Carleton University

AI-Assisted Marking with Humans-in-the-Loop

A transparent, auditable workflow for large-scale engineering assessment

Leonard MacEachern, PhD, P.Eng.(APENS)
Associate Professor & Chair
Department of Electronics, Carleton University

Carleton University

Context and Motivation

1) Teaching context

- Fall-term co-instructor for a large engineering course
- Approx. 650 students across the course
- 10 labs per student

2) Resource constraint

- Historical TA allocation: 14 to 20 TAs
- Actual allocation that term: 10 TAs
- **I prefer to use TA time for student interaction**

3) Scale of marking challenge

- Each lab: 5 multiple-choice questions, marked conventionally
- Each lab: 5 long-form engineering questions requiring rubric-based judgment
- Long-form written responses across the term: about 32,500
- **Prototype question: can AI assist first-pass marking without surrendering human control?**

This was a workflow experiment under real teaching constraints, not a procurement recommendation.

Carleton University

Human-in-the-Loop Marking Workflow

5 multiple-choice questions were already marked automatically; the workflow focused on the 5 long-form questions.

1. Student submission
2. Anonymization by codeword
   - Rubrics were shared after submission
3. Rubric-based AI first pass
4. Semantic consistency audit
   - Semantically similar answers with divergent marks were flagged
5. Instructor verification in HTML review interface
6. Selective no-downgrade remarking
   - Marks were adjusted upward only during consistency review
7. Transparent release to students
8. Student remark request if needed

Architecture is model-agnostic and could be implemented with local models.

Carleton University

Boundaries and Safeguards

Where AI assistance fits

- Best for structured, verifiable responses
- Examples: formulas, derivations, numerical results, fill-in responses, and rubric-based engineering explanations
- Even complicated reasoning can often be checked against an explicit framework
- Reasoning is not the same as thinking
- Poor fit for creative writing or work intended to evoke human emotion

Safeguards for fairness and trust

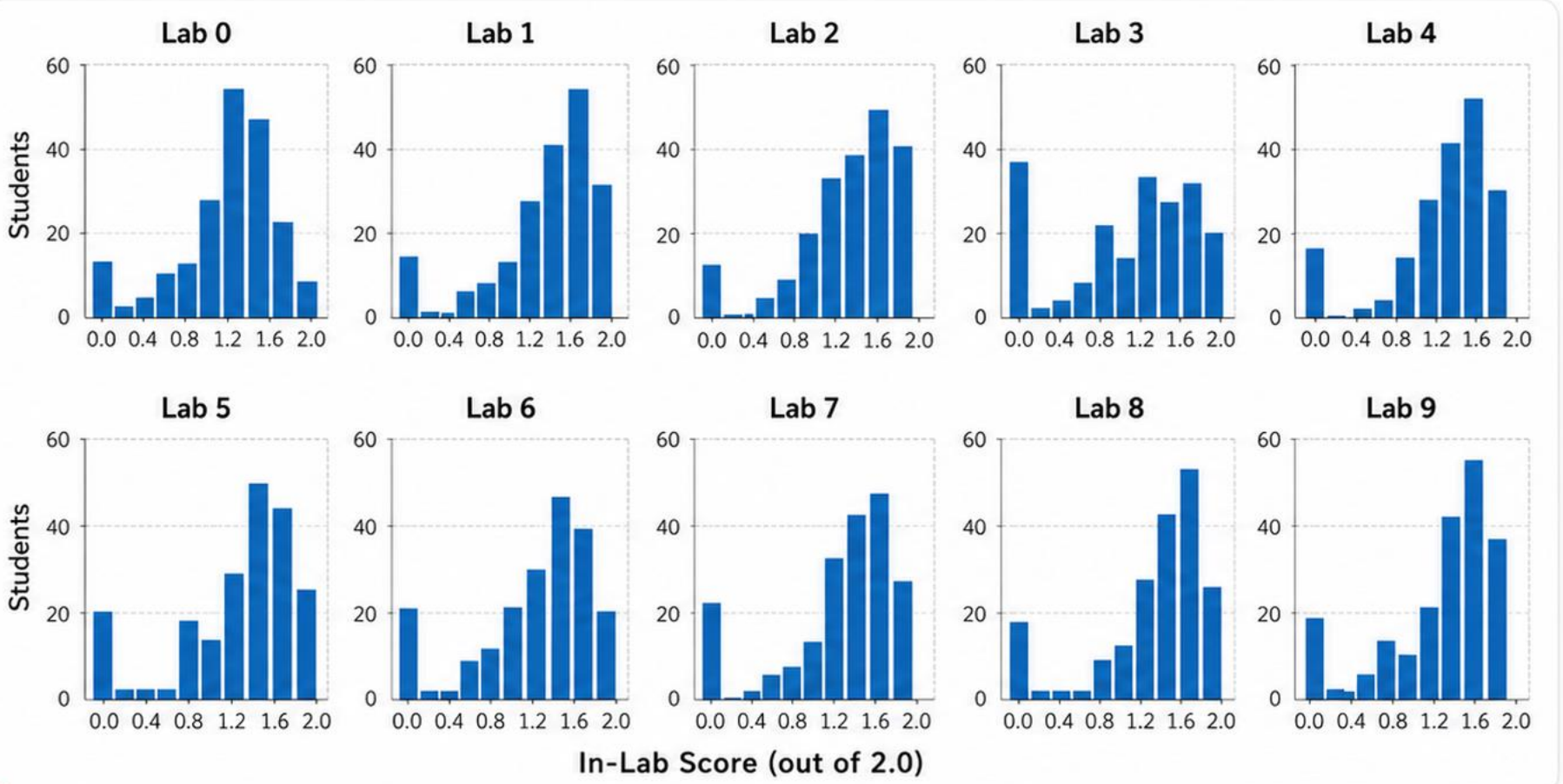
- Anonymized student work
- Detailed rubrics, like those given to TAs
- Independent semantic consistency checks
- Instructor review before release
- No score reductions during consistency re-review
- Transparent student-facing release using codewords
- Remark path for comparable cases

The model mattered less than the surrounding process: anonymization, rubrics, auditing, verification, and transparency.

Section-Level Results: In-Lab Histograms

Pre-lab marks excluded, example section n = 202

Carleton University



Observations

- Most students cluster between 1.2 and 1.8 / 2.0
- Lab 3 shows the widest spread and the most zero-score outcomes
- Later labs shift upward, with Lab 9 showing stronger high-end performance
- Distribution views help identify difficulty spikes, anomalies, and non-participation

Each histogram shows the number of students in each 0.2-wide score bin. Scores are out of 2.0.

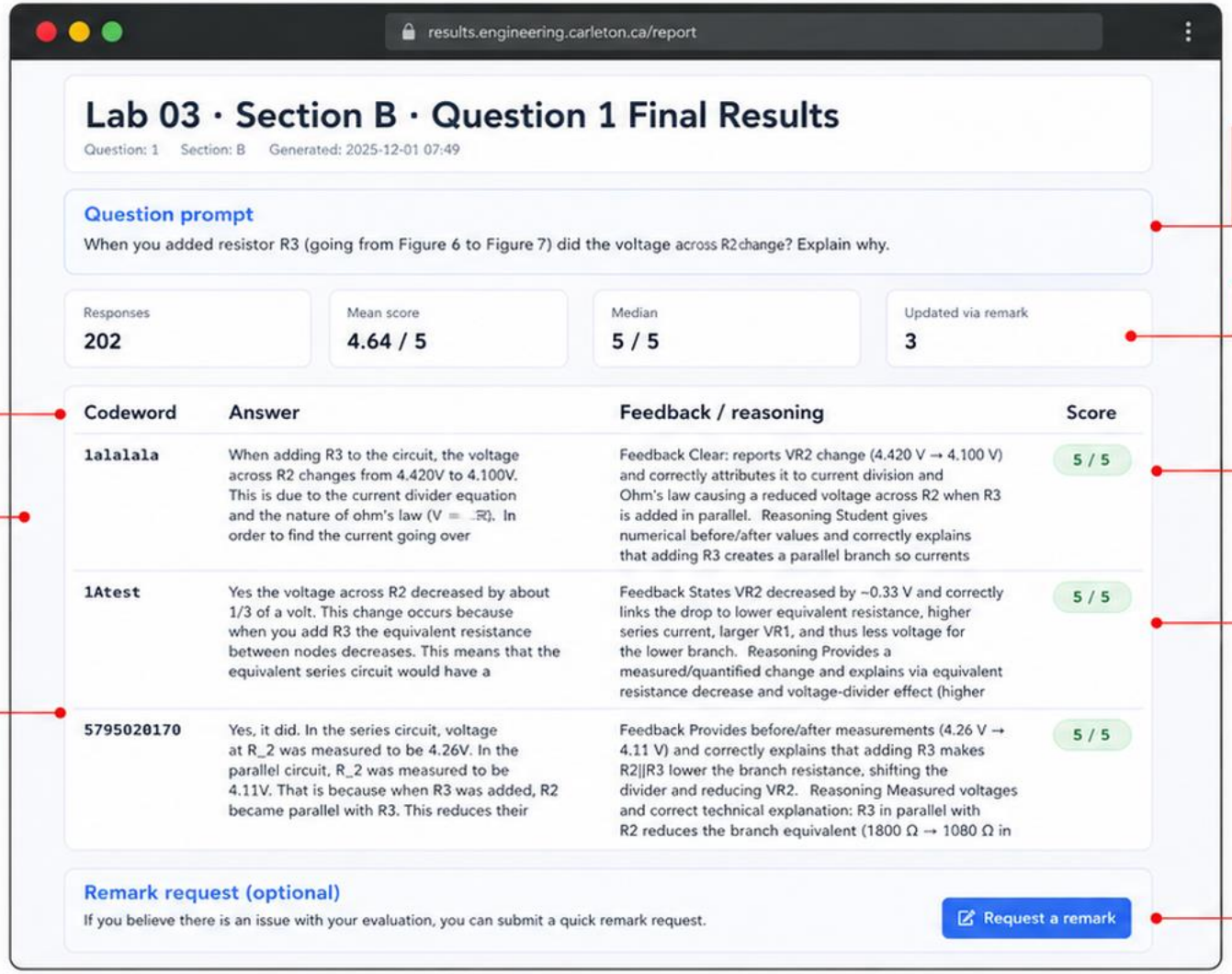
What Students Received

Carleton University

Anonymized by student-chosen **codeword**

Exact **student answer** shown

Students could **compare anonymized cases** for consistency



Exact **question prompt** shown

Updated-via-remark tags where applicable

Rubric-referenced **feedback** and **reasoning**

Final **score** clearly displayed

A simple form allowed **remark requests**

Transparency was used as a fairness mechanism, not just a reporting feature.

Carleton University

What I Learned

What worked

- AI assistance can work for bounded, rubric-based engineering marking
- The gain comes from shifting work from first-pass marking to faster human verification
- A cautious semantic audit can surface potential inconsistencies before release
- Transparency and clear rubrics improved trust and auditability

Caveats

- Not suitable for all forms of assessment
- Human judgment remains essential
- The process must include appeal or remark pathways
- Local institutional models may be preferable for governance reasons
- The surrounding workflow matters at least as much as the model

My conclusion: AI-assisted marking is viable for structured assessments when humans remain in control.