README

Code and Data for “Scaling up classroom observation with large language models: A reliability study of the World Bank’s Teach framework in Peru” by Matthew D. Krasnow, Carolina Lopez, Jostin Kitmang, William J. Krasnow, Carolina Moreira Vasquez, and Ezequiel Molina.

Overview

This document provides the information needed to replicate all tables and figures in “Scaling up classroom observation with large language models: A reliability study of the World Bank’s Teach framework in Peru”. The replication package is organized into the following folders:

- code/
- data/
- output/
- paper/

Files and Contents

All analysis is conducted in Python (pandas / numpy / scipy / matplotlib; pipeline stages are Jupyter notebooks executed programmatically via nbconvert).

code/00_master.py: Master script. Running this file from top to bottom reproduces all tables and figures in the paper. It runs the numbered pipeline stages first and then one script per exhibit (see “How to Run” below). No paths need to be edited; the script locates the package root from its own location. Estimated runtime in cached mode is under five minutes.

code/01_video_assignment.ipynb – *code/05_evaluation.ipynb*: Pipeline stages 1–5 (video assignment, audio conversion, transcript generation, cleaning and EDA, LLM evaluation). These stages require the restricted raw recordings and/or live LLM API access; in cached mode (recommended for review) they are skipped and the pipeline starts from the cached de-identified intermediates shipped in data/.

code/06_performance.ipynb: Stage 6, performance analysis. Reads data/model_evaluation_data/ and the de-identified human scores, runs the canonical Monte Carlo certification-exam simulation (1,000 iterations, seed 42, 2 attempts), and writes data/model_performance_metrics.csv (the Figure 4 source) plus per-model diagnostic reports to output/reports/.

code/07_statistical_analysis.ipynb: Stage 7, statistical analysis and robustness checks. Writes data/model_comparison_summary.csv (an alternative-rule Monte Carlo used as a robustness check; the manuscript headline numbers come from stage 6).

code/tables/: One script per table in the paper (table2.py, table3.py, table4.py, table5.py, table6.py), each writing a LaTeX table to output/tables/. table2.py–table5.py compute (and verify against the manuscript) their statistics from the shipped data; table6.py emits the cost table, whose components are external April-2025 list prices. Table 1 (the Teach framework structure) is a hand-authored LaTeX table shipped directly as output/tables/table1.tex.

code/tables/generate_tables_s1_s5.py: Generates the five supplementary tables (Tables S.1–S.5) from the same shipped data, with no API calls. Each table is written to output/tables/ as both a CSV and a LaTeX file, and all of them are additionally collected into a single Excel workbook, output/tables/supp_tables.xlsx, one sheet per table. Table S.5 is written as two files, matching its two panels in the supplementary materials: the per-school clip distribution and the five-fold school-grouped cross-validation. The script recomputes the Table S.4 and Table S.5 statistics and verifies them against the values in the submitted supplementary materials, raising an error if any of them drift.

code/figures/: One script per figure (figure2.py, figure3.py, figure4.py, figure5.py), writing to output/figures/. Figure 1 (the pipeline diagram) is a hand-authored TikZ figure shipped directly as output/figures/figure1.tex.

code/models/: The LLM scoring pipeline used by stage 5: the Gemini API client and prompt templates, the control evaluator (all configurations except M5) and the timstamped evaluator (M5), and the rubric JSON configurations in code/models/_context/.

code/verify_numerics.py: Regression gate that cross-checks the manuscript's headline numerics against the regenerated CSVs.

data/: Contains all datasets used by the code (see the Data Availability Statement below). peru_anonymized_transcripts.csv is the de-identified analytic dataset (199 clip rows; anonymized transcript text of the scored segment on each row — see the Data Note below — plus human Teach scores). test_only_peru_anonymized_transcripts.csv is its test-split subset (84 clips). min15_clips.csv lists the clips of at least 15 minutes used by stage 6. TEACH_Final_Scores_4 - Peru.csv contains the de-identified human-expert Teach scores. model_evaluation_data/ contains the cached LLM evaluation outputs (one CSV per model configuration) so stage 5 does not need to be re-run. The remaining CSVs are exhibit sources (fig2_domain_accuracy.csv for Figure 2 and fig5_significance_matrix.csv

for Figure 5) and derived metrics (model_performance_metrics.csv, model_comparison_summary.csv). model_id_map.json maps the manuscript model IDs M0–M7 to the internal run names and is read by code/tables/table5.py, code/tables/generate_tables_s1_s5.py, code/figures/figure2.py, code/figures/figure5.py and code/verify_numerics.py. Tables 2 and 4 are computed directly from peru_anonymized_transcripts.csv. Three further files — Table1_Numeric_Dimensions.csv, Table2_Categorical_Dimensions.csv and TEACH_Numeric_Dimensions_Summary.csv — are precomputed summary tables from an earlier descriptive analysis, retained for reference; no script in this package reads them. The confidential source recordings and unredacted transcripts are not distributed with this package.

output/tables/ and output/figures/: Exhibit files produced by the scripts in code/tables/ and code/figures/, one file per exhibit, named after the corresponding table or figure. output/reports/ contains regenerable per-model diagnostic reports (not paper exhibits).

paper/replication.tex: A standalone LaTeX file that compiles all tables and figures from the paper in order, including the supplementary Tables S.1–S.5 (numbered S.1–S.5 in the compiled PDF; the two panels of Table S.5 share one float). Compile with pdflatex (run twice). A compiled copy of that document, paper/replication.pdf, ships with the package.

Software Requirements

- Python 3.10 or later (the package was verified with Python 3.11).
- Python packages pinned in code/requirements.txt (pandas 2.2.1, numpy 1.26.4, scipy 1.15.3, scikit-learn 1.8.0, matplotlib 3.10.1, seaborn 0.13.2, jupyter/nbconvert, and others). Install with: pip install -r code/requirements.txt
- A Google Gemini API key is required only for a live re-run of stage 5 (LLM evaluation); it is not required in cached mode.

The shipped LLM outputs were generated with pinned model snapshots (gemini-2.5-flash-preview-04-17 and gemini-2.0-flash-001, temperature 0, seed 42). Because provider-side model updates can shift free-text outputs, the cached outputs in data/model_evaluation_data/ are the canonical stage-5 results; reviewers re-running stage 5 live should expect minor drift.

How to Run

1. Install dependencies: pip install -r code/requirements.txt

2. Recommended (cached mode, no API keys, starts from the shipped de-identified intermediates): set the environment variable AITEACH_USE_CACHED_LLM=1 and run python code/00_master.py
3. Full pipeline with live LLM API calls (requires API keys): python code/00_master.py

Examples:

```
# macOS / Linux
AITEACH_USE_CACHED_LLM=1 python code/00_master.py

# Windows (PowerShell)
$env:AITEACH_USE_CACHED_LLM = "1"; python code\00_master.py
```

Tables and figures are saved to output/tables/ and output/figures/. Per-stage logs are written to a _runlog/ folder, which the master script creates at the package root on the first run. To produce a PDF with all exhibits, compile paper/replication.tex with pdflatex (run twice). To verify the manuscript's headline numerics after a run: python code/verify_numerics.py (expected: 5/5 OK).

In cached mode, stages 1–5 are skipped by design: stages 1–3 require the restricted raw recordings, stage 4 starts from data already shipped, and stage 5 requires live LLM API access. Stages 6–7 and all exhibit scripts execute and regenerate every output exhibit deterministically (fixed seeds; pandas/scipy/statsmodels/scikit-learn/matplotlib, no network access).

Relevant Sample

The sample comprises 199 Peruvian primary-classroom observations, retained from 363 raw video-observation entries (354 distinct classroom observations after removing duplicate clip-mappings) collected under the Peru Teach pilot. Each observation is one row of data/peru_anonymized_transcripts.csv and is scored from a single 15-minute audio segment — the first or the last part of a lesson recording — as described in the Data Note below. Each observation was scored by certified human Teach raters on the Teach 1.0 rubric and by eight LLM configurations (Gemini 2.0 Flash and 2.5 Flash, M0–M7) varying rubric quality, prompt format, and reasoning depth. Predicted scores are benchmarked against the World Bank Teach certification exam via Monte Carlo simulation (1,000 iterations, seed 42, 2 attempts to certify).

Variable Construction and Estimation

Teach domain/element scores are on the Teach 1.0 rubric's 1–5 scale (see the rubric JSONs in code/models/_context/). The train/test split is 60/40 stratified with seed 42;

the split indicator is the split column of peru_anonymized_transcripts.csv. Within-one-point agreement is computed per Teach element as the share of LLM-predicted scores within ±1 point of the human-rater score. Segment reliability follows the Teach certification rule (agreement on at least 8 of 10 elements, with Time-on-Task snapshots requiring exact agreement on 2 of 3).

Two Monte Carlo certification-exam simulations exist and should not be confused: the canonical simulation in stage 6 (Teach 8-of-10 rule including Time-on-Task; source of the manuscript headline 92.9% [91.3, 94.5] for M3, in data/model_performance_metrics.csv) and a robustness check in stage 7 (alternative 8-of-9 rule excluding Time-on-Task, 10,000 iterations, in data/model_comparison_summary.csv). Manuscript headlines should be compared against the stage-6 file.

De-identification

The transcript text in peru_anonymized_transcripts.csv was manually de-identified: real person names and sub-department place names were replaced with consistent placeholder tags ([Student A], [Teacher], [Fieldworker], [School Name], [Local Person A], [Family Member A], [Local Place]). Redaction was performed by readers reviewing each transcript in full rather than by pattern matching, followed by an independent adversarial QA pass over every row and a final full-corpus verification that cross-referenced every known real name against the final text. Department/province/region-level place names (e.g., Lima, Cusco) are retained, matching how the paper describes sampling regions; names of characters in printed reading passages, fables, word problems, prayers, or songs read aloud are also retained, as they are text content rather than identifying information. Content is otherwise preserved verbatim (pauses, disfluencies, pedagogical content, punctuation). All other columns (Teach scores, clip metadata, split) are unchanged from the source file.

The following are not distributed with this package: the underlying audio/video recordings, the unredacted transcripts and their intermediate variants, and any per-teacher identifying records.

Data Note: Transcript Fields on Clip Rows

Each observation session contributes two rows (Clip 1 / Clip 2), and both rows carry the session-level “First Audio Transcript …” and “Last Audio Transcript …” field groups. This layout comes from the original export (one wide row per clip carrying both of the session’s audio segments), not from the anonymization. The evaluation pipeline selects

the segment per clip: Clip 1 rows are scored from “First Audio Transcript Text” and Clip 2 rows from “Last Audio Transcript Text” (see the transcript-selection logic in the evaluation stage, stage 5). In the distributed files the unused transcript-text field on each row has been blanked; it is never read by any script, so no result is affected. Teach scores are clip-specific on every row.

Data Availability Statement

The data used in this paper come from classroom observations collected in Peru by the World Bank’s Global Education Policy Dashboard team (https://www.educationpolicydashboard.org/), and access to all datasets is restricted. The underlying classroom video recordings contain personally identifiable information (PII) about teachers and minors and are not distributed; the recordings and all data-collection information and materials may be requested from the Dashboard team. The de-identified datasets included in data/ are shared solely to enable replication of the results in this paper: even after de-identification, classroom transcripts could in principle be re-identified by someone with contextual knowledge of the classrooms, so they will not be deposited publicly and may not be redistributed.

Detailed information for the main datasets:

File name: data/peru_anonymized_transcripts.csv. Description: De-identified analytic dataset; one row per classroom video clip (199 clips), with anonymized transcript text of the scored segment, Teach 1.0 rubric scores by certified human raters, clip metadata, and the train/test split indicator. Citation: Krasnow, Matthew D., Carolina Lopez, Jostin Kitmang, William J. Krasnow, Carolina Moreira Vasquez, and Ezequiel Molina. “Scaling up classroom observation with large language models: A reliability study of the World Bank’s Teach framework in Peru.” Access: Restricted. Included in this package solely for replication; not publicly deposited due to re-identification risk; no redistribution.

File name: data/TEACH_Final_Scores_4 - Peru.csv. Description: De-identified human-expert Teach scores for the retained sessions. Access: Restricted. Included in this package solely for replication; no redistribution.

File name: data/model_evaluation_data/ (18 CSV files). Description: Cached LLM evaluation outputs, one file per model configuration, generated in April 2025 with the pinned Gemini snapshots listed above. These are the canonical stage-5 outputs. Access: Restricted. Included in this package solely for replication (de-identified); no redistribution.

File name: raw classroom video recordings and data-collection materials (not distributed). Description: Original classroom videos (~363 raw video-observation entries) and the associated data-collection information and documentation. Access: Restricted (PII). May be requested from the World Bank Dashboard team.

The Teach 1.0 rubric and framework are publicly available from the World Bank under CC-BY 4.0.

List of Exhibits

The following table maps each output file generated by the replication code to the corresponding table or figure in the paper.

Output file	Paper exhibit	Generating script	Description
output/tables/table1.tex	Table 1	shipped (hand-authored)	Teach framework: areas, elements, and observable behaviors
output/tables/table2.tex	Table 2	code/tables/table2.py	Descriptive statistics for the Peru analytic sample
output/tables/table3.tex	Table 3	code/tables/table3.py	The eight LLM scoring configurations (M0–M7), with Monte Carlo certification rates
output/tables/table4.tex	Table 4	code/tables/table4.py	Mean Teach domain scores from human evaluators
output/tables/table5.tex	Table 5	code/tables/table5.py	Element-level within-one-point accuracy: baseline (M1) vs. best configuration (M3)
output/tables/table6.tex	Table 6	code/tables/table6.py	Marginal

			cost of LLM-augmented Teach observation (per lesson)
output/figures/figure1.tex	Figure 1	shipped (hand-authored TikZ)	Data processing pipeline from raw classroom videos to analytic sample
output/figures/figure2.png (.pdf)	Figure 2	code/figures/figure2.py	Domain-level within-one-point accuracy by model (M0–M7)
output/figures/figure3.png (.pdf)	Figure 3	code/figures/figure3.py	Monte Carlo exam pass probability by model, with 95% CIs
output/figures/figure4.png (.pdf)	Figure 4	code/figures/figure4.py	Monte Carlo certification outcomes (pie charts per configuration)
output/figures/figure5.png (.pdf)	Figure 5	code/figures/figure5.py	Pairwise statistical-significance matrix (M0–M7)
output/tables/table_s1.csv (.tex)	Table S.1	code/tables/generate_tables_s1_s5.py	Selection check: the 155 excluded vs. the 199 retained sessions on the nine Teach elements
output/tables/table_s2.csv (.tex)	Table S.2	code/tables/generate_tables_s1_s5.py	Development-vs-test balance check on the nine Teach elements
output/tables/table_s3.csv (.tex)	Table S.3	code/tables/generate_tables_s1_s	Per-element

		5.py	signed deviation Mean(LLM − Human) and SD, for M1 and M3
output/tables/table_s4.csv (.tex)	Table S.4	code/tables/generate_tables_s1_s5.py	Design-dimension OLS with school-clustered standard errors
output/tables/table_s5_clip_distribution.csv (.tex)	Table S.5, clip-distribution panel	code/tables/generate_tables_s1_s5.py	Test clips per anonymized school (46 schools contribute one clip, 19 contribute two)
output/tables/table_s5.csv (.tex)	Table S.5, cross-validation panel	code/tables/generate_tables_s1_s5.py	Five-fold school-grouped cross-validation of M3's within-one-point accuracy
output/tables/supp_tables.xlsx	Tables S.1–S.5 (workbook)	code/tables/generate_tables_s1_s5.py	All six supplementary tables collected in one Excel workbook, one sheet per table

Model-ID Mapping

The mapping of manuscript model IDs M0–M7 to internal run names is data/model_id_map.json; the eight runs are exactly those of the Monte Carlo reliability exam (data/model_performance_metrics.csv), and the same mapping is encoded in code/tables/table3.py, which verifies each configuration's certification rate and confidence interval against the regenerated metrics on every run. All tables, figures, and in-text statistics of the manuscript were verified against the data regenerated by this package.

References

Krasnow, Matthew D., Carolina Lopez, Jostin Kitmang, William J. Krasnow, Carolina Moreira Vasquez, and Ezequiel Molina. “Scaling up classroom observation with large language models: A reliability study of the World Bank’s Teach framework in Peru.”

World Bank. Teach 1.0: Helping countries track and improve teaching quality. Washington, DC: World Bank.