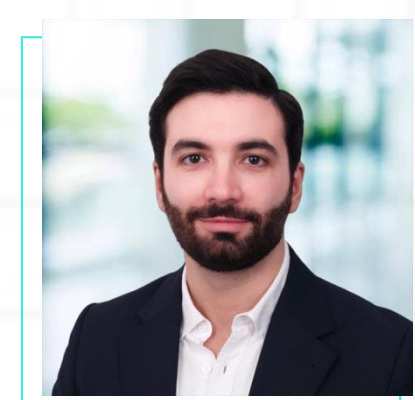
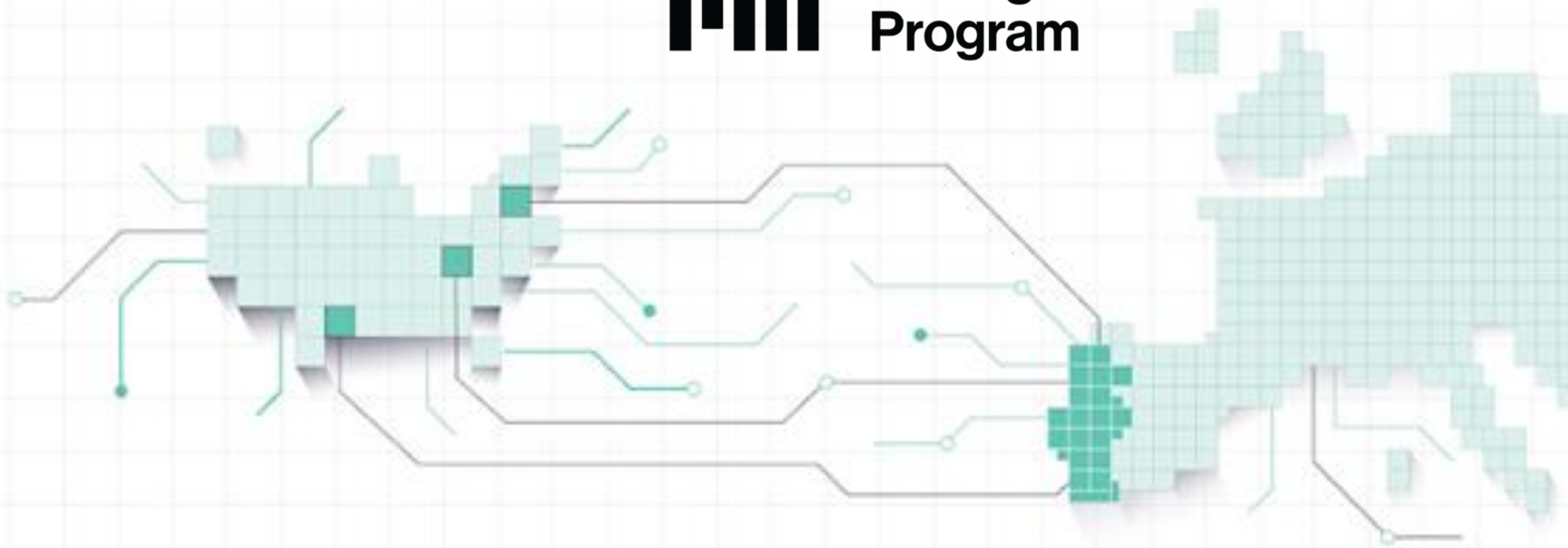


AI-Assisted Grading of Handwritten Exams in Higher Education: From OCR to Individualized Feedback



Leonardo Franco Cruz
leonardo.franco.cruz@tecnico.ulisboa.pt

Supervisor(s): Miguel Mira da Silva¹, Nuno Pombo²
1. Instituto Superior Técnico – Universidade de Lisboa
2. Universidade da Beira Interior



MIT Portugal Program Research Area
Advanced Computing and AI

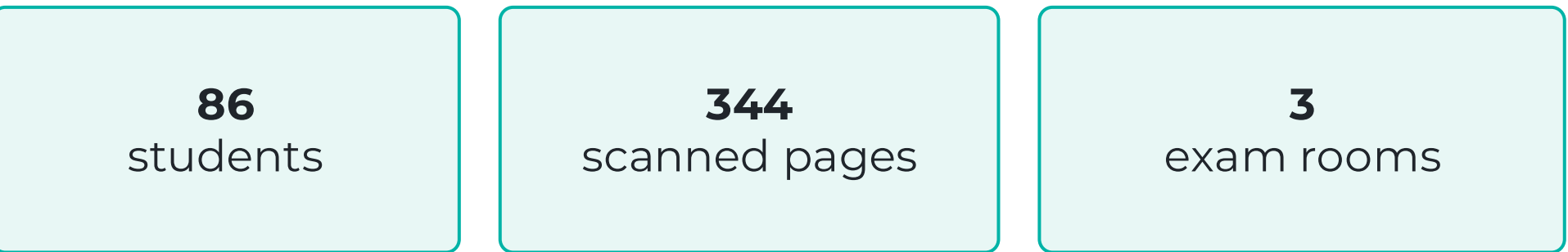
Research question and objectives

Can an instructor-supervised OCR and LLM pipeline grade handwritten exams on exam day while keeping direct student identifiers out of model inputs?

- Return criterion-level scores and individualized written feedback.
- Keep identity handling local while preserving instructor authority over release.
- Measure turnaround, score patterns and the manual intervention required.

Deployment context

Information Systems Project Management at Instituto Superior Técnico, 2026. A real cohort spanning six degree programs completed the handwritten exam in three rooms.



Privacy-by-design workflow

- Identify and anonymize**
Offline OCR reads each header, creates an anonymous submission ID and removes the header before downstream processing.
- Organize by question**
Anonymized pages are grouped into question packets with traceable local manifests.
- Transcribe**
Part A remains local; only anonymized open-answer pages are sent for model-assisted transcription.
- Grade**
Objective items use the answer key. Five open questions use six rubric criteria each.
- Review and release**
The instructor inspects a sample and all exceptions, then authorizes publication.

Privacy and governance controls

- Names, IST identifiers and header crops remain in a private local mapping.
- Direct identifiers are blocked in filenames, manifests and model prompts.
- Part A is processed locally; model calls receive anonymized Part B content only.
- Published grades are protected from automated overwrite and require instructor sign-off.

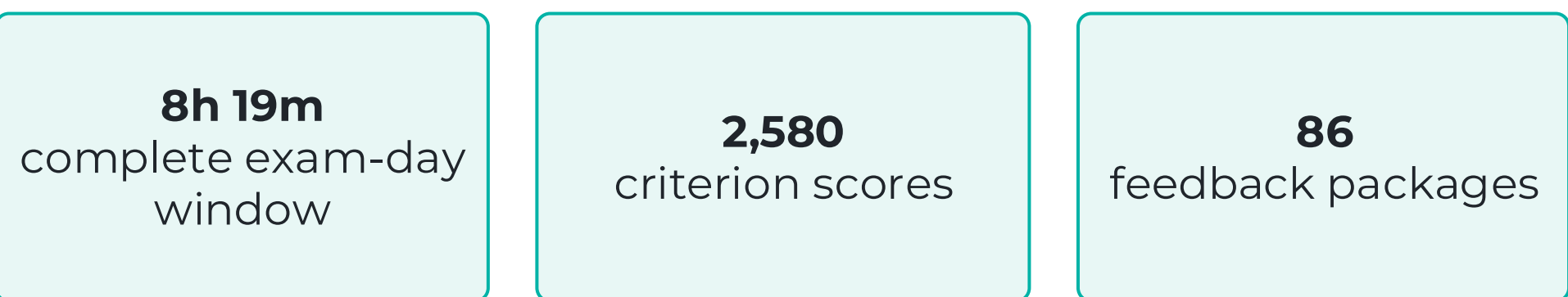
Assessment design

40% of the final course grade

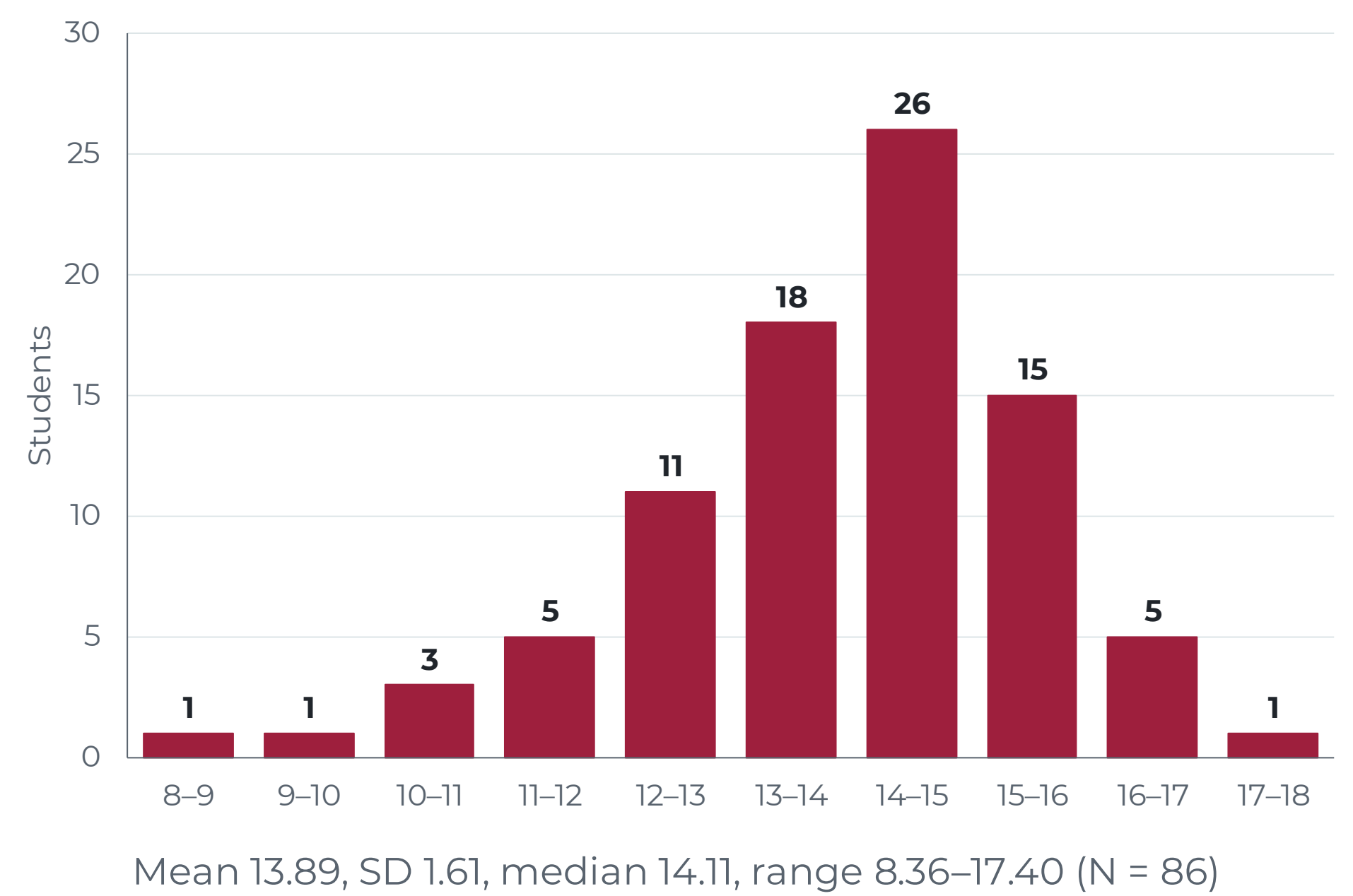
Part A contains 17 objective items worth 8.5 points. **Part B** contains five case-study questions scored on scope, theory, examples, personal opinion, organization and objectivity.

Exam grade (0–20)
Part A 8.5 + Q1 2.0 + Q2 2.0 + Q3 2.5 + Q4 2.5 + Q5 2.5
Every criterion included an AI-generated justification sentence. The instructor could use it to explain the score when a student asked.

Deployment results



Raw pipeline grade distribution



Grade bands		
Band	Students	Share
<10	2	2.3%
10–12.99	19	22.1%
13–15.99	59	68.6%
16–17.99	6	7.0%
18–20	0	0.0%

Instructor validation



The manual sample covered 15.1% of the cohort.
The sole exception recovered missing transcribed passages. It was not a scoring disagreement.
All scripts remained subject to exception checks before release.
An in-person exam and grade review session was open to every student.
No student challenged the published grade.

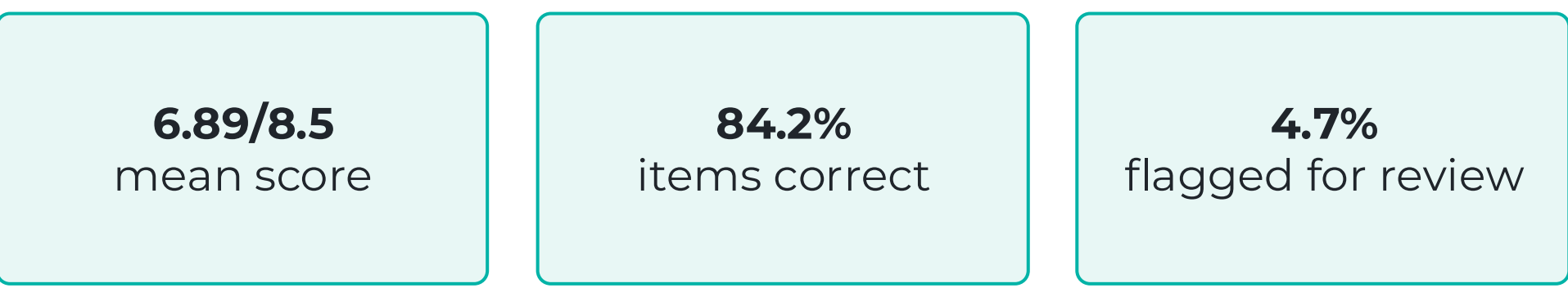
Post-review calibration

Raw mean 13.89. Published mean 14.95.
Average adjustment +1.06 points (SD 0.22; range +0.32 to +1.76).
The narrow spread indicates a near-uniform course-level calibration rather than systematic student-specific rewriting.

Operational impact

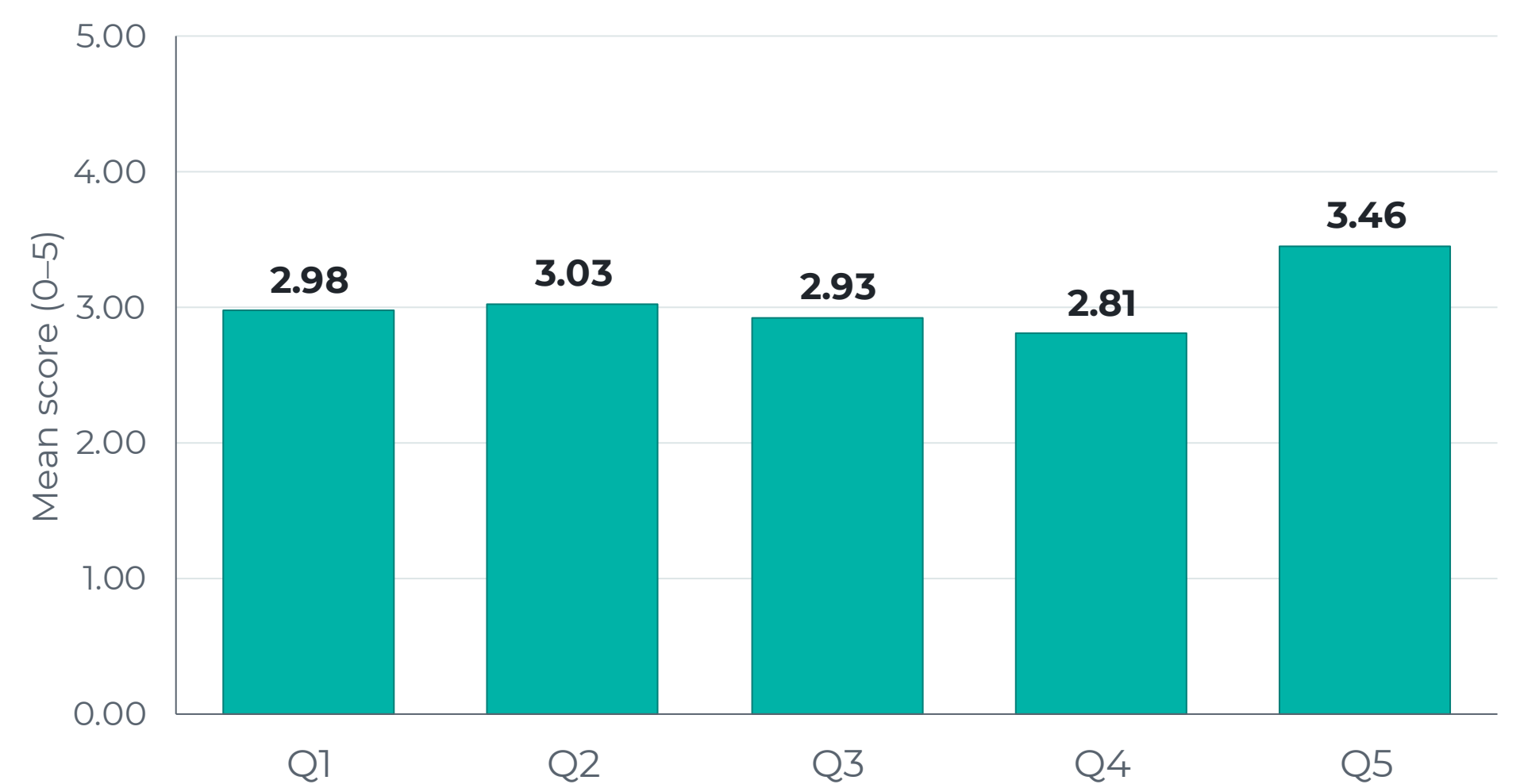
Same-day, criterion-level feedback was available for the full cohort.
The instructor redirected about 48 hours saved per trimester to supporting students with their course projects.
Instructor work shifted from full manual grading to targeted validation and exception handling.

Part A performance



Students answered 14.31 of 17 objective items correctly on average, with 0.48 blank responses. Four scripts were flagged for targeted Part A review.

Part B question results



Q	Mean	SD	Topic
Q1	2.98	0.53	Effect of backlog instability on sprint commitment and predictability
Q2	3.03	0.55	Shift from peer coordination to status reporting in daily stand-ups
Q3	2.93	0.56	Why retrospectives failed to produce continuous improvement
Q4	2.81	0.76	Effect of a remote product owner on backlog clarity and decisions
Q5	3.46	0.62	Impact of adopting Scrum ceremonies without its underlying values

Rubric signal

Q5 had the highest mean (3.46); Q4 had the lowest mean and the largest spread.
Criterion 3, examples and references, was consistently weakest: question means ranged from 1.63 to 2.13.
Objectivity was fixed at 5.00 by project configuration for all 430 open-answer evaluations and therefore had no discriminative power in this deployment.

Scientific contribution

- Demonstrates same-day feedback in a real 86-student higher-education cohort.
- Separates identity handling from model calls through local anonymization.
- Operationalizes instructor supervision before grades reach students.

Limitations and next steps

- Remove the fixed Objectivity score and validate a revised criterion that can distinguish answer quality.
- Retain direct scan access so reviewers can recover occasional transcription gaps.
- Benchmark against a larger blind sample of independent human graders and replicate in other courses.

Conclusion
Same-day first-pass grading was operationally feasible when privacy controls, traceability and instructor sign-off were built into the workflow.

Funded by:

