

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

Abel Brodeur, **David Valenta (dvalenta@uottawa.ca)**, Alexandru Marcoci, Juan P. Aparicio, Derek Mikola, Bruno Barbarioli, Rohan Alexander, Lachlan Deer, Tom Stafford, Lars Vilhuber, Gunther Bensch, et al.

EEA Congress 2025, Bordeaux  
August 25, 2025

# Introduction

- **(Computational) Reproducibility** = Ability obtain the same results as a prior study using the same data
- **Robustness** = Assessing whether changes to the model specification (e.g., changes to control variables or weighting) or data lead to the same conclusion
- Reproducibility and robustness is critical for empirical social-sciences research and enable confidence in published studies (Gertler et al. 2018)

# Our experiment

- Teams of researchers assessing:
  - reproducibility
  - coding errors
  - robustness
- Randomized into 3 treatment arms:
  - **Human-only**: no use of ChatGPT or other AI
  - **AI-assisted**: may use ChatGPT to assist the replication
  - **AI-led**: may not read the paper and must only follow instruction by ChatGPT
- Over 270 participating researchers in 136 teams
- 8 events between February 2024 and April 2025

# Reproducibility and robustness remain an issue

- Brodeur et al. (2024)
  - Assessed 110 recent papers from leading economics and political science journals
  - **Reproducibility: 85%**
  - **Robustness: 70%**
  - **Coding errors: 25%**
- Campbell et al. (2024)
  - Assessed robustness of 17 non-experimental papers from AER
  - Half of key results of remained statistically significant at 5% level
  - **Surveyed economists overestimated robustness of the results**

# Reproductions are costly

- Reproducibility is expensive and requires technical skills
  - Reproducing all top-10 econ journal articles estimated to cost about \$330 per article (Colliard et al. 2023)
  - AEA data editor budget about \$750 per published paper (Brodeur et al. 2025)
- Additional effort to check for coding errors and assess robustness
  - Replication games: about 3 team members work 7 hours during the event, and usually spend more time before and after

## Goal of this study

- Can AI help in assessing reproducibility and robustness?
- Two ways AI could help:
  - Aid researchers when reproducing studies
  - Directly reproduce the studies without researchers' involvement

# Background

- AI/LLMs can make workers more productive (Brynjolfsson et al. 2025, Dell'Acqua et al. 2023), including software developers (Peng et al. 2023)
- People may over-rely on AI (Buçinca et al. 2021) and AI can worsen performance of software developers (Becker et al. 2025)
- AI can be used to detect errors in code (Wadhwa et al. 2024, Zhang 2024)
- And for replicating scientific studies
  - Starace et al. (2025) found that while AI can reproduce (parts of) studies in CS, it does not yet outperform PhD students

# Study Setting: AI Replication Games

- 7-hour long event
- Teams of 2–4 researchers
  - Graduate students, postdoctoral fellows, professors, and researchers
- Task
  - Reproduce a few pre-defined numerical result
  - Detect coding errors and data irregularities
  - Suggest two robustness checks and run those checks if possible
- In-person or online

# Sample

## Study 1

- 103 teams
- 271 researchers
- 7 events
- February-November 2024
- 12 papers
- No reasoning models

## Study 2

- 34 teams
- 95 researchers (all part of Study 1)
- 1 event (virtual)
- April 2025
- 6 papers
- Models with reasoning available

## Studies Reproduced

- From leading behavioral, economics, political science, or psychology journals
- Previously reproduced
- Known to contain coding errors
- Code can be executed
- Replication reports not yet publicly released
- The teams did not know the study they will reproduce before the event
- The paper (PDF), screenshots of relevant tables/figures and replication package provided to the teams

# Treatment Arms

- Researchers randomly assigned to one of three treatment conditions:
  - Human-only
  - AI-Assisted
  - AI-Lead
- Teams formed based on
  - Software preference (R/Stata)
  - Mode of participation

# 1. Human-only

- **No Usage of ChatGPT or any other AI**

## 2. AI-Assisted

- **Free to use ChatGPT at their discretion**
- Provided ChatGPT team subscription
  - Access to newest models
  - Higher usage limits compared to the free version
- Mandatory ChatGPT training before the games
- All AI-assisted teams used ChatGPT at least a bit

### 3. AI-Lead

- **Must only follow instructions by ChatGPT**
- **Not allowed to read the paper**
- **Only allowed to interact with the paper through ChatGPT**
- First try to reproduce the results using ChatGPT's Python module
- Allowed to run code on their computer but only as instructed by ChatGPT
- Provided ChatGPT team subscription
- Mandatory ChatGPT training before the games

### 3. AI-Lead


- Why this group?
  - Test the capabilities of AI when operating with minimum human oversight
  - Simulate environment where AI has all the ability needed to conduct reproduction (e.g. execute code R or Stata)
  - Without fully developing automated system
  - Learn which prompting strategies work


ChatGPT x +

https://chatgpt.com/?model=gpt-4o

ChatGPT 4o Temporary

## How can I help, David?

 **Paper.pdf**  
PDF

 **replication data.zip**  
Zip Archive

Take a look at this paper? What is in table 3? Can you find the do file and corresponding data that run model 2 of table 3 and run that regression using Python?

+ Search Deep research Create image ...

ChatGPT can make mistakes. OpenAI doesn't use Institute for Replication test workspace data to train its models.

Table 3 Model 2 Replication x +

https://chatgpt.com/c/682e523e-3060-8012-be14-63b429218e5e

ChatGPT 4o

Share

Analyzed ^

```
python

import statsmodels.api as sm
import statsmodels.formula.api as smf

# Filter for females only as done in the original Stata command
df_model2 = df[df['female'] == 1].copy()

# Drop missing values in the outcome and key independent variables
df_model2 = df_model2.dropna(subset=['worked', 'garmentvill', 'dgarmentyears10to23', 'dgarmentyears24to39'])

# Run the regression: Model 2 of Table 3
model2 = smf.ols(
    formula='worked ~ garmentvill + dgarmentyears10to23 + dgarmentyears24to39 + C(yearofbirth)',
    data=df_model2
).fit(cov_type='cluster', cov_kwds={'groups': df_model2['sibgroup']})

model2.summary()
```

Always show details  Copy

Ask anything

+ Search Deep research Create image

ChatGPT can make mistakes. OpenAI doesn't use Institute for Replication test workspace data to train its models.

Table 3 Model 2 Replication

https://chatgpt.com/c/682e523e-3060-8012-be14-63b429218e5e

ChatGPT 4o

```

C(yearofbirth)[T.108.0]    -1.0455    0.039    -26.075    0.000    -1.122    -0.969
C(yearofbirth)[T.109.0]    -1.0455    0.039    -26.675    0.000    -1.122    -0.969
garmentvill                0.0455     0.039     1.161     0.246    -0.031    0.122
dgarmentyears10to23        0.1266     0.054     2.357     0.018     0.021    0.232
dgarmentyears24to39        0.0677     0.059     1.152     0.249    -0.047    0.183
-----
Omnibus:                   103.966   Durbin-Watson:           1.683
Prob(Omnibus):              0.000   Jarque-Bera (JB):       135.535
Skew:                        0.928   Prob(JB):                 3.71e-30
Kurtosis:                    2.683   Cond. No.                  321.
-----
Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
****

```

### Table 3, Model 2 — Analysis Summary and Replication in Python

#### Content of Table 3

From the *paper on page 8*, Table 3 presents regression results on the probability that a girl has ever worked out.

Ask anything

+ Search Deep research Create image

ChatGPT can make mistakes. OpenAI doesn't use Institute for Replication test workspace data to train its models.

# Results: Computational Reproducibility

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

| Variable                                | Human-Only       | AI-Assisted      | AI-Led           | Human-Only<br>vs<br>AI-Assisted | Human-Only<br>vs<br>AI-Led | AI-Assisted<br>vs<br>AI-Led |
|---|------------------|------------------|------------------|---------------------------------|----------------------------|-----------------------------|
| Reproduction                            | 0.956<br>(0.208) | 0.935<br>(0.250) | 0.478<br>(0.505) | 0.021<br>(0.668)                | 0.477<br>( $<0.001$ )      | 0.457<br>( $<0.001$ )       |
| Minutes to reproduction                 | 83.9<br>(50.6)   | 86.1<br>(78.1)   | 134.8<br>(86.1)  | -2.2<br>(0.877)                 | -51.0<br>(0.004)           | -48.7<br>(0.025)            |
| Number of minor errors                  | 1.222<br>(1.521) | 0.891<br>(1.303) | 0.478<br>(0.863) | 0.331<br>(0.268)                | 0.744<br>(0.005)           | 0.413<br>(0.076)            |
| Minutes to first minor error            | 104.5<br>(73.5)  | 97.8<br>(67.2)   | 141.0<br>(99.1)  | 6.8<br>(0.733)                  | -36.5<br>(0.185)           | -43.2<br>(0.118)            |
| Number of major errors                  | 1.222<br>(1.363) | 0.652<br>(0.849) | 0.304<br>(0.591) | 0.570<br>(0.018)                | 0.918<br>( $<0.001$ )      | 0.348<br>(0.025)            |
| Minutes to first major error            | 146.0<br>(86.9)  | 134.8<br>(69.4)  | 148.1<br>(101.3) | 11.2<br>(0.626)                 | -2.1<br>(0.950)            | -13.3<br>(0.661)            |
| At least one good robustness check      | 1.000<br>(0.000) | 1.000<br>(0.000) | 0.870<br>(0.341) | -<br>(-)                        | 0.130<br>(0.012)           | 0.130<br>(0.011)            |
| At least two good robustness checks     | 0.911<br>(0.288) | 0.891<br>(0.315) | 0.696<br>(0.465) | 0.020<br>(0.755)                | 0.215<br>(0.010)           | 0.196<br>(0.020)            |
| Ran at least one good robustness check  | 0.933<br>(0.252) | 0.957<br>(0.206) | 0.674<br>(0.474) | -0.023<br>(0.632)               | 0.259<br>(0.002)           | 0.283<br>( $<0.001$ )       |
| Ran at least two good robustness checks | 0.822<br>(0.387) | 0.848<br>(0.363) | 0.543<br>(0.504) | -0.026<br>(0.745)               | 0.279<br>(0.004)           | 0.304<br>(0.001)            |

Note: Standard errors in parentheses for individual branches (Human-only, AI-Assisted, and AI-Led); p-values in parentheses for branch comparisons (Human-Only vs AI-Assisted, Human-Only vs AI-Led, and AI-Assisted vs AI-Led).  
 Good robustness check: must be clear, feasible, not done by authors, and focused on validity of the empirical strategy.

Major coding error: have potential to impact the claims tested.

Minor coding error: no impact (e.g., missing packages, incorrect paths).

# Results: Coding Errors

Table 2: Comparison of Human, AI-Assisted, and AI-Led Teams

| Variable                                | Human-Only       | AI-Assisted      | AI-Led           | Human-Only<br>vs<br>AI-Assisted | Human-Only<br>vs<br>AI-Led | AI-Assisted<br>vs<br>AI-Led |
|---|------------------|------------------|------------------|---------------------------------|----------------------------|-----------------------------|
| Reproduction                            | 0.956<br>(0.208) | 0.935<br>(0.250) | 0.478<br>(0.505) | 0.021<br>(0.668)                | 0.477<br>( $<0.001$ )      | 0.457<br>( $<0.001$ )       |
| Minutes to reproduction                 | 83.9<br>(50.6)   | 86.1<br>(78.1)   | 134.8<br>(86.1)  | -2.2<br>(0.877)                 | -51.0<br>(0.004)           | -48.7<br>(0.025)            |
| Number of minor errors                  | 1.222<br>(1.521) | 0.891<br>(1.303) | 0.478<br>(0.863) | 0.331<br>(0.268)                | 0.744<br>(0.005)           | 0.413<br>(0.076)            |
| Minutes to first minor error            | 104.5<br>(73.5)  | 97.8<br>(67.2)   | 141.0<br>(99.1)  | 6.8<br>(0.733)                  | -36.5<br>(0.185)           | -43.2<br>(0.118)            |
| Number of major errors                  | 1.222<br>(1.363) | 0.652<br>(0.849) | 0.304<br>(0.591) | 0.570<br>(0.018)                | 0.918<br>( $<0.001$ )      | 0.348<br>(0.025)            |
| Minutes to first major error            | 146.0<br>(86.9)  | 134.8<br>(69.4)  | 148.1<br>(101.3) | 11.2<br>(0.626)                 | -2.1<br>(0.950)            | -13.3<br>(0.661)            |
| At least one good robustness check      | 1.000<br>(0.000) | 1.000<br>(0.000) | 0.870<br>(0.341) | -<br>(-)                        | 0.130<br>(0.012)           | 0.130<br>(0.011)            |
| At least two good robustness checks     | 0.911<br>(0.288) | 0.891<br>(0.315) | 0.696<br>(0.465) | 0.020<br>(0.755)                | 0.215<br>(0.010)           | 0.196<br>(0.020)            |
| Ran at least one good robustness check  | 0.933<br>(0.252) | 0.957<br>(0.206) | 0.674<br>(0.474) | -0.023<br>(0.632)               | 0.259<br>(0.002)           | 0.283<br>( $<0.001$ )       |
| Ran at least two good robustness checks | 0.822<br>(0.387) | 0.848<br>(0.363) | 0.543<br>(0.504) | -0.026<br>(0.745)               | 0.279<br>(0.004)           | 0.304<br>(0.001)            |

Note: Standard errors in parentheses for individual branches (Human-only, AI-Assisted, and AI-Led); p-values in parentheses for branch comparisons (Human-Only vs AI-Assisted, Human-Only vs AI-Led, and AI-Assisted vs AI-Led).  
 Good robustness check: must be clear, feasible, not done by authors, and focused on validity of the empirical strategy.

Major coding error: have potential to impact the claims tested.  
 Minor coding error: no impact (e.g., missing packages, incorrect paths).

# Results: Robustness Checks

Table 3: Comparison of Human, AI-Assisted, and AI-Led Teams

| Variable                                | Human-Only       | AI-Assisted      | AI-Led           | Human-Only<br>vs<br>AI-Assisted | Human-Only<br>vs<br>AI-Led | AI-Assisted<br>vs<br>AI-Led |
|---|------------------|------------------|------------------|---------------------------------|----------------------------|-----------------------------|
| Reproduction                            | 0.956<br>(0.208) | 0.935<br>(0.250) | 0.478<br>(0.505) | 0.021<br>(0.668)                | 0.477<br>( $<0.001$ )      | 0.457<br>( $<0.001$ )       |
| Minutes to reproduction                 | 83.9<br>(50.6)   | 86.1<br>(78.1)   | 134.8<br>(86.1)  | -2.2<br>(0.877)                 | -51.0<br>(0.004)           | -48.7<br>(0.025)            |
| Number of minor errors                  | 1.222<br>(1.521) | 0.891<br>(1.303) | 0.478<br>(0.863) | 0.331<br>(0.268)                | 0.744<br>(0.005)           | 0.413<br>(0.076)            |
| Minutes to first minor error            | 104.5<br>(73.5)  | 97.8<br>(67.2)   | 141.0<br>(99.1)  | 6.8<br>(0.733)                  | -36.5<br>(0.185)           | -43.2<br>(0.118)            |
| Number of major errors                  | 1.222<br>(1.363) | 0.652<br>(0.849) | 0.304<br>(0.591) | 0.570<br>(0.018)                | 0.918<br>( $<0.001$ )      | 0.348<br>(0.025)            |
| Minutes to first major error            | 146.0<br>(86.9)  | 134.8<br>(69.4)  | 148.1<br>(101.3) | 11.2<br>(0.626)                 | -2.1<br>(0.950)            | -13.3<br>(0.661)            |
| At least one good robustness check      | 1.000<br>(0.000) | 1.000<br>(0.000) | 0.870<br>(0.341) | -<br>(-)                        | 0.130<br>(0.012)           | 0.130<br>(0.011)            |
| At least two good robustness checks     | 0.911<br>(0.288) | 0.891<br>(0.315) | 0.696<br>(0.465) | 0.020<br>(0.755)                | 0.215<br>(0.010)           | 0.196<br>(0.020)            |
| Ran at least one good robustness check  | 0.933<br>(0.252) | 0.957<br>(0.206) | 0.674<br>(0.474) | -0.023<br>(0.632)               | 0.259<br>(0.002)           | 0.283<br>( $<0.001$ )       |
| Ran at least two good robustness checks | 0.822<br>(0.387) | 0.848<br>(0.363) | 0.543<br>(0.504) | -0.026<br>(0.745)               | 0.279<br>(0.004)           | 0.304<br>(0.001)            |

Note: Standard errors in parentheses for individual branches (Human-only, AI-Assisted, and AI-Led); p-values in parentheses for branch comparisons (Human-Only vs AI-Assisted, Human-Only vs AI-Led, and AI-Assisted vs AI-Led).

**Good robustness check: must be clear, feasible, not done by authors, and focused on validity of the empirical strategy.**

Major coding error: have potential to impact the claims tested.

Minor coding error: no impact (e.g., missing packages, incorrect paths).

# Model with Control Variables

Table 4: Results from OLS regressions predicting reproduction outcomes

| Panel B: Study 1 and 2 combined |   |   |   |  |  |   |  |
|---------------------------------|---|---|---|--|--|---|--|
|                                 | (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                     | -0.027<br>( 0.049)<br>[-0.124; 0.069]     | -0.345*<br>( 0.204)<br>[-0.750; 0.060]    | -0.510**<br>( 0.211)<br>[-0.928; -0.092]  | 0.001<br>( 0.020)<br>[-0.038; 0.040]     | -0.008<br>( 0.075)<br>[-0.155; 0.140]    | 0.008<br>( 0.050)<br>[-0.091; 0.108]      | 0.026<br>( 0.085)<br>[-0.142; 0.193]     |
| AI-Led                          | -0.493***<br>( 0.077)<br>[-0.646; -0.340] | -0.823***<br>( 0.203)<br>[-1.225; -0.422] | -0.869***<br>( 0.201)<br>[-1.267; -0.470] | -0.129**<br>( 0.051)<br>[-0.229; -0.028] | -0.203**<br>( 0.082)<br>[-0.367; -0.040] | -0.228***<br>( 0.079)<br>[-0.385; -0.071] | -0.251**<br>( 0.104)<br>[-0.458; -0.045] |
| Controls                        | ✓   | ✓   | ✓   | ✓  | ✓  | ✓   | ✓  |
| Mean of dep. var                | 0.788                                     | 0.861                                     | 0.723                                     | 0.956                                    | 0.832                                    | 0.854                                     | 0.737                                    |
| p-val (AI-Assisted vs. AI-Led)  | 0.000                                     | 0.001                                     | 0.025                                     | 0.011                                    | 0.012                                    | 0.001                                     | 0.002                                    |
| Obs.                            | 137                                       | 137                                       | 137                                       | 137                                      | 137                                      | 137                                       | 137                                      |

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

Controls include number of teammates; game–software, baseline skill, and mode of attendance fixed effects.

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

# Do AI-assisted and AI-led Teams Improve Over Time?

Table 5: Results of Study 1 compared to Study 2

| Panel A: Study 1               |   |   |  |  |  |   |  |
|--------------------------------|---|---|--|--|--|---|--|
|                                | (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<br>(0.063)                       | -0.487*<br>(0.270)                      | -0.646**<br>(0.254)                      | -0.009<br>(0.027)                      | -0.014<br>(0.103)                      | -0.032<br>(0.061)                       | -0.009<br>(0.113)                      |
| AI-Led                         | [-0.144; 0.107]<br>-0.593***<br>(0.090) | [-1.025; 0.051]<br>-1.050***<br>(0.258) | [-1.153; -0.139]<br>-1.136***<br>(0.235) | [-0.063; 0.046]<br>-0.167**<br>(0.068) | [-0.220; 0.191]<br>-0.250**<br>(0.107) | [-0.155; 0.090]<br>-0.323***<br>(0.098) | [-0.233; 0.216]<br>-0.290**<br>(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 vs. AI-Led) | 0.000                                   | 0.022                                   | 0.018                                    | 0.023                                  | 0.036                                  | 0.004                                   | 0.019                                  |
| Obs.                           | 103                                     | 103                                     | 103                                      | 103                                    | 103                                    | 103                                     | 103                                    |
| Panel B: Study 2               |   |   |  |  |  |   |  |
|                                | (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                    | -0.017<br>(0.064)                       | 0.283<br>(0.191)                        | 0.012<br>(0.281)                         | 0.000<br>(.)                           | -0.046<br>(0.043)                      | 0.124<br>(0.130)                        | 0.034<br>(0.151)                       |
| AI-Led                         | [-0.149; 0.114]<br>-0.190<br>(0.134)    | [-0.113; 0.679]<br>-0.085<br>(0.229)    | [-0.570; 0.594]<br>-0.170<br>(0.354)     | [0.000; 0.000]<br>0.000<br>(.)         | [-0.136; 0.043]<br>-0.065<br>(0.084)   | [-0.145; 0.393]<br>0.111<br>(0.117)     | [-0.279; 0.347]<br>-0.039<br>(0.170)   |
|                                | [-0.467; 0.086]                         | [-0.558; 0.389]                         | [-0.904; 0.563]                          | [0.000; 0.000]                         | [-0.240; 0.109]                        | [-0.132; 0.353]                         | [-0.390; 0.313]                        |
| Controls                       | ✓                                       | ✓                                       | ✓  | ✓                                      | ✓                                      | ✓                                       | ✓                                      |
| Mean of dep. var               | 0.939                                   | 0.576                                   | 0.727                                    | 1.000                                  | 0.970                                  | 0.970                                   | 0.909                                  |
| p-val (AI-Assisted vs. AI-Led) | 0.245                                   | 0.119                                   | 0.572                                    | .                                      | 0.718                                  | 0.685                                   | 0.395                                  |
| Obs.                           | 33                                      | 33                                      | 33                                       | 33                                     | 33                                     | 33                                      | 33                                     |

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

Controls include number of teammates; game-software, baseline skill, and mode of attendance fixed effects.

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

# Does experience with ChatGPT matter? Study I Only

Table 6: AI-Assisted and AI-Led Metrics by Experience Level (Study I only)

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

Note: Group columns display mean (SD). The two right-most columns show High – Low/Medium differences within each branch, with two-sided Welch p-values in parentheses. Study I teams only.

# Limitations

- We only test ChatGPT
  - Chosen for having most suitable environment (upload files, execute code)
  - Might not be the best AI models
- Out of the box AI: No custom training focused on reproducibility
- Limited time for reproduction of 7 hours
  - Reproductions often take longer
- Focus on limited number of social sciences papers
- Small time frame in fast moving field

# Conclusion

- AI-Assisted teams do not outperform human-only teams
  - But the usage of AI is associated with a cost
  - Assistance by AI does not currently seem to be a solution to reducing cost
- Some AI-led teams finished the tasks successfully
  - AI-driven reproduction has a potential to make an impact
  - At least for a subset of studies or subset of the tasks
  - Could be used as 1st check
- AI is not reliable and significant human component is still required

# Thank you!

- Contact me at [dvalenta@uottawa.ca](mailto:dvalenta@uottawa.ca)
- Preprint of our paper is available at:  
<https://econpapers.repec.org/paper/zbwi4rdps/195.htm>

# References I

- Becker, J., Rush, N., Barnes, E. and Rein, D.: 2025, Measuring the impact of early-2025 ai on experienced open-source developer productivity, *arXiv preprint arXiv:2507.09089* .
- Brodeur, A., Mikola, D. and Cook, N.: 2024, Mass reproducibility and replicability: A new hope.
- Brodeur, A., Sung, S. Y., Miguel, E., Vilhuber, L. and de la Guardia, F. H.: 2025, Assessing reproducibility in economics using standardized crowd-sourced analysis, *Technical report*, National Bureau of Economic Research.
- Brynjolfsson, E., Li, D. and Raymond, L.: 2025, Generative ai at work, *The Quarterly Journal of Economics* p. qjae044.
- Buçinca, Z., Malaya, M. B. and Gajos, K. Z.: 2021, 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.
- Campbell, D., Brodeur, A., Dreber, A., Johannesson, M., Kopecky, J., Lusher, L. and Tsoy, N.: 2024, The robustness reproducibility of the american economic review.
- Colliard, J.-E., Hurlin, C. and Pérignon, C.: 2023, The economics of computational reproducibility, *HEC Paris Research Paper No. FIN-2019-1345* .

## References II

- Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F. and Lakhani, K. R.: 2023, Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality, *Harvard Business School Technology & Operations Mgt. Unit Working Paper* (24-013).
- Gertler, P., Galiani, S. and Romero, M.: 2018, How to make replication the norm, *Nature* **554**(7693), 417–419.
- Peng, S., Kalliamvakou, E., Cihon, P. and Demirer, M.: 2023, The impact of ai on developer productivity: Evidence from github copilot, *arXiv preprint arXiv:2302.06590* .
- Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J. S., Maksin, L., Dias, R., Mays, E., Kinsella, B., Thompson, W. et al.: 2025, Paperbench: Evaluating ai's ability to replicate ai research, *arXiv preprint arXiv:2504.01848* .
- Wadhwa, N., Pradhan, J., Sonwane, A., Sahu, S. P., Natarajan, N., Kanade, A., Parthasarathy, S. and Rajamani, S.: 2024, Core: Resolving code quality issues using llms, *Proceedings of the ACM on Software Engineering* **1**(FSE), 789–811.
- Zhang, Y.: 2024, Detecting code comment inconsistencies using llm and program analysis, *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, pp. 683–685.

# Table 7: Balance of Team-Level Characteristics by Branch

| Variable                           | Branches         |                  |                   | Differences               |                      |                       |
|------------------------------------|------------------|------------------|-------------------|---------------------------|----------------------|-----------------------|
|                                    | Human-Only       | AI-Assisted      | AI-Led            | Human-Only vs AI-Assisted | Human-Only vs AI-Led | AI-Assisted vs AI-Led |
| Number of teammates                | 2.622<br>(0.490) | 2.587<br>(0.686) | 2.783<br>(0.554)  | 0.035<br>(0.779)          | -0.160<br>(0.148)    | -0.196<br>(0.136)     |
| Minimum academic level: Professor  | 0.089<br>(0.288) | 0.065<br>(0.250) | 0.109<br>(0.315)  | 0.024<br>(0.676)          | -0.020<br>(0.755)    | -0.043<br>(0.465)     |
| Minimum academic level: Postdoc    | 0.022<br>(0.149) | 0.152<br>(0.363) | 0.065<br>(0.250)  | -0.130<br>(0.029)         | -0.043<br>(0.323)    | 0.087<br>(0.184)      |
| Minimum academic level: Researcher | 0.222<br>(0.420) | 0.174<br>(0.383) | 0.065<br>(0.250)  | 0.048<br>(0.568)          | 0.157<br>(0.033)     | 0.109<br>(0.110)      |
| Minimum academic level: Student    | 0.667<br>(0.477) | 0.609<br>(0.493) | 0.761<br>(0.431)  | 0.058<br>(0.570)          | -0.094<br>(0.325)    | -0.152<br>(0.119)     |
| Maximum academic level: Professor  | 0.578<br>(0.499) | 0.565<br>(0.501) | 0.674<br>(0.474)  | 0.013<br>(0.905)          | -0.096<br>(0.349)    | -0.109<br>(0.288)     |
| Maximum academic level: Postdoc    | 0.156<br>(0.367) | 0.217<br>(0.417) | 0.196<br>(0.401)  | -0.062<br>(0.455)         | -0.040<br>(0.620)    | 0.022<br>(0.799)      |
| Maximum academic level: Researcher | 0.111<br>(0.318) | 0.087<br>(0.285) | 0.000<br>(0.000)  | 0.024<br>(0.703)          | 0.111<br>(0.020)     | 0.087<br>(0.041)      |
| Maximum academic level: Student    | 0.156<br>(0.367) | 0.130<br>(0.341) | 0.130<br>(0.341)  | 0.025<br>(0.736)          | 0.025<br>(0.736)     | 0.000<br>(1.000)      |
| Average years of coding experience | 9.915<br>(4.735) | 8.895<br>(3.612) | 10.212<br>(3.478) | 1.020<br>(0.250)          | -0.297<br>(0.733)    | -1.317<br>(0.078)     |
| Min ChatGPT level: Never           | 0.222<br>(0.420) | 0.130<br>(0.341) | 0.217<br>(0.417)  | 0.092<br>(0.255)          | 0.005<br>(0.956)     | -0.087<br>(0.276)     |
| Min ChatGPT level: Beginner        | 0.422<br>(0.499) | 0.565<br>(0.501) | 0.522<br>(0.505)  | -0.143<br>(0.176)         | -0.100<br>(0.347)    | 0.043<br>(0.680)      |
| Min ChatGPT level: Intermediate    | 0.267<br>(0.447) | 0.283<br>(0.455) | 0.239<br>(0.431)  | -0.016<br>(0.867)         | 0.028<br>(0.766)     | 0.043<br>(0.639)      |
| Min ChatGPT level: Advanced        | 0.089<br>(0.288) | 0.000<br>(0.000) | 0.022<br>(0.147)  | 0.089<br>(0.039)          | 0.067<br>(0.163)     | -0.022<br>(0.320)     |
| Max ChatGPT level: Never           | 0.000<br>(0.000) | 0.022<br>(0.147) | 0.022<br>(0.147)  | -0.022<br>(0.325)         | -0.022<br>(0.325)    | 0.000<br>(1.000)      |
| Max ChatGPT level: Beginner        | 0.111<br>(0.318) | 0.109<br>(0.315) | 0.087<br>(0.285)  | 0.002<br>(0.971)          | 0.024<br>(0.703)     | 0.022<br>(0.729)      |
| Max ChatGPT level: Intermediate    | 0.511<br>(0.506) | 0.500<br>(0.506) | 0.565<br>(0.501)  | 0.011<br>(0.917)          | -0.054<br>(0.609)    | -0.065<br>(0.536)     |
| Max ChatGPT level: Advanced        | 0.378<br>(0.490) | 0.348<br>(0.482) | 0.326<br>(0.474)  | 0.030<br>(0.769)          | 0.052<br>(0.610)     | 0.022<br>(0.828)      |

Note: Each cell in the first three columns shows the mean (top) and standard deviation (bottom) of the characteristic for the indicated branch. Cells in the last three columns show the mean difference between branches with the corresponding two-sided t-test p-value in parentheses.

# Reproduction Rate by Event

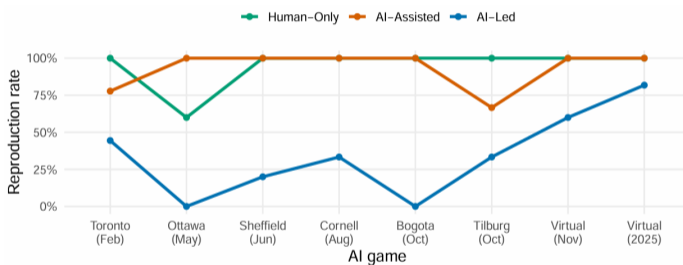
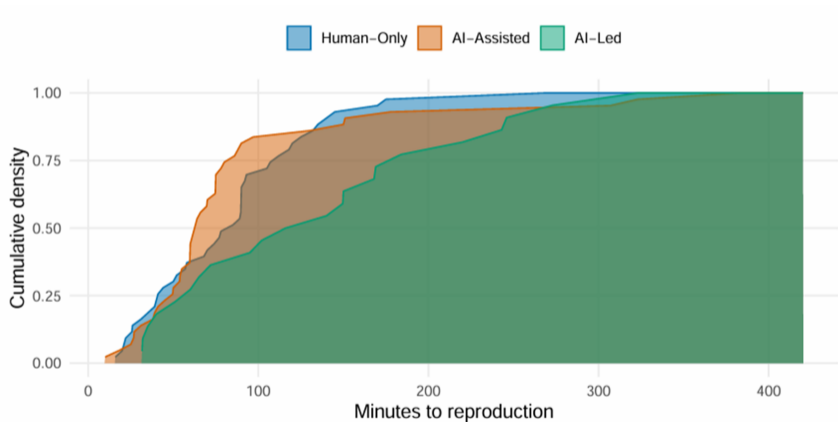


Figure 1: Computational reproducibility rate across treatment groups for each event.

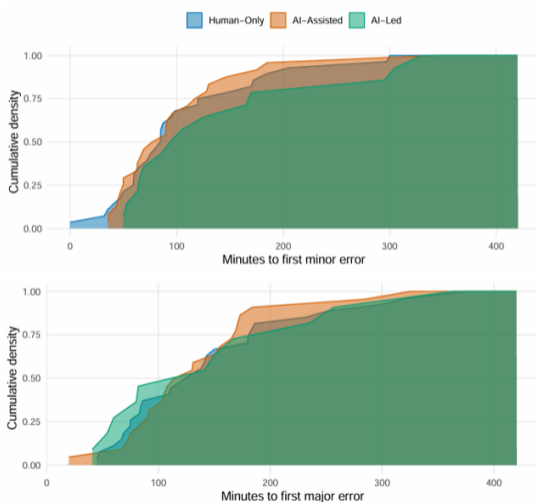
# Computational Reproducibility - Time to Reproduction



**Figure 1:** Cumulative density plot depicting the time (in minutes) taken by each treatment arm—Human, AI-assisted, and AI-Led—to computationally reproduce the main findings across both Study 1 and Study 2.

# Results: Coding Errors - Time

**Figure 2:** Cumulative density plots showing the time (in minutes) taken by each treatment arm— Human, AI-assisted, and AI-Led—to identify the first minor (left) and major (right) coding errors across both Study 1 and Study 2.



# Minor Coding Errors by Event

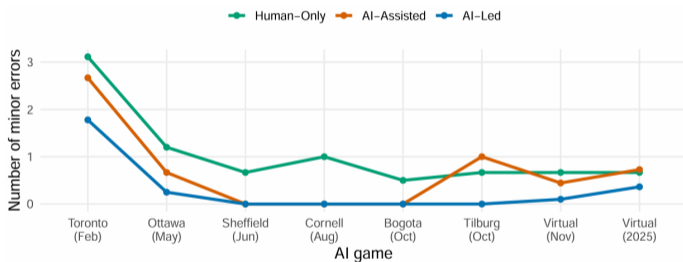


Figure 2: Differences in the number of minor errors detected across groups for each event.

# Major Coding Errors by Event

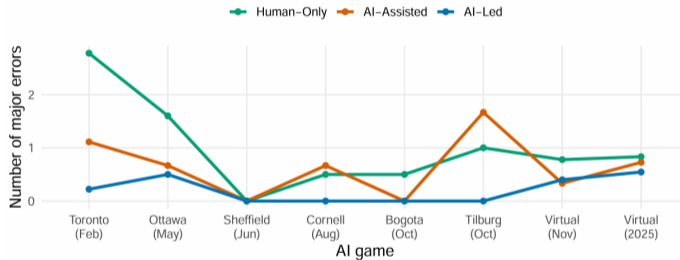


Figure 3: Number of major errors detected across groups for each event.

# Improvement between Study 1 and Study 2

**Table S10:** Results from OLS regressions predicting reproduction outcomes for study 2

|                                | (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                    | -0.022<br>(0.060)<br>[-0.141; 0.097]     | -0.467*<br>(0.275)<br>[-1.011; 0.078]    | -0.660**<br>(0.257)<br>[-1.170; -0.150]  | -0.009<br>(0.023)<br>[-0.055; 0.037]    | -0.013<br>(0.098)<br>[-0.207; 0.181]    | -0.024<br>(0.057)<br>[-0.138; 0.089]     | 0.009<br>(0.108)<br>[-0.205; 0.222]     |
| AI-Led                         | -0.595***<br>(0.089)<br>[-0.772; -0.418] | -0.948***<br>(0.241)<br>[-1.426; -0.469] | -1.089***<br>(0.241)<br>[-1.566; -0.612] | -0.172**<br>(0.067)<br>[-0.306; -0.039] | -0.258**<br>(0.102)<br>[-0.461; -0.056] | -0.325***<br>(0.095)<br>[-0.514; -0.137] | -0.312**<br>(0.120)<br>[-0.550; -0.075] |
| AI-Assisted × Study 2          | -0.022<br>(0.085)<br>[-0.191; 0.148]     | 0.491<br>(0.473)<br>[-0.446; 1.429]      | 0.604<br>(0.375)<br>[-0.139; 1.348]      | 0.042<br>(0.035)<br>[-0.027; 0.111]     | 0.021<br>(0.111)<br>[-0.199; 0.242]     | 0.132<br>(0.121)<br>[-0.107; 0.371]      | 0.068<br>(0.158)<br>[-0.245; 0.382]     |
| AI-Led × Study 2               | 0.395**<br>(0.159)<br>[0.080; 0.710]     | 0.506<br>(0.366)<br>[-0.219; 1.230]      | 0.884*<br>(0.467)<br>[-0.041; 1.809]     | 0.170**<br>(0.069)<br>[0.034; 0.307]    | 0.213*<br>(0.124)<br>[-0.032; 0.458]    | 0.382***<br>(0.132)<br>[0.120; 0.645]    | 0.238<br>(0.189)<br>[-0.136; 0.613]     |
| Mean of dep. var               | 0.788                                    | 0.861                                    | 0.723                                    | 0.956                                   | 0.832                                   | 0.854                                    | 0.737                                   |
| p-val (AI-Assisted vs. AI-Led) | 0.000                                    | 0.024                                    | 0.024                                    | 0.015                                   | 0.019                                   | 0.002                                    | 0.006                                   |
| Observations                   | 137                                      | 137                                      | 137                                      | 137                                     | 137                                     | 137                                      | 137                                     |

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

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

# Results by Software

Table 8: Results from OLS regressions predicting reproduction outcomes for different software

| Panel B: Study 2 combined      |   |   |   |                                       |                                       |  |  |
|--------------------------------|---|---|---|---------------------------------------|---------------------------------------|--|--|
|                                | (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                    | -0.010<br>( 0.049)<br>[-0.108; 0.088]     | -0.619**<br>( 0.308)<br>[-1.229; -0.009]  | -0.672**<br>( 0.293)<br>[-1.252; -0.092]  | 0.007<br>( 0.024)<br>[-0.040; 0.055]  | 0.101<br>( 0.077)<br>[-0.051; 0.253]  | 0.063<br>( 0.064)<br>[-0.064; 0.190]     | 0.200*<br>( 0.103)<br>[-0.004; 0.405]    |
| AI-Led                         | -0.475***<br>( 0.119)<br>[-0.710; -0.240] | -1.075***<br>( 0.353)<br>[-1.774; -0.376] | -1.172***<br>( 0.290)<br>[-1.747; -0.597] | -0.100<br>( 0.065)<br>[-0.229; 0.028] | -0.180<br>( 0.114)<br>[-0.406; 0.046] | -0.233**<br>( 0.108)<br>[-0.447; -0.018] | -0.220<br>( 0.155)<br>[-0.526; 0.086]    |
| R                              | -0.090<br>( 0.062)<br>[-0.213; 0.034]     | 0.023<br>( 0.376)<br>[-0.721; 0.768]      | -0.237<br>( 0.347)<br>[-0.924; 0.451]     | -0.002<br>( 0.027)<br>[-0.055; 0.052] | 0.029<br>( 0.093)<br>[-0.156; 0.214]  | -0.028<br>( 0.078)<br>[-0.182; 0.127]    | 0.051<br>( 0.124)<br>[-0.193; 0.296]     |
| AI-Assisted × R                | -0.047<br>( 0.099)<br>[-0.243; 0.148]     | 0.568<br>( 0.430)<br>[-0.283; 1.420]      | 0.295<br>( 0.434)<br>[-0.564; 1.154]      | -0.009<br>( 0.035)<br>[-0.078; 0.061] | -0.219<br>( 0.138)<br>[-0.491; 0.054] | -0.121<br>( 0.104)<br>[-0.326; 0.084]    | -0.359**<br>( 0.161)<br>[-0.677; -0.040] |
| AI-Led × R                     | -0.051<br>( 0.161)<br>[-0.370; 0.269]     | 0.432<br>( 0.432)<br>[-0.423; 1.288]      | 0.557<br>( 0.410)<br>[-0.255; 1.369]      | -0.049<br>( 0.092)<br>[-0.232; 0.133] | -0.058<br>( 0.158)<br>[-0.370; 0.254] | -0.013<br>( 0.149)<br>[-0.309; 0.283]    | -0.083<br>( 0.195)<br>[-0.469; 0.303]    |
| Controls                       | ✓   | ✓   | ✓   | ✓                                     | ✓                                     | ✓  | ✓  |
| Mean of dep. var               | 0.788                                     | 0.861                                     | 0.723                                     | 0.956                                 | 0.832                                 | 0.854                                    | 0.737                                    |
| p-val (AI-Assisted vs. AI-Led) |   |   |   |                                       |                                       |  |  |
| Obs.                           | 137                                       | 137                                       | 137                                       | 137                                   | 137                                   | 137                                      | 137                                      |

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

Controls include number of teammates; game, skill, and attendance fixed effects.

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

# Results by Prompt Number

Table 9: Comparison of Key Metrics by Prompt Levels within AI-Assisted Branch

| Variable                                       | Above median<br>(n=23) | Below/equal to median<br>(n=23) | Difference        |
|--|------------------------|---------------------------------|-------------------|
| Reproduction                                   | 1.000<br>(0.000)       | 0.870<br>(0.344)                | 0.130<br>(0.083)  |
| Minutes to reproduction                        | 94.9<br>(88.2)         | 75.9<br>(65.4)                  | 19.0<br>(0.423)   |
| Number of minor errors                         | 0.826<br>(0.984)       | 0.957<br>(1.581)                | -0.130<br>(0.739) |
| Minutes to first minor error                   | 119.7<br>(77.7)        | 71.8<br>(42.2)                  | 47.9<br>(0.071)   |
| Number of major errors                         | 0.696<br>(0.926)       | 0.609<br>(0.783)                | 0.087<br>(0.733)  |
| Minutes to first major error                   | 141.2<br>(87.2)        | 127.2<br>(42.8)                 | 14.0<br>(0.632)   |
| At least one appropriate robustness check      | 1.000<br>(0.000)       | 1.000<br>(0.000)                | 0.000<br>(-)      |
| At least two appropriate robustness checks     | 0.826<br>(0.388)       | 0.957<br>(0.209)                | -0.130<br>(0.164) |
| Ran at least one appropriate robustness check  | 0.957<br>(0.209)       | 0.957<br>(0.209)                | 0.000<br>(1.000)  |
| Ran at least two appropriate robustness checks | 0.783<br>(0.422)       | 0.913<br>(0.288)                | -0.130<br>(0.228) |

Note: Group columns show mean (SD); the Difference column is Above Below with a two-sided Welch p-value in parentheses. Groups are defined by the median number of prompts ( 25.5 ) in the AI-Assisted sample.

# Does experience with ChatGPT matter? Study I and II combined

Table 10: AI-Assisted and AI-Led Metrics by Experience Level (Study I and II)

| Variable                                       | AI-Assisted<br>high experience<br>(n=16) | AI-Assisted<br>low/medium experience<br>(n=29) | AI-Led<br>high experience<br>(n=15) | AI-Led<br>low/medium experience<br>(n=31) | AI-Assisted<br>High vs Low | AI-Led<br>High vs Low |
|--|--|--|-------------------------------------|---|----------------------------|-----------------------|
| Reproduction                                   | 1.000<br>(0.000)                         | 0.897<br>(0.310)                               | 0.600<br>(0.507)                    | 0.419<br>(0.502)                          | 0.103<br>(0.083)           | 0.181<br>(0.265)      |
| Minutes to reproduction                        | 79.5<br>(75.7)                           | 78.7<br>(56.6)                                 | 102.7<br>(84.1)                     | 157.1<br>(83.3)                           | 0.8<br>(0.971)             | -54.4<br>(0.152)      |
| Number of minor errors                         | 1.125<br>(1.668)                         | 0.793<br>(1.082)                               | 0.400<br>(0.910)                    | 0.516<br>(0.851)                          | 0.332<br>(0.481)           | -0.116<br>(0.682)     |
| Minutes to first minor error                   | 120.0<br>(89.7)                          | 78.9<br>(33.5)                                 | 94.2<br>(30.3)                      | 159.7<br>(111.9)                          | 41.1<br>(0.176)            | -65.4<br>(0.116)      |
| Number of major errors                         | 0.812<br>(0.911)                         | 0.586<br>(0.825)                               | 0.533<br>(0.743)                    | 0.194<br>(0.477)                          | 0.226<br>(0.416)           | 0.340<br>(0.122)      |
| Minutes to first major error                   | 127.3<br>(87.2)                          | 140.0<br>(57.4)                                | 87.7<br>(46.6)                      | 220.6<br>(104.4)                          | -12.7<br>(0.709)           | -132.9<br>(0.043)     |
| At least one appropriate robustness check      | 1.000<br>(0.000)                         | 1.000<br>(0.000)                               | 0.867<br>(0.352)                    | 0.871<br>(0.341)                          | 0.000<br>(-)               | -0.004<br>(0.969)     |
| At least two appropriate robustness checks     | 0.938<br>(0.250)                         | 0.862<br>(0.351)                               | 0.800<br>(0.414)                    | 0.645<br>(0.486)                          | 0.075<br>(0.408)           | 0.155<br>(0.270)      |
| Ran at least one appropriate robustness check  | 1.000<br>(0.000)                         | 0.966<br>(0.186)                               | 0.733<br>(0.458)                    | 0.645<br>(0.486)                          | 0.034<br>(0.326)           | 0.088<br>(0.553)      |
| Ran at least two appropriate robustness checks | 0.938<br>(0.250)                         | 0.828<br>(0.384)                               | 0.667<br>(0.488)                    | 0.484<br>(0.508)                          | 0.110<br>(0.253)           | 0.183<br>(0.250)      |

Note: Group columns display mean (SD). The two right-most columns show High – Low/Medium differences within each branch, with two-sided Welch p-values in parentheses.

# ChatGPT Models Available by Training

| 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 <sup>†</sup> | Yes    |                            | Yes | Yes | Yes     | Yes                    | Yes <sup>‡</sup> |                                    |
| Tilburg    | Oct 18 | Sep 30              | Yes    |                            | Yes | Yes | Yes     | Yes                    | Yes <sup>‡</sup> |                                    |
| Virtual    | Nov 22 | Nov 8               | Yes    |                            | Yes | Yes | Yes     | Yes                    | Yes              |                                    |
| 2025 Event | Apr 30 | Apr 28              | Yes    |                            | Yes | Yes | Yes     |                        | Yes <sup>§</sup> | Yes                                |

\* Image upload part of 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.

# ChatGPT Model Capabilities

| 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 <sup>1</sup>       | Yes         | Yes                     | Yes          | No        |
| GPT-4o-mini        | July 18, 2024 <sup>2</sup>      | Yes*        | Yes*                    | Yes*         | No        |
| o1-preview         | September 12, 2024 <sup>3</sup> | No          | No                      | No           | Yes       |
| o1-mini            | September 12, 2024 <sup>3</sup> | No          | No                      | No           | Yes       |
| GPT-4o with canvas | October 3, 2024 <sup>4</sup>    | Yes         | Yes                     | No           | No        |
| GPT-4.5 (preview)  | February 27, 2025 <sup>5</sup>  | Yes         | Yes                     | Yes          | No        |
| o3                 | April 16, 2025 <sup>6</sup>     | Yes         | Yes                     | Yes          | Yes       |
| o4-mini            | April 16, 2025 <sup>6</sup>     | Yes         | Yes                     | Yes          | Yes       |
| o4-mini-high       | April 16, 2025 <sup>6</sup>     | 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/>

◀ ▶ ⏪ ⏩ 🔍 ↻

# Full List of Authors (1/3)

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 (University of Melbourne), Tom Stafford (Sheffield University), Lars Vilhuber (Cornell University), Gunther Bensch (RWI - Leibniz Institute for Economic Research), 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), Igor Asanov (International Center for Higher Education Research and Faculty of Economics, University of Kassel), Anastasiya-Mariya Asanov Nohad (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 (NHH Norwegian School of Economics), Louis-Philippe Beland (Carleton University), Carsten 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 (Faculty of Biology, Medicine and Health, The University of Manchester), Katherine Brennan (Bank of Canada), 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), 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 (Goethe-Universität Frankfurt), Carlos Daniel Bermúdez Jaramillo (Universidad del Rosario), Hannes Datta (Tilburg University), Lien Denoo (Tilburg University), Arshia Dhaliwal (Carleton University), Nancy 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 Elmenejad (Nazarbayev University), Mahmoud Elsherif (University of 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),

## Full List of Authors (2/3)

Victoria Fernandes (Bank of Canada), Lenka Fiala (University of Bergen), Lukas Fink (FU Berlin), 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 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; Liège University), 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), Edmund Kelly (University of Oxford), Eva Kimel (University of York), Sorraich Kingswankul (Vrije Universiteit Amsterdam), Valentin Klotzbücher (University of Freiburg), Daniel Krämer (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 München) 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), 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 (University of Notre Dame), 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), Fabio Motoki (University of East Anglia), Lucija Muehlenbachs (University of Calgary), Gastón Mullin (Tilburg University), Andreea Musulan (University of Montreal), Sara Muzzi (University of Milano Bicocca), James A. C. Myers (University of Sheffield), Florian Neubauer (RWI - Leibniz Institute for Economic Research),

# Full List of Authors (3/3)

Tuan Nguyen (Hasselt University), Ali Niazi (University of Calgary), Ardyn Nordstrom (Carleton University), Bartłomiej Nowak (Cardinal Stefan Wyszyński University), Daneal O'Habib (Bank of Canada), Tim Ölkens (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 Phielor (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, 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 (Bank of Canada), 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 Sudhakaran (Tilburg University), Pu Sun (University of Ottawa), 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, QLD), 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), 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 Zirculis (ISM University of Management and Economics), Jiacheng Zou (Columbia University), Floris Zoutman (NHH Norwegian School of Economics), Christelle Zozoungbo (Penn State University).