



Scaling up classroom observation with large language models: A reliability study of the World Bank's Teach framework in Peru

Second Submission: RR_PER_2026_700

Vaishnavi Dhas

reproducibility@worldbank.org

August 10, 2026

This review verifies the reproducibility of the exhibits included in the paper “Scaling up classroom observation with large language models: A reliability study of the World Bank's Teach framework in Peru”.

Contents in this review:

1. Main findings
2. List of exhibits and reproducibility status
3. Reproduction Environment

Main findings

- **Every exhibit in the main section of the paper and the appendix were reproduced accurately**
- The code was successfully executed on a new computer after:
 1. Restoring the environment in a clean conda environment using "requirements.txt"
 2. Setting the environment variable AITEACH_USE_CACHED_LLM=1 to run the cached version of the pipeline using the shipped de-identified intermediate data (without requiring live LLM API calls).
 3. Running "oo_master.py"
- The output demonstrates consistent stability across multiple runs. Specifically, executing the code two times consecutively yielded identical results.
- The code takes approximately 5 minutes to run.
- We conducted our reproducibility analysis based on the paper shared by the authors by email on July 10, 2026.
- **Verification Process and Data Handling:**
 - The reproducibility package relies on 1 type of data: restricted data (available only from the data owners).
 - Reviewers used an intermediate version of restricted data provided directly by the authors to conduct the reproducibility verification, , and this is the starting point of the public reproducibility package. The data is forthcoming on the World Bank Microdata Library.

- *data_hash_report.csv* lists the SHA256 hashes of all files in the Data folder to support data integrity checks. Users who acquire the forthcoming data can use this file to verify that the data has not been altered.
- **Reproducibility Summary:**
 - **Data:** Some data is restricted and has not been included in the reproducibility package. For more details, please refer to the README file. Intermediate data is forthcoming in the World Bank Microdata Library under a license and can be used to fully reproduce the results in the paper.
 - **Code:** All code files (from cleaning to analysis) are included in the reproducibility package.
 - **Outputs:** All outputs are generated by code included in the reproducibility package.
 - **Dependencies environment:** The reviewers created a new environment using the latest versions of dependencies available at the moment of the review.

List of exhibits and reproducibility status

Results in the Main Section of the Paper

- **Figure 1** Does not apply. The exhibit does not contain analytical output.
- **Figure 2** Reproduced.
- **Figure 3** Reproduced.
- **Figure 4** Reproduced.
- **Figure 5** Reproduced.
- **Table 1** Does not apply. The exhibit does not contain analytical output.
- **Table 2** Reproduced.
- **Table 3** Reproduced.
- **Table 4** Reproduced.
- **Table 5** Reproduced.
- **Table 6** Reproduced.

Results in the Annex

- **Table S1** Reproduced with changes.
- **Table S2** Reproduced with changes.
- **Table S3** Reproduced.
- **Table S4** Reproduced.
- **Table S5** Reproduced with changes.

Reproduction Environment

Paper exhibits were reproduced on a computer with the following specifications:

- OS: Windows 11 Enterprise
- Processor: Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz (2.30 GHz) (2 processors)
- Memory available: 16.0 GB
- Software version: Python 3.11.15