188

Transforming GTA Feedback Consistency and Usefulness with LLM: Improving Clarity, Consistency, and Actionability

Zohaib Akhtar, Iro Ntonia
Imperial College London, London, United Kingdom

Research Domains

Learning, teaching and assessment (LTA)

Abstract

This paper addresses a key challenge in higher education: ensuring students receive consistent, clear, and pedagogically useful feedback, particularly in large cohorts assessed by multiple Graduate Teaching Assistants (GTAs). At Imperial College, GTAs provide feedback for over 200 students in lab oral assessments, making it difficult to ensure quality and uniformity across the marking cohort, while moderation currently only takes into account final marks rather than quality and consistency of feedback Here, we report on our recent initiative to employ a Large Language Model (LLM)-powered tool to enhance the clarity, consistency, and actionability of GTA-generated feedback. The tool uses Natural Language Processing (NLP) to refine language, improve tone, align comments with rubrics, and promote feedforward and reflection. After testing the model in a two-year pilot study examining cohort-wide performance patterns, we present evidence for the tool's ability to deliver high-quality feedback and improve specificity, transparency and motivational value.

Full paper

Providing pedagogically useful feedback across large cohorts remains a persistent challenge in higher education. At Imperial College London, Graduate Teaching Assistants (GTAs) play a major role in delivering feedback; the consistency in quality and actionable feedforward value of GTA feedback has however frequently resulted in student complaints and dissatisfaction. Our GTA assessors do not undertake any formal training in feedback best practice, which in turn results in lack of motivation and underestimation of the importance of iterative, actionable feedback as standard. To tackle this problem, we deployed a large language model (LLM)-powered tool, tuned to improve and standardise the quality of feedback delivered specifically by GTAs. The tool was designed to enhance actionability of feedback delivered, encourage student reflection, and moderate the tone of language used. In this paper, we use lab-based modules in Electrical and Electronic Engineering as a typical example of the disciplinary and institutional context. In our second-year Communication Lab modules, 200 students are assessed by 7 GTAs. While we

have established processes for moderating marks, we previously had no practical method to moderate the written feedback each student received. Each GTA assessed a different subset of students, which made it difficult to establish consistency, clarity, and fairness in feedback across the full cohort. GTA feedback often varied in tone, specificity, and alignment with marking rubrics and learning outcomes. Crucially, it lacked a comparative element, offering students no insight into how their work compared to others in the cohort. This undermined the educational value of feedback and limited its impact on student learning, particularly when students were not equipped to interpret vague or inconsistent comments (Carless & Boud, 2018).To address these challenges, we piloted a Large Language Model (LLM)-powered tool to support moderation and enhancement of GTA-generated feedback. The tool uses Natural Language Processing (NLP) to analyse and refine raw feedback comments. It corrects language and structural issues, improves clarity and tone, aligns comments with intended learning outcomes, and adds actionable suggestions. It also reduces GTA workload by streamlining the feedback process while increasing its pedagogical usefulness.

Student feedback to date reports on students wanting to get a meaningful sense of where they 'stand' in relation to the rest of the cohort. Ensuring this information is fed back to them in a pedagogically-meaningful way is notoriously difficult, as lack of specificity and clear feedforward suggestions often leaves students feeling anxious, lost and demotivated. We posit that a key novelty of our tool is its ability to generate such meaningful cohort-comparative feedback. Unlike traditional moderation practices, this LLM can compare performance across the full cohort and incorporate peer-based insights into individual feedback. For example, a student receiving a mid-range mark might be told: "Compared to the rest of your cohort, your explanation of the signal processing steps could be more clearly linked to the theoretical concepts introduced in Lab 2." This comparative element adds motivational and diagnostic value that GTAs alone cannot offer at scale and could enhance student self-efficacy (Bandura, 1977), planning (Nicol, 2010) and self-regulated learning (Zimmerman, 2002).

In our pilot evaluation we conducted a textual analysis comparing original GTA-written comments with LLM-enhanced versions. The analysis examined five key areas: feedback length, tone, specificity, actionability, and alignment with assessment rubrics. Specifically, we found that LLM-enhanced feedback was, on average, 40% longer, reflecting more detailed and coherent explanations without excessive verbosity.
Additionally, the proportion of positively framed comments increased by over 60% following moderation, shifting from often neutral or critical phrasing to supportive, constructive language. Originally, only about 50% of original comments contained specific suggestions for improvement. After LLM moderation, this rose to 100%, ensuring every student received clear guidance on how to improve. While none of the original GTA-written feedback included peer comparison. With the LLM, 100% of comments now include personalised comparative statements, offering students insight into how their work differs

from that of higher-performing peers. Finally, references to intended learning outcomes and assessment criteria increased by 70–80%, resulting in more structured feedback.

Our planned next steps involve scaling the pilot across departments, formal integration into GTA training, and additional evaluation using student surveys and focus groups, to specifically explore how this is received by students and whether it influences their engagement, confidence, and performance over time.

We envisage that in the long term, this tool will offer a practical model for AI-supported feedback moderation at scale in line with recent recommendations for designing feedback processes that students can act on (Winstone & Carless, 2019).