Pre-Analysis Plan

# Reproducing with AI Across the Expertise Ladder

***Version 0.3 – 12 Jun 2025***

## Principal Investigators & Co-authors

Ghina Abdul Baki (University of Ottawa)

Juan P. Aparicio (University of Ottawa)

Bruno Barbarioli (University of Ottawa)  
Abel Brodeur (University of Ottawa)

Lenka Fiala (University of Ottawa)  
Derek Mikola (University of Ottawa)

David Valenta (University of Ottawa)  
et al. (All participants who comply with protocol will be listed after the PIs).

## Funding & Registration

Funded by Open Philanthropy and the Institute for Replication (I4R).

This PAP will be preregistered on OSF and GitHub immediately after pilot power simulations (target September 2025).

## 1. Background & Rationale

Our 2024–25 AI‑Replication Games (Brodeur et al. [econstor.eu/bitstream/10419/308508/1/I4R-DP195.pdf](https://www.econstor.eu/bitstream/10419/308508/1/I4R-DP195.pdf)) showed that AI assistance changes between‑team performance in reproduction tasks. This study examines the vertical dimension of expertise—professors, post‑docs, PhD students, master’s students, undergraduates—on some of the same tasks. Large‑language models (LLMs) may be a “great equalizer”, narrowing productivity gaps, or could amplify expert advantages through better prompting. Randomizing individuals to AI versus control (human-only) within expertise strata lets us test the following hypotheses.

## 2. Research Questions and Hypotheses

| **ID** | **Research Question** | **Directional Hypothesis** |
| --- | --- | --- |
| H1 | Does ChatGPT Plus raise average performance in reproduction tasks? | Yes: β1 > 0 |
| H2 | Does AI compress performance gaps between expertise tiers? | Yes: δk < 0 for lower tiers |
| H3 | Where is compression largest? | Greatest for MA/UG tiers |
| H4 | Does AI accelerate task completion? | Yes: β\_time < 0 |

Exploratory: (i) dose–response by years of coding; (ii) AI prompt counts; (iii) learning across events.

The main outcome variables will be achieving computational reproducibility, detecting coding errors, ***proposing*** robustness checks and XXX.

## 3. Experimental Design

### 3.1 Treatments

* Control — Human‑only: participants pledge not to use AI tools.
* Treatment — Human + ChatGPT Plus: full subscription (GPT‑4o and successors, code interpreter, vision).

### 3.2 Expertise Strata

| Tier | Definition | Target N |
| --- | --- | --- |
| P | Professor (tenure‑track/tenured) | 60 |
| PD | Post‑doctoral researcher / non‑TT faculty | 60 |
| PhD | Doctoral candidate | 60 |
| MA | Master’s student | 60 |
| UG | Final‑year undergraduate | 60 |

Randomization: 1:1 AI vs. control within each stratum.

### 3.3 Sample Size & Power

With a planned sample of 300 individuals, roughly 60 per expertise tier, our 10,000-iteration power simulation starts from three empirical premises: (i) in the control arm, professors outperform undergraduates by about 15 percentage points in successfully reproducing results; (ii) granting ChatGPT Plus is expected to shrink that gap by at least 40 percent (a six-point improvement for undergraduates); and (iii) hypothesis tests are run at the 5 percent level with standard errors clustered by event-software cells. Under those assumptions, the design attains ≈ 82 percent power to detect the key interaction for undergraduates (δ\_UG = –0.06) and over 90 percent power for the joint “compression across all tiers” test, providing ample sensitivity without exceeding the logistical limits of four one-day events.

## 4. Setting, Participants, & Authorship

| Item | Details |
| --- | --- |
| Events | Four one‑day games Oct–Nov 2025. |
| Recruitment | I4R mailing lists, departmental calls, social media. |
| Eligibility | Quant methods background; own laptop; data‑sharing agreement. |
| Authorship | All compliant participants listed alphabetically after PIs. |
| Ethics | No sensitive data; IRB exemption anticipated. |

## 5. Tasks & Materials

* Computational reproduction of a prespecified result.
* Error detection (classify major versus minor).
* Propose and attempt at least two robustness checks.
* One-page referee report (tbd)

Materials: PDF, replication package, screenshot, Excel logbook, and ChatGPT transcript template (treatment only). Work window: 7 h (09:00–16:00). Treatment arm completes 1‑h ChatGPT training beforehand.

## 6. Outcome Variables

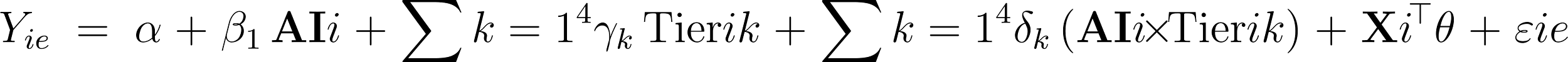
| **Domain** | **Variable** | **Scale** | **Source** |
| --- | --- | --- | --- |
| Reproduction | Success (0/1) | Binary | Log sheet |
| Reproduction | Minutes‑to‑success (censor 420) | Continuous | Log sheet |
| Error detection | # Minor errors | Count | Error tab |
| Error detection | # Major errors | Count | Error tab |
| Robustness | ≥1 / ≥2 “good” checks | Binary | Review panel |
| Implementation | Ran ≥1 proposed check | Binary | Code evidence |
| Clarity | Mean 0‑5 blind score | Continuous | Grading rubric |

## 7. Covariates

* Years of coding experience (self‑reported).
* Preferred software (R, Stata, Python).
* Event and article fixed effects.
* Prior ChatGPT familiarity (exploratory).

## 8. Statistical Analysis Plan

Our core estimation equation is

[](https://www.codecogs.com/eqnedit.php?latex=Y_%7Bie%7D%20%5C%3B%3D%5C%3B%20%5Calpha%20%5C%3B%2B%5C%3B%20%5Cbeta_%7B1%7D%5C%2C%5Cmathbf%7BAI%7D%7Bi%7D%20%5C%3B%2B%5C%3B%20%5Csum%7Bk%3D1%7D%5E%7B4%7D%20%5Cgamma_%7Bk%7D%5C%2C%5Ctext%7BTier%7D%7Bik%7D%20%5C%3B%2B%5C%3B%20%5Csum%7Bk%3D1%7D%5E%7B4%7D%20%5Cdelta_%7Bk%7D%5C%2C(%5Cmathbf%7BAI%7D%7Bi%7D%5C!%5Ctimes%5C!%5Ctext%7BTier%7D%7Bik%7D)%20%5C%3B%2B%5C%3B%20%5Cmathbf%7BX%7D%7Bi%7D%5E%7B%5C!%5Ctop%7D%5Ctheta%20%5C%3B%2B%5C%3B%20%5Cvarepsilon%7Bie%7D#0)

where

| **Symbol** | **Definition** | **Interpretation** |
| --- | --- | --- |
|  | Outcome for individual *i* in event *e*. |  |
|  | Intercept | Mean outcome for **undergraduates in the control group** (reference category). |
|  | Treatment dummy (= 1 if the participant had ChatGPT Plus access) | captures the average treatment effect for undergraduates (the omitted tier). |
|  | Four dummies identifying the higher-ranked tiers—Master’s, PhD, Post-doc, and Professor (undergraduates are the baseline) | Each  measures how much better or worse that tier performs relative to undergraduates when **no AI** is available. |
|  | Interaction between treatment and tier | Each  tells us whether ChatGPT Plus changes the *gap* between tier *k* and undergraduates—negative values indicate “compression.” |
|  | Control vector (preferred software, years of coding experience, prior ChatGPT familiarity) | absorbs any confounding variation unrelated to the experiment. |
|  | Error term | Assumed mean 0; we cluster its variance by **event × software** to accommodate intra-cluster correlation. |

### Estimation method.

* For binary and continuous-time outcomes, we use ordinary least squares (OLS).
* For count outcomes (e.g., number of errors) we use Poisson regression with the same right-hand-side.  
  Standard errors are Huber–White and clustered at the event–software level, reflecting that participants working in the same software environment during the same event share local shocks.

### Secondary analyses.

1. Replace the five tier dummies with a continuous measure—self-reported years of coding—to estimate a dose–response curve.
2. Within the AI arm, regress outcomes on [](https://www.codecogs.com/eqnedit.php?latex=%5Clog(%5Ctext%7Bnumber%20of%20ChatGPT%20prompts%7D%2B1)#0) to study over- or under-use.
3. Add an interaction between treatment and event order to see whether learning accumulates across events.
4. For time-to-event outcomes, we complement OLS with Kaplan–Meier survival curves and log-rank tests (reported in the appendix).

Together, these specifications isolate the causal impact of AI access, reveal whether it narrows vertical expertise gaps, and check how robust the findings are to alternative functional forms and behavioral metrics.

## 9. Compliance, Deviations, & Monitoring

* Compare assignment with ChatGPT logs and observer notes.
* Any protocol change preregistered on OSF and GitHub before data access.

## 10. Data Management & Sharing

* Raw logs, transcripts, cleaned data on OSF and GitHub (view‑only until publication).
* Replication packages under journal licenses; provide download links.
* De‑identified grading sheets included.

## 11. Timeline

| **Milestone** | **Target date** |
| --- | --- |
| Pilot sims & final SAP | 31 Aug 2025 |
| OSF and GitHub preregistration | 15 Sep 2025 |
| Event 1 | 07 Oct 2025 |
| Event 2 | 21 Oct 2025 |
| Event 3 | 04 Nov 2025 |
| Event 4 | 18 Nov 2025 |
| Cleaning & grading | Dec 2025 |
| Unblinding & analysis | Jan 2026 |
| Draft manuscript | Mar 2026 |

## 12. Limitations & Risk Mitigation

* Self‑report bias mitigated by continuous experience measure.
* Control contamination checked via pledge plus random screen‑shares.
* LLM drift documented by logging GPT model names/dates.

## 13. References

1. AI‑Replication Games (2025), Nature.

2. 2024 Pre‑Analysis Plan (OSF).