.381]
Game: Virtual North America × Software: R	0.076	-1.145	-0.216	0.002	-0.312	0.110	-0.210
	(0.285)	(0.723)	(0.836)	(0.138)	(0.316)	(0.275)	(0.377)
	[-0.482; 0.634]	[-2.562; 0.272]	[-1.855; 1.423]	[-0.268; 0.272]	[-0.931; 0.307]	[-0.429; 0.649]	[-0.949; 0.528]
Controls	✓	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.738	0.951	0.728	0.942	0.786	0.816	0.680
Obs.	103	103	103	103	103	103	103

Note: Standard errors in parentheses, confidence intervals in brackets; human-only group omitted.
Controls include number of teammates; game-by-software fixed effects; maximum and minimum position skill fixed effects; attendance fixed effects.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S7: Causal relationship between treatment groups and minor and major errors (using shares)

	(1)	(2)
	Minor errors	Major errors
AI-Assisted	-0.521	-0.668
	(0.268)	(0.256)
	[-1.055; 0.014]	[-1.177; -0.158]
AI-Led	-0.968	-1.082
	(0.251)	(0.232)
	[-1.468; -0.467]	[-1.544; -0.620]
Controls	✓	✓
Mean dep. var	0.951	0.728
p-val (AI-Assisted vs. AI-Led)	0.029	0.044
Obs.	103	103

Note: Standard errors in parentheses, confidence intervals in brackets; human-only group omitted.

Controls include number of teammates; game-by-software fixed effects; maximum and minimum position skill fixed effects; attendance fixed effects.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S8: Causal relationship between treatment groups and reproducibility outcomes for different softwares

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Reproduction	Minor errors	Major errors	One good robustness	Two good robustness	Ran one robustness	Ran two robustness
AI-Assisted	-0.019	-0.858	-1.026	0.000	0.148	0.024	0.245
	(0.062)	(0.421)	(0.368)	(0.035)	(0.107)	(0.059)	(0.130)
AI-Led	-0.554	-1.440	-1.658	-0.141	-0.192	-0.365	-0.185
	(0.146)	(0.503)	(0.344)	(0.099)	(0.160)	(0.136)	(0.192)
R	-0.146	-0.327	-0.564	-0.012	0.051	-0.074	0.062
AI-Assisted × R	-0.022	0.806	0.707	-0.011	-0.322	-0.128	-0.511
	(0.138)	(0.587)	(0.486)	(0.051)	(0.204)	(0.125)	(0.226)
AI-Led × R	-0.072	0.659	0.897	-0.037	-0.106	0.052	-0.192
	(0.194)	(0.566)	(0.445)	(0.126)	(0.210)	(0.184)	(0.247)
Controls	✓	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.738	0.951	0.728	0.942	0.786	0.816	0.680
Obs.	103	103	103	103	103	103	103

Note: Standard errors in parentheses, confidence intervals in brackets; human-only group omitted; Stata papers omitted.

Controls include number of teammates; game and software fixed effects; maximum and minimum position skill fixed effects; attendance fixed effects.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S9: AI-Assisted and AI-Led Metrics by Experience Level

Variable	AI-Assisted high experience (n=10)	AI-Assisted low/medium experience (n=24)	AI-Led high experience (n=9)	AI-Led low/medium experience (n=26)	AI-Assisted High vs Low	AI-Led High vs Low
Reproduction	1.000 (0.000)	0.875 (0.338)	0.444 (0.527)	0.346 (0.485)	0.125 [0.083]	0.098 [0.631]
Minutes to reproduction	89.2 (85.5)	81.4 (60.6)	178.2 (69.2)	180.3 (72.3)	7.8 [0.800]	-2.1 [0.962]
Number of minor errors	1.400 (2.066)	0.792 (1.141)	0.556 (1.130)	0.500 (0.860)	0.608 [0.399]	0.056 [0.895]
Minutes to first minor error	104.7 (59.4)	66.0 (22.5)	108.0 (15.7)	180.9 (116.2)	38.7 [0.143]	-72.9 [0.123]
Number of major errors	0.900 (1.101)	0.542 (0.884)	0.444 (0.726)	0.154 (0.368)	0.358 [0.376]	0.291 [0.278]
Minutes to first major error	102.4 (29.5)	158.4 (58.2)	117.0 (49.6)	255.2 (80.8)	-56.0 [0.034]	-138.2 [0.039]
At least one appropriate robustness check	1.000 (0.000)	1.000 (0.000)	0.778 (0.441)	0.846 (0.368)	0.000 [-]	-0.068 [0.684]
At least two appropriate robustness checks	0.900 (0.316)	0.833 (0.381)	0.667 (0.500)	0.615 (0.496)	0.067 [0.604]	0.051 [0.794]
Ran at least one appropriate robustness check	1.000 (0.000)	0.958 (0.204)	0.556 (0.527)	0.577 (0.504)	0.042 [0.328]	-0.021 [0.917]
Ran at least two appropriate robustness check	0.900 (0.316)	0.792 (0.415)	0.444 (0.527)	0.462 (0.508)	0.108 [0.417]	-0.017 [0.934]

Table S10: Comparison of Key Metrics by Prompt Levels within AI-Assisted Group

Variable	Above median (n=17)	Below/equal to median (n=18)	Difference
Reproduction	1.000 (0.000)	0.833 (0.383)	0.167 [0.083]
Minutes to reproduction	102.6 (95.8)	82.7 (73.8)	20.0 [0.511]
Number of minor errors	0.824 (1.131)	1.056 (1.731)	-0.232 [0.640]
Minutes to first minor error	104.8 (54.1)	61.7 (21.1)	43.1 [0.064]
Number of major errors	0.706 (1.105)	0.556 (0.784)	0.150 [0.648]
Minutes to first major error	155.3 (68.5)	121.6 (37.8)	33.7 [0.282]
At least one appropriate robustness check	1.000 (0.000)	1.000 (0.000)	0.000 [-]
At least two appropriate robustness checks	0.765 (0.437)	0.944 (0.236)	-0.180 [0.146]
Ran at least one appropriate robustness check	0.941 (0.243)	0.944 (0.236)	-0.003 [0.968]
Ran at least two appropriate robustness check	0.706 (0.470)	0.889 (0.323)	-0.183 [0.192]

Note: Columns 2–3 present means and standard errors in parentheses for individual groups (Human-only, AI-Assisted, and AI-Led); column 4 shows mean differences and p-values in brackets for the indicated group comparison. Groups are defined by the median number of prompts (22) in the AI-Assisted sample.

Table S11: Ex-post power calculation for Table 2 analyses

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Reproduction	Minor errors	Major errors	One good robustness	Two good robustness	Ran one robustness	Ran two robustness
AI-Assisted power	0.061	0.603	0.838	0.050	0.059	0.077	0.056
AI-Led power	1.000	0.995	0.999	0.860	0.785	0.984	0.822

Table S12: Restricted-mean time without success (minutes) and contrasts

Variable	Human only	AI-Assisted	AI-Led	Human vs AI-Assisted	Human vs AI-Led	AI-Assisted vs AI-Led
	102	121	331	-18.8	-228	-209
Minutes to reproduction	(15.5)	(20.6)	(20.8)	(0.465)	(<0.001)	(<0.001)
	217	256	339	-39	-122	-82.8
Minutes to first minor error	(28.7)	(29)	(22.4)	(0.339)	(<0.001)	(0.024)
	258	307	375	-49.1	-117	-67.8
Minutes to first major error	(25.4)	(24)	(16.6)	(0.160)	(<0.001)	(0.020)

Note: Each cell shows the mean RMST in minutes with the standard error below in parentheses. Contrast columns present the mean difference with its two-sided p*-value below. Times are right-censored at 420 minutes (7 hours).*