IMPERIAL

FLAIR
FEEDBACK AND LEARNING: AI-ASSISTED & REIMAGINED

Transforming GTA Feedback with LLMs:
Improving Clarity, Consistency, and Actionability (at Scale)

Zohaib Akhtar, Iro Ntonia, Stella Eva Tsiapali

Overview

A Human-in-the-Loop AI Approach to Feedback Moderation at Scale

- **Motivation:**

  Inconsistent, vague, and low-utility feedback across large, multi-marker cohorts

- **Our Approach:**

  Leveraging LLMs to support, standardise, and enhance GTA-written feedback — without replacing human judgement

- **System Design:**

  A three-stakeholder tool built for markers, students, and course leads

- **Results:**

  Evidence of improvement in feedback quality, consistency, and actionability

- **Limitations & Future Work:**

  Current boundaries of the tool and the road ahead

Motivation

Motivation

Why does feedback quality matter — and what's going wrong?

Feedback is one of the most powerful influences on student learning — yet one of the most inconsistently delivered.

"Feedback is among the most critical influences on student learning" — Hattie & Timperley (2007)

Students expect feedback that is: — Li & De Luca (2014)

Quality	What it means in practice
Timely	Received while the work is still fresh
Personal	Relevant to the individual student's performance
Explicable	Clear reasoning behind the grade
Criteria-referenced	Linked explicitly to the marking rubric
Actionable	Tells students what to do differently next time

Hattie, J. and Timperley, H., 2007. The power of feedback. Review of educational research, 77(1), pp.81-112.
Li, J. and De Luca, R., 2014. Review of assessment feedback. Studies in higher education, 39(2), pp.378-393.

Motivation
The Challenges We Faced

With 200+ students assessed by 6 GTAs, three key problems emerged:

Challenge 1: Variability in Feedback Quality

Same mark, very different feedback:

"Answered all but one question correctly and showed great understanding of both theory and lab. Did extra research on root cosine filters and presented relevant work well."

”Answered questions well; did Bonus 1, 2, 3."

Challenge 2: Feedback–Grade Mismatch After Moderation

Marks were moderated, but written feedback was not:

Feedback: "Excellent work" | Original mark: 90 | Moderated mark: 75

Motivation

The Challenges We Faced (continued)

Challenge 3: Student Scepticism Towards AI-Assisted Feedback

"Students who failed to identify the feedback provider correctly tended to rate AI feedback higher, whereas those who succeeded preferred human feedback" — Nazaretsky et al. (2024)

Scepticism driven by **lack of transparency** about AI involvement

Students had **no comparative insight** across markers

Can LLMs help us do better — at scale, without losing the human touch?

Nazaretsky et al., 2024. AI or human? Evaluating student feedback perceptions in higher education. ECTEL (pp. 284–298). Springer Nature Switzerland.

Our Approach

Our Approach

Goals

We set out to develop an LLM-powered tool with three core goals:

- **Goal 1: Improve Feedback Quality:**

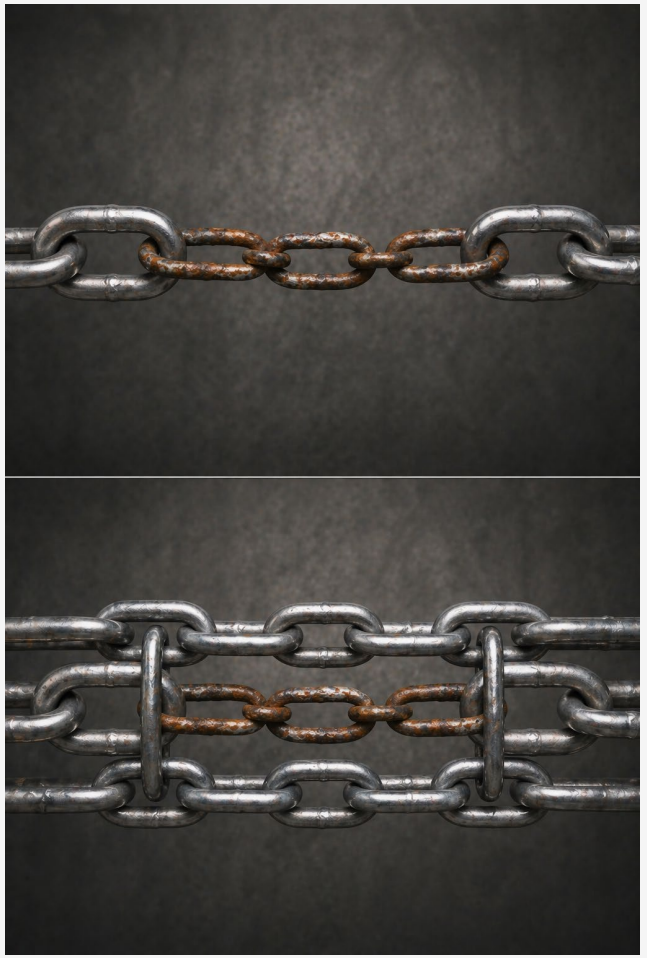
Enhance the clarity, tone, and actionability of GTA-written feedback — without replacing the human marker

- **Goal 2: Ensure Consistency Across Markers:**

Raise the floor of feedback quality so that all students — regardless of which GTA marked them — receive equally useful feedback - A chain as strong as the strongest links

- **Goal 3: Make Feedback Useful and Perceived as Such:**

Ensure students not only receive better feedback, but recognise and engage with it as genuinely helpful



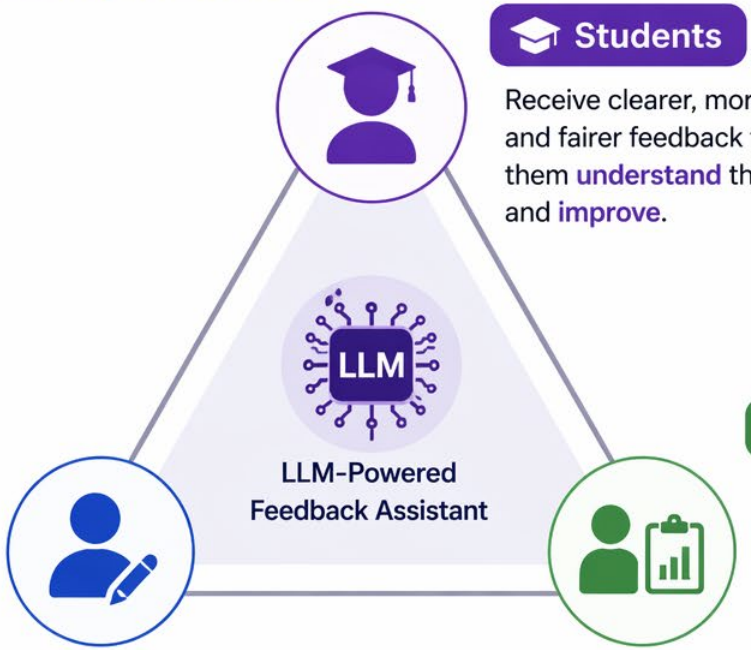
The LLM does not assess student work. It only improves the feedback written by human markers.

System Design

System Design

Who Is This Tool Built For?

The tool is designed around three key stakeholders, each with distinct needs:



Students

Receive clearer, more actionable, and fairer feedback that helps them **understand** their performance and **improve**.

Markers (GTAs)

Get real-time LLM assistance to improve the **quality** and **consistency** of the feedback they write — while retaining **full control** over the final output.

Course Leads

Gain **visibility** into feedback quality across the cohort, enabling **informed moderation** decisions and **systemic improvements**.

DESIGN PRINCIPLE:

The tool operates as a **human-in-the-loop** system — LLM suggestions are always subject to **human review, approval, and correction**.

System Design

How the Tool Helps Markers (GTAs)

- **Language Correction:**
  Automatically fixes grammar, spelling, and punctuation errors in written feedback
- **Rubric Alignment:**
  Prompts markers to provide a comment for each rubric criterion
  Ensures feedback coverage is complete and consistent across all markers
- **Tone Improvement:**
  Refines the tone of feedback to ensure it is constructive, professional, and appropriate for student reception
- **LLM-Generated Improvement Prompts:**
  Identifies weaknesses in the written feedback and prompts the marker to address them, for example:
  “This comment may be too vague — would you like to add a specific example?"
  "Would you like to add a concrete suggestion for improvement?"
- **Feedback for the Markers:**
  Detailed feedback is provided to the GTAs on the quality of their written feedback.

System Design

A Unique Capability: Comparative Feedback

Unlike GTA training, our tool can generate personalised comparative feedback for every student — automatically and at scale.

The Core Idea: Every student receives a contextualised comment that situates their performance relative to the cohort — based on their mark and the feedback written for higher and lower performing peers.

Step	What Happens
1. Calculate	Compute the cohort mean and standard deviation for each marking criterion
2. Compare	Identify where each student sits relative to the cohort
3. Generate	LLM writes a personalised comparative comment based on their position

Example Comments:

For a student performing above average:
"Your logbook was more detailed and better structured than most of your peers. This is likely why you scored above the cohort average on this criterion."

For a student performing below average:
"Students who scored higher on this criterion typically provided more detailed experimental analysis. Focusing on this in future assessments could significantly improve your mark."

WHY THIS MATTERS:

Gives every student **personalised context** for their grade

Highlights exactly **what higher performers** did differently

Motivates improvement through **concrete, evidence-based** comparison

Something that is **simply not scalable** through human marking alone

System Design

Usefulness for student

Students receive their grade and feedback through the platform, benefitting from three key improvements:

- **Structured, Rubric-Aligned Feedback:**
  - Feedback is organised around each marking criterion
  - Every student receives comments on every aspect of their performance
  - Provides a clear and concrete pathway for improvement

- **Consistent, Constructive Tone:**
  - Feedback tone is standardised across all markers
  - Language is professional, empathetic, and appropriate for student reception
  - Reduces the emotional impact of poorly worded or harsh comments

- **Dual Presentation of Feedback:**
  - Students see both the original GTA comment and the LLM-enhanced version:
  - Makes the marker's contribution visible and transparent
  - Builds trust by showing AI as a tool that enhances — not replaces — human feedback

Every student — regardless of which GTA marked them — receives feedback of the same quality, depth, and usefulness

System Design

Help Course Leads

Course leads gain unprecedented visibility into feedback quality across the entire cohort — before and after moderation.

- **Grade & Feedback Moderation:**
  - Apply grade moderation algorithms tailored to the module's needs
  - Feedback is automatically updated to reflect any moderated grade
  - Eliminates the disconnect between marks and written comments

- **Marker Insights:**
  - Aggregated metrics per marker, grade band, and rubric criterion, including
  - Number of actionable points per feedback comment
  - Tone and specificity scores across the marking cohort

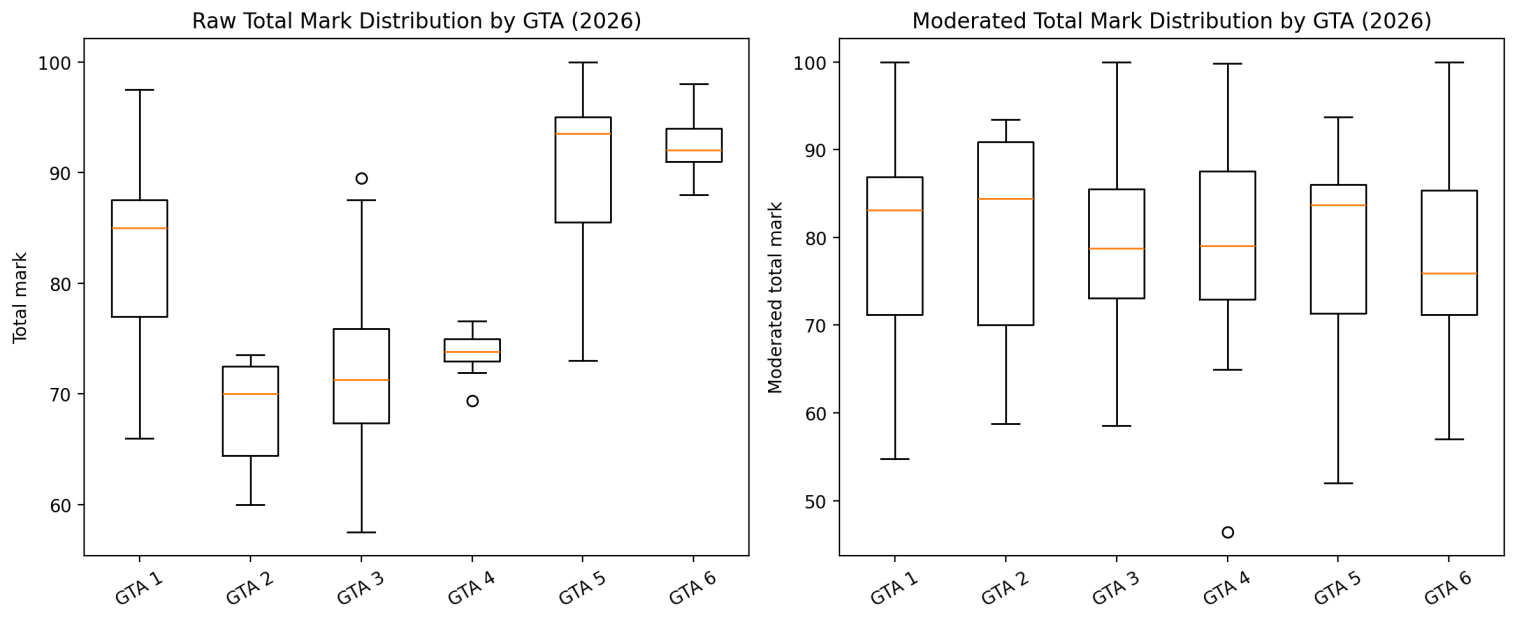
- **Anomaly Detection & Recommendations:**
  - Flags inconsistencies across markers, grade bands, and rubric criteria
  - Generates targeted recommendations, for example:
  - *"GTA 3 consistently writes fewer actionable points for lower-grade students"*
  - *“Students in the 60–70% band receive significantly shorter feedback on average"*

Results

Results - Comms Lab – Yr 2 - 2026

Marks Moderation

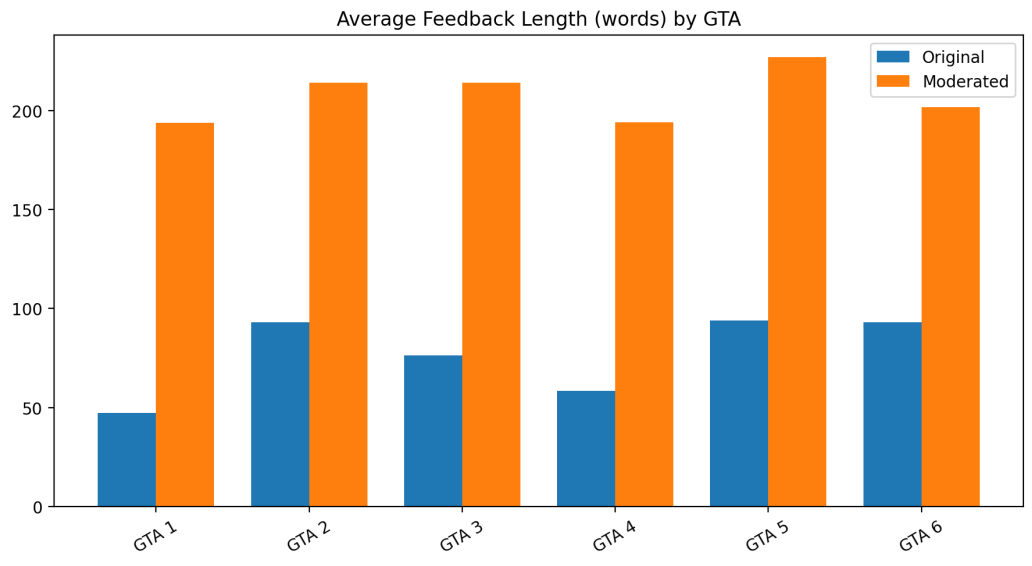
200+ students assessed by 6 GTAs

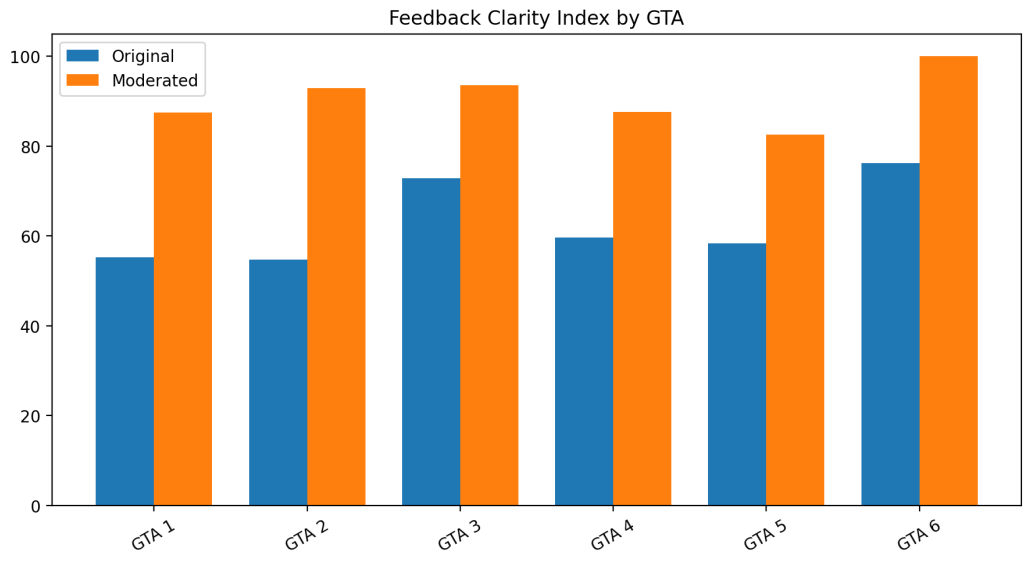


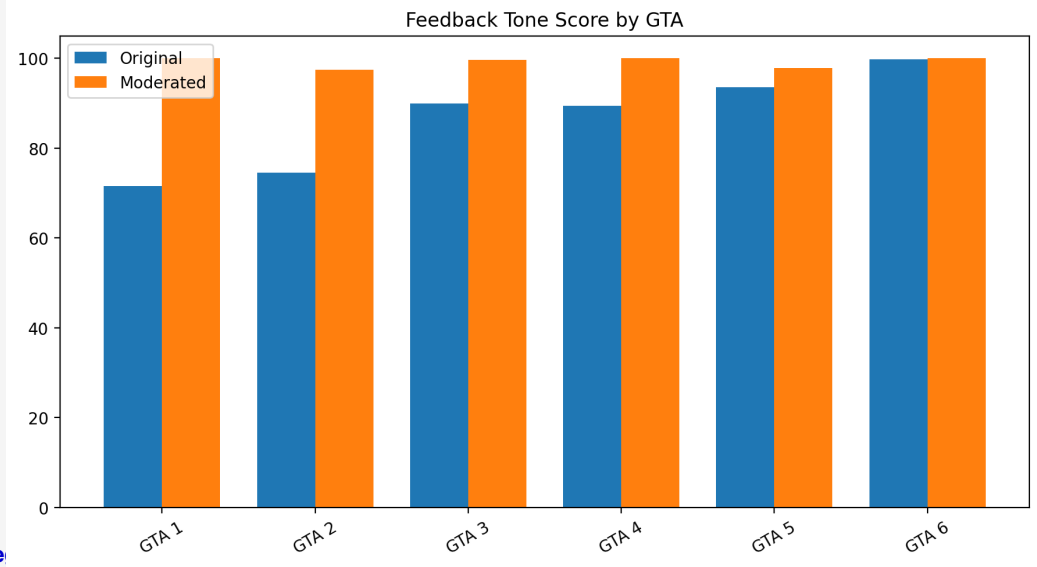
1. 🔴 GTA 6 required the largest downward adjustment (-12.6 points) — most generous marker
2. 🔴 GTA 2 required the largest upward adjustment (+12.0 points) — most strict marker
3. 🟡 GTA 4 showed extreme grade compression — std dev of only 1.6 points
4. 🟢 After moderation, all six GTAs align to the cohort mean of 79.8%
5. 🟢 Relative student performance is preserved — only the mean shifts, not the spread

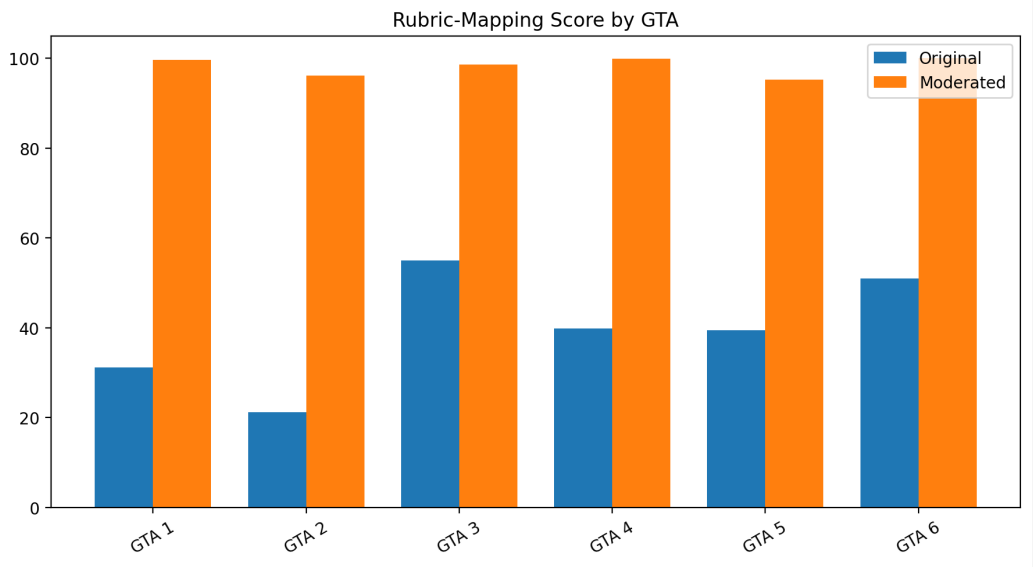
Results - Comms Lab – Yr 2 - 2026

Feedback Moderation









Results - Comms Lab – Yr 2 - 2026

GTA Feedback

GTA	Most Common Issues	Recommendations
GTA 1	Identical comments across groups; very short comments; lacks actionability	Add student-specific observations; include at least one actionable suggestion per criterion
GTA 2	Missing understanding comments; progress comments too brief; some comments too long	Ensure all three criteria are commented on; balance comment length
GTA 3	Spelling/grammar errors; informal language; identical group comments	Proofread before submission; personalise comments per student
GTA 4	Very narrow mark range; repetitive comments; overuse of "needs help from teammates"	Differentiate marks more; replace vague phrases with specific observations
GTA 5	Identical logbook/progress comments; marks very high relative to cohort	Personalise logbook and progress comments; calibrate marks against cohort
GTA 6	Highly formulaic comments; identical across students; lacks actionability	Add specific observations per student; include improvement suggestions

Limitations and Future Work

Limitations and Future Work

While the tool shows clear promise, there are important limitations to acknowledge:

- **Sparse Feedback Cannot Be Meaningfully Enhanced:**
  - When a GTA provides very little information, the LLM has insufficient material to work with
  - The tool can improve how feedback is written — but it cannot invent substance that was never there
  - *"Garbage in, garbage out"* — the tool is only as good as the raw feedback it receives
- **No Subject-Specific Training:**
  - The LLM is intentionally not trained on subject-specific material
  - This keeps the tool transferable across disciplines — but means it cannot verify the technical accuracy of feedback
  - The marker remains solely responsible for the correctness of subject-level comments
- **Risk of Perceived Inauthenticity:**
  - Although students interact with a human marker during the oral assessment, the written feedback may feel impersonal or AI-generated
  - This perception risk is real — even when the LLM is only enhancing human-written comments
  - Transparency about AI involvement is essential to managing student trust
- **Early Stage Evidence Base:**
  - Results currently reflect one lab assessment in one department
  - Broader conclusions require replication across multiple modules, cohorts, and disciplines

IMPERIAL

Thank you