Coaching AI to be a Team Player

A. A. Tomlinson[1], N. K. Wilkin[1], M. Kainth[2], and R. Stanyon[2]

[1] *University of Birmingham, B15 2TT, UK*
[2] *Graide, UK*

Abstract

In this paper, we explore how an 'AI marking engine' (*Aime*) can be effectively, and ethically, incorporated into an assessment and feedback cycle with an agent-based model. We consider *Aime* as an additional team member who needs to be on-boarded. In the scenario we explore, *Aime* sits in the education hierarchy alongside the GTA graders while the overall decision making and moderation remains with the module lead. We implement, and comment, on the expectations of the humans, and the human-AI interactions via Graide, a subscription platform, that incorporates an AI engine which is optimised for mathematical disciplines. We propose that by treating the AI engine as a persona with a well-defined role and reporting structure in the human- AI ecosystem, one can optimise the grading and, most importantly, the individual student experience where the goal is to rapidly provide personalised, in-depth, and consistent feedback.

Keywords

Human-centered computing, computer-aided assessment, instructional coaching

1. Introduction

In order for AI to be an effective tool in education, there is an acknowledgement that new protocols are required in order that the AI both enhances the workflow and works with the human team [7]. There is understandably a distrust in AI and concerns around the transparency and accuracy of decision making [5, 6]. There is also a clear need for the proliferation of AI tools to support teaching and learning for large cohorts of students where individuals grading work struggle to return consistent and timely feedback [12]. Within the context of AIED, it is crucial that grading precision and the quality of feedback generated by AI assessment tools is reliable and similar in depth and nuance to that expected of an expert human grader. Retaining human oversight can provide assurances to leadership and frontline educators that the grading and feedback process is at least as fair and accurate to students as traditional processes. Benefits include meaningful pedagogical insights alongside significant time and cost savings. In this paper, we discuss how an AI marking engine (*Aime*) can fit into existing grading workflows to enhance all aspects of assessment for students and educators. We consider *Aime* to be a member of the team, who is 'coached' to become fully-integrated with carefully demarked responsibilities. We discuss the design of the handover points between humans (students, module leads, and Graduate Teaching Assistants (GTAs)) and *Aime*. In particular, we highlight the controls differentially owned by human users and how this is leveraged to speed up the grading process while maintaining confidence in the AI decision-making. We will exemplify this in the context of an undergraduate mathematical subject.

Proceedings Name, Month XX–XX, YYYY, City, Country
EMAIL: a.a.tomlinson@bham.ac.uk (A. 1)
ORCID: 0000-0002-1767-8032 (A. 1); 0000-0001-8073-7173 (A. 2); 0000-0002-5906-3888 (A. 3); 0000-0003-0995-4559 (A. 4)


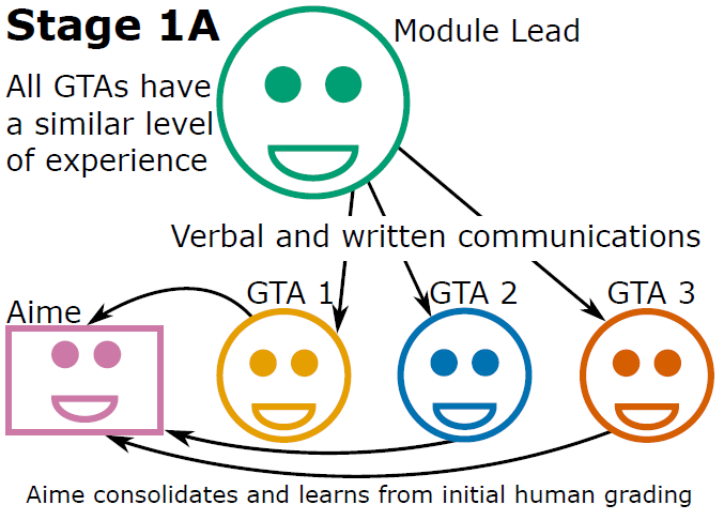

Figure 1. The first stage of the grading process is led by the module lead who communicates with their GTAs to set grading expectations for the assignment.

The platform we have chosen for our ecosystem is Graide. As showcased at L@S2022 [10], Graide, is a now commercialised grading and personalised feedback system which is a spin-out from the University of Birmingham, UK. This close working between academics and the system developers has enabled this model to develop nimbly.

However, the underlying methodology of ascertaining what roles *Aime* is empowered (or prevented) from undertaking transfers to any grading ecosystem that has access to an AI engine that has been made suitably adaptable to work with and for the humans, rather than as a central system with which the humans must learn to engage

1.1. Including *Aime* in an idealised grading workflow for large undergraduate cohorts

In the team hierarchy that we have implemented in Fig. 1, the humans always retain overall moderation control, intentionally implemented as a safety feature. In stage 1, as with traditional grading of large cohorts, there is written and verbal communication and discussion between the module lead and the GTAs (for simplicity, we consider 3 GTAs in the figures shown). The GTAs then commence grading, to the best of their ability, without an implied time constraint per script.

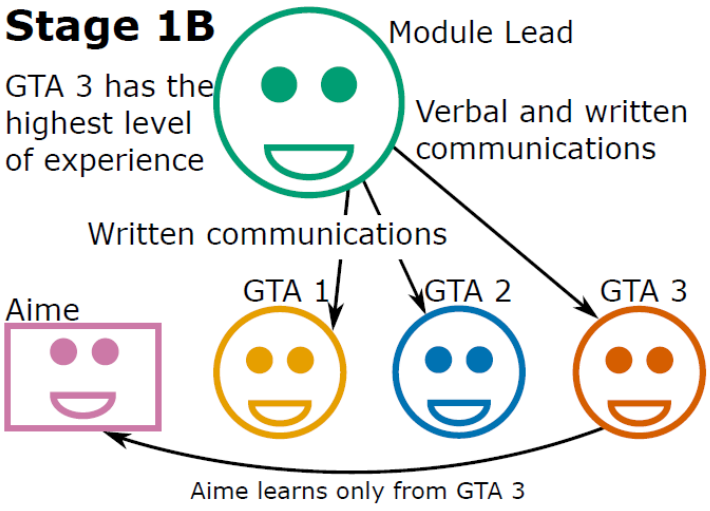

Figure 2. Alternative to stage 1A, the hierarchy of experience in the GTAs can be exploited to benefit from their experience.

In a non-AI scenario, each GTA grades work individually and, without a carefully agreed-upon mark scheme or rubric, graders can produce inconsistent feedback [3]. In addition to this, as graders become more familiar with a question paper, their grading speed should increase as they proceed, potentially leading to a reduction in the depth of feedback on individual scripts as the grader 'gets into the swing' of grading. This increase in speed or cumulative grading time could cause a grader to become careless or fatigued leading to variations in their decision making [9]. From an individual student perspective, this traditional grading set up leads to an appearance of inconsistent care and attention when comparing their feedback.

Aime sits alongside the GTAs and initially learns from the human grading they undertake. GTAs grade work and *Aime* rapidly predicts feedback they are likely to apply to the next submission with a similar algebraic structure. Feed- back is then suggested to the grader (with an associated confidence level) which is then reviewed and, ideally, confirmed. If the grader rejects the suggested feed- back, the AI model adapts to provide refined feedback on the next approach. Work by Benton has shown that the accuracy of grades produced by a collective team of graders is better than a single senior grader [1]; *Aime* facilitates a collective approach to grading work by adapting to the entire scope of the feedback applied by the team of graders.

For large student cohorts, the team of GTAs will be of significant size itself, often with a designated lead GTA, who has prior experience of working on the module and a deeper familiarity with the material. In this scenario (in a non-AI setting), the lead GTA (GTA 3 in Fig. 2) will be the one who liaises with the module lead. They would then instruct or supervise the other GTAs, potentially by physically co-locating to grade.

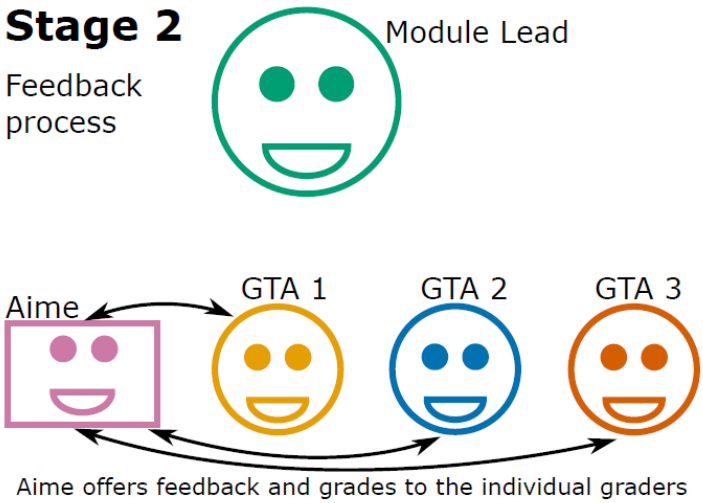

Figure 3. The second stage of the grading process is led by the GTAs and supported by *Aime*. Given the prior training, *Aime* can now suggest helpful feedback to the GTAs and further refine the suggestions when they encounter new methods. If a GTA makes a poor suggestion, *Aime* can unlearn this when a different GTA modifies the feedback when it is suggested to them.

The process can be streamlined with *Aime* in the team. GTA 3 grades scripts first, such that *Aime* only learns from GTA 3. Once GTA 3 is satisfied that the feedback coverage is sufficient for the standard solution paths students offer, the other GTAs can commence grading with high quality feedback offered to them to select for the students. In effect, via *Aime*, the GTAs are being coached 'on the job' to become as good as the senior GTA.

Stage 2 of the process is where the bulk of the scripts are graded. There is a choice to be made in terms of workload distribution between the GTAs. One can either allocate groups of students to each GTA or give all the answers to a particular question to a GTA. The second option provides greater consistency at a question level, but the whole script approach enables overall synoptic feedback to be additionally offered.

With *Aime* as part of the team, the same feedback is offered to each GTA (Fig. 3) as they review similar answers from different students. Similarly, if a grading error is uncovered by a GTA, this is

comments are then easily added by GTA 3, where these overall comments themselves are ensured to be consistent via *Aime*'s support.

Stage 3 is a moderation process. In a traditional non-AI set up with finite time, and a large cohort of students, this will only be a statistical spot test of scripts, with a follow up of remarking if significant discrepancies are uncovered. Particularly for in-semester class tests or similar, where rapid feedback to students is also a driver, there is a need for feedback for student learning, rather than just a statement of attainment to date. Further discrepancies (perceived or real) will also be picked up by students comparing their feedback with that of their peers. This latter issue is a particular concern in the UK where the National Student Survey (NSS) has an entire section set on assessment and feedback [8]. Students are asked a series of questions where they reflect on the quality of the provision they receive from their HE provider. Findings are taken seriously by institutions and can impact league table positions and prospective student choices [11]. With the ideal *Aime* in the team, the module lead can interrogate all scripts and *Aime*'s narrative feedback as a summarised visual report including average and individual performance. The module lead can then apply moderation (for instance: reallocating marks between sections or changing the sentiment of a narrative comment) to all scripts simultaneously.

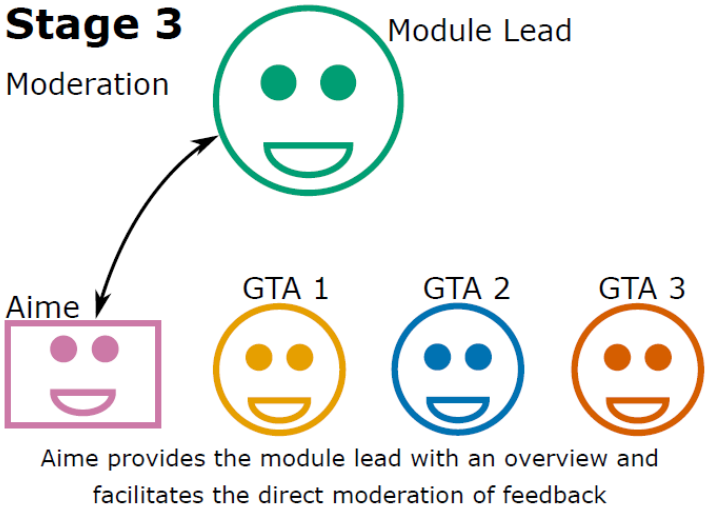

Figure 4. The third stage of the grading process is led by the module lead. *Aime* provides a high-level overview of the grading including statistics and feedback trends. The module lead may then moderate the feedback at the individual student level, if required.

We have observed, as both GTAs and instructors, that stage 4 is frequently only given cursory attention. It is the closing of the feedback loop between the module lead and the GTA, where improvements to questions for later years, and discussing aspects that students have consistently found difficult, should happen. This is a human-only discussion, but informed by the visual summary that *Aime* provides.

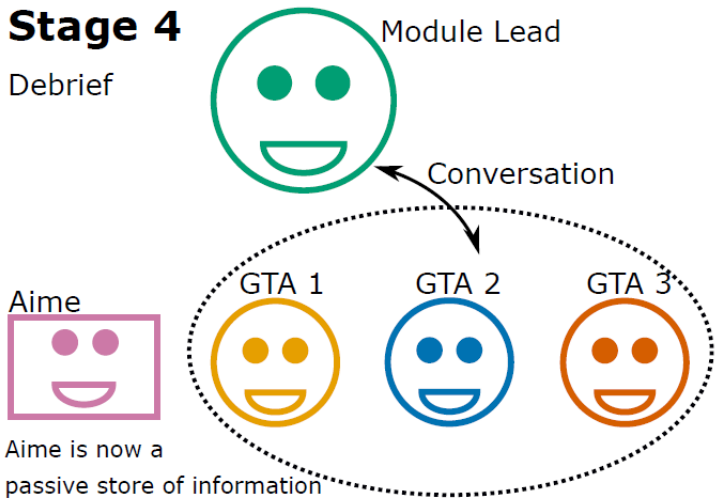

Figure 5. The fourth stage of the grading process is led by the module lead in the form of a supportive discussion with the GTAs. *Aime* is not involved at this stage as the loop is closed by and with the humans.

To note, so far we have coached *Aime* to become a member of the grading team. We have carefully defined *Aime*'s role at the different stages of the grading process and who they are learning from, recommending information to, and reporting to. It should be emphasised that none of this section is dependent on either a particular platform, or the discipline.

In summary, when incorporating *Aime* into an existing workflow, we should carefully consider the requirements of each stakeholder to ensure they are all met. *Aime* needs to integrate into this network of stakeholders and improve the grading experience for everyone to justify implementing such a system. In the theoretical framework that we have outlined above, this is visibly true for all except the administrators. This group is dependent on having well-designed Application Programming Interfaces (APIs) of both the Learning Management System (LMS) at their institution, and *Aime*. Although important, it is a technical issue and out with the discussion of this paper.

2. Incorporating *Aime* into the team: a proposed case study

So far we have described the team interactions in an assessment and feedback ecosystem that include an *Aime*. *Aime* is, ideally, agnostic to the type of assessment and completely capable for all tasks required.

The case study focuses on mathematical subjects, the predominant area in which we have taught and graded as subject-matter experts. At scale, this often comprises aspects of pre-calculus and calculus, where students are required to produce a moderate amount of working to solve a problem. This is commonly assessed via problem sheets and closed book examinations which are typically hand-graded by a team of GTAs or the module lead. Previous attempts to in- crease the rapidity of marking have diminished the personalisation, and depth of feedback. Hahn et al. report that automatic feedback systems risk reducing human interactions and could reduce the amount of personalised feedback given to students who need it [4]. Instead, using a grading team that includes *Aime* has enabled us to change the workflow to facilitate grading traditional assessments more efficiently, while providing additional training opportunities to GTAs.

Students need: Fast, helpful, and friendly feedback; opportunities to submit formative work in advance; authentic assessments to prepare them for the future.

Graders need: Clear grading schemes and the ability to grade anywhere; peer communities of practice to develop professionally; support, engagement, and praise from module leaders.

Module leads need: to facilitate timely and consistent grading; areas of student strengths and weaknesses to be highlighted; happy students, even when the work is challenging.

Administrators need: Easy integration into an existing LMS; simple moderation tools to view and access submissions and grading; accessible grade books for data processing.

Figure 6. Key stakeholder needs for *Aime*

In the case study proposed, *Aime* is Graide, which itself has been designed to be a team player who always follows the rules and guidance and defers to humans for final decisions. To date, the focus has been on ensuring that Graide is optimised for the key stakeholder interactions, as shown in Fig. 6.

To enable interaction with the submitted scripts, they need to be digitised, but can be initially handwritten. It could be tempting to require students to learn and become proficient in digital equation input, although this would add significant time and cognitive load to the assessment, inadvertently changing the learning outcomes away from solving problems and towards computer input and required syntax [2]. By requiring students (or administrators, depending on the assessment) to upload scans of work, solutions can be digitised and interpreted by Graide. This means that educators can continue to develop their assessments using well-understood and accepted methods (only scanning is added to the workflow) without having to consider additional complications of a new system. Hence Graide does not require alterations in traditional workflows in order to be incorporated as a team member and we are able to implement stages 1-4 of the team marking that were described in the previous section for real assessments.

To contextualise this, we will consider a scenario with a typical question set in a first year undergraduate mathematics module for physics students. Considering a single problem (where the details of the mathematics are not particularly important):

$$\text{Differentiate } \frac{1}{\sqrt{x}} \tag{1}$$

This is considered a straightforward problem at this level and many students should be able to correctly answer it in one or two lines of working. Assuming we have 150 students and 3 GTAs, stage 1A would begin with the module lead informing the GTAs of the problems to be marked and associated deadlines for returning feedback to students. They would share a set of model solutions so the GTAs can inspect the standard method and calibrate how they award marks. In this case, this would be the first time the GTAs have worked with *Aime*, so the module lead gives a further briefing on what to expect.

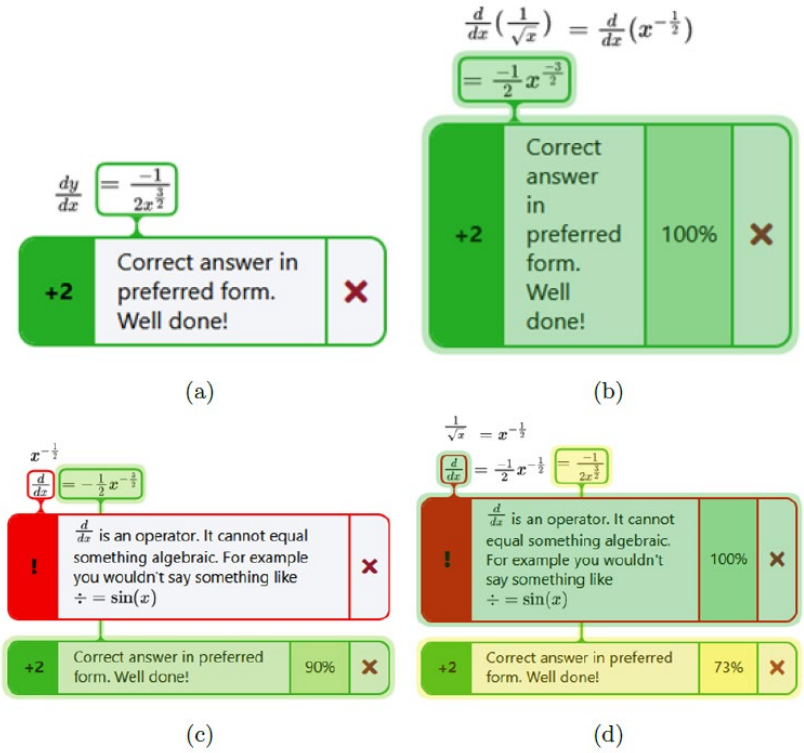

Figure 7. Examples of the grading process. Feedback with a percentage confidence is suggested by *Aime* and confirmed by a GTA, feedback without a percentage confidence is given by a GTA. (a) *Aime* is initially trained by GTA. (b) *Aime* suggests learned feedback to another GTA. (c) A GTA notices an error in a student's notation and further trains *Aime*. (d) *Aime* suggests the constructive feedback to another GTA when the notation error is re-encountered.

Assuming all the GTAs have a similar level of experience, one of them (say, GTA 1) will likely start stage 2 of the grading before the others and begin training *Aime* on their allocation of 50 pieces of work (Fig. 7a). *Aime* quickly starts suggesting feedback to GTA 1 and the other GTAs when they start grading (Fig. 7b). GTA 1 notices that a few students use the notation incorrectly and instructs *Aime* to provide some constructive (but not punitive) feedback for students (Fig. 7c). When GTA 2 starts grading, *Aime* automatically suggests this feedback to them, modelling good feedback practice and ensuring that no penalty is consistently applied to the overall mark for that particular error (Fig. 7d).

Once the GTAs have finished marking, the module lead begins stage 3 and inspects the grade distribution and the scheme of feedback for a general overview. They notice that a GTA has written some unsupportive feedback so decides to modify the tone of it before releasing it to students (Fig. 8).

Figure 8. (a) The module lead notices a GTA has produced an unsupportive piece of feedback. (b) The module lead updates the tone of the feedback to be more supportive and *Aime* updates all the work it was applied to.

The module lead, satisfied with the quality of the feedback, meets with the GTAs for stage 4. They praise them for their prompt return of marks and discuss the modified feedback and why a supportive tone is more helpful to encourage students to learn from their mistakes. The GTAs take this on board for the next assessment cycle and the module lead is reassured when *Aime* shows more supportive feedback the next time.

All of these experiences at the different stages have been implemented and the planned case study will observe the full set of stages and hand-over points in detail.

3. Conclusions

In this paper, we have considered the AI marking engine, *Aime*, as a team member, rather than a platform that one interacts with. Through this view of the whole team, we have been able to ensure that *Aime* fits into the team hierarchy and that the human members of the team always retain the decision making authority. Detailing the key aspects of the ecosystem in this way will be helpful to institutional providers who need reassurance of the guide rails that have been put in place to ensure that humans have the ultimate control of the moderation processes.

By drawing on examples from our use and development of Graide, we have demonstrated that this viewpoint of the grading team with *Aime* as a member, is relevant to a commercially available system. Scenarios beyond mathematics require different technical specifications for *Aime*, for instance narrative or figures. Furthermore, recent extensions to Graide, in beta, enable short answer text questions, but there is still much to do in implementing this paradigm with *Aime* as a team member for assessments across the breadth of disciplines that are actively assessed via digital or paper submissions.

4. References

[1] Benton, T.: Which is better: one experienced marker or many inexperienced markers? Research Matters: A Cambridge Assessment publication 28, 2–10 (2019).

[2] Chan, C. K. Y.: A systematic review–handwritten examinations are becoming outdated, is it time to change to typed examinations in our assessment policy?. Assessment & Evaluation in Higher Education, 1-17 (2023).

[3] Popenici, S. A., Kerr, S.: Exploring the impact of artificial intelligence on teaching and learning in higher education. Research and Practice in Technology Enhanced Learning, 12(1), 1-13 (2017).

[4] Hahn, M.G., Navarro, S.M.B., Valentin, L.D.L.F., Burgos, D.: A sys- tematic review of the effects of automatic scoring and automatic feed- back in educational settings. IEEE Access 9, 108190–108198 (2021).

[5] Kaplan, A., Haenlein, M.: Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business horizons 62(1), 15-25 (2019).

[6] Omrani, N., Rivieccio, G., Fiore, U., Schiavone, F., Agreda, S. G.: To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. Technological Forecasting and Social Change, 181, 121763 (2022).

[7] Popenici, S.A., Kerr, S.: Exploring the impact of artificial intelligence on teaching and learning in higher education. Research and Practice in Technology Enhanced Learning 12 (2017).

[8] Richardson, J. T., Slater, J. B., Wilson, J.: The National Student Survey: development, findings and implications. Studies in Higher Education, 32(5), 557-580 (2007).

[9] Sadler, D. R.: Indeterminacy in the use of preset criteria for assessment and grading. Assessment & Evaluation in Higher Education, 34(2), 159-179 (2009).

[10] Stanyon, R., Tomlinson, A. A., Kainth, M., Wilkin, N. K.: Providing individual student feedback at scale for mathematical disciplines. In Proceedings of the Ninth ACM Conference on Learning@ Scale, pp. 400-404 (2022).

[11] Williams, J., Kane, D., Sagu, S., Smith, E.: Exploring the national student survey: Assessment and feedback issues. The Higher Education Academy, Centre for Research into Quality (2008).

[12] Xu, Y., Harfitt, G.: Is assessment for learning feasible in large classes? Challenges and coping strategies from three case studies. Asia-Pacific Journal of Teacher Education, 47(5), 472-486 (2019).