I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17511

Man vs Machine: Can AI Grade and Give Feedback Like a Human?

Arnaud Chevalier
Jakub Orzech
Petar Stankov

DECEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17511

Man vs Machine: Can AI Grade and Give Feedback Like a Human?

Arnaud Chevalier
University of London, IZA, CESIfo and Vive

Jakub Orzech
University of London

Petar Stankov
University of London

DECEMBER 2024

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Man vs Machine: Can AI Grade and Give Feedback Like a Human?*

Grading and providing feedback are two of the most time-consuming activities in education. We developed a randomised controlled trial (RCT) to test whether they could be performed by generative artificial intelligence (Gen-AI). We randomly allocated undergraduate students to feedback provided either by a human instructor, ChatGPT 3.5, or ChatGPT 4. Our results show that: (i) Students treated with the freely accessible ChatGPT 3.5 received lower grades in subsequent assessments than their peers in the control group who always received human feedback; (ii) No such penalty was observed for ChatGPT 4. Separately, we tested the capacity of Gen-AI to grade student work. Gen-AI grades and ranks were significantly different than human-generated grades. Overall, while the newest LLM helps learning as well as a human, its ability to grade student work is still inferior.

JEL Classification: A22, C93, I23, I24

Keywords: feeback, grading, Artificial Intelligence, learning with Gen-AI

Corresponding author:
Arnaud Chevalier
Economics Department
Royal Holloway
University of London (RHUL)
TW20 0EX Egham
Great Britain

E-mail: arnaud.chevalier@rhul.ac.uk

* We are grateful to colleagues at the Centre for Research in Economics Education (CREEd), the CTaLE CoP AI in Education seminar series, as well as the CORE Econ workshop for helpful comments and suggestions. The experiments were financially supported by the Reid Research Fund and approved by the RHUL Ethics Committee Project ID 3907 and Project ID 4072, respectively.

I. Introduction

Feedback is essential for student performance (Bandiera et al. 2015, Hattie & Timperley, 2007; Wisniewski et al., 2020; Wang & Zhang, 2020). However, its production is time consuming, especially when more personalised feedback is required. To reduce time costs and improve personalisation of feedback it might be possible to rely on generative artificial intelligence (Gen-AI). The performance of Gen-AI has increased in leaps and bounds since the release of ChatGPT, a chatbot based on Large Language Model (LLM) in November 2022, given its almost human-like ability to converse (Bansal et al. 2024). To assess the potential of ChatGPT to provide feedback to students, we designed a randomised controlled trial (RCT) involving a group of undergraduate students who were randomly assigned to have feedback on their assessments generated by either AI (ChatGPT 3.5 or a newer generation model, ChatGPT 4), or a human grader. While others have tried to directly measure the quality of AI-generated feedback (Banihashem et al., 2024; Dai et al., 2023; Steiss et al., 2024), we focus on the core function of feedback---helping student learning. We assess whether the feedback improved student learning by measuring grades at a subsequent, related assignment.

In the second part of the paper, and outside of this experimental set-up, we also assess the grading performance of ChatGPT versus a human grader. Altogether, theses evidences allow us to assess whether current LLMs could boost productivity in the education sector by saving instructors' time invested in feedback and grading, and reallocating it to activities that could further improve students' engagement and understanding.

In its most basic form, feedback might simply consist of a grade summarising the student's overall performance at an assessment. Even this simplest form of feedback can improve subsequent performance (Bandiera et al. 2015). More detailed and personalised feedback might be even more effective but is more time-consuming to provide, especially at scale. To reduce this cost, alternative feedback methods have been implemented. For example,

peer feedback, regular multiple-choice quizzes or adaptive learning environments have been effective at enhancing student performance (Li et al., 2019[3] , Chevalier et al. 2018, Ippinaye and Risquez, 2024 or Kabudi et al., 2021), while Cavalcanti et al. (2021) and Keuning et al. (2018) offer broader critical reviews of feedback and learning methods.

Following the public release of ChatGPT 3.5, much attention has been given to ways it could affect students[4]. Our contribution is to the emerging field that has focused on the effects of deploying Gen-AI in teaching, particularly in assessment, feedback and learning. For example, feedback provided by GPT3.5 was found to be of inferior quality to that of a peer (Banihashem et al., 2024) or instructors (Dai et al., 2023, Steiss et al., 2024). However, these papers did not study the impact of the source of feedback on student outcomes, such as academic achievement.

The closest study to ours is Escalante et al. (2023). It examines the performance of ChatGPT 4 in giving feedback to 48 students participating in an English writing course, where feedback provision was divided between a human and ChatGPT 4. Using pre- and post-assessments, they report no significant differences in learning outcomes between the two groups. However, while human feedback was provided during one-on-one tuition, ChatGPT 4 was provided by email, which may have been a confounding factor.

Our paper is the first to provide causal evidence from a double-blind RCT experiment on the effect of AI-generated feedback on student grades from two rapidly evolving LLM-based technologies – ChatGPT 3.5 and ChatGPT 4. Undergraduate students studying for an economics degree who agreed to participate were randomly allocated to a control group, where

[3] However, providing quality feedback for more advanced assignments requires a certain level of expertise which students may lack (Valero Haro et al., 2018) and may induce low compliance (Hovardas et al., 2014) while carrying risks of racial and intra-group biases (Dancer & Dancer, 1992).

[4] Predictably, grade improvements have been reported for students allowed to use Chat-GPT to answer quizzes (Choi and Schwarcz, 2023), Nakavachara et. al., 2024), translation tasks (Godwin-Jones, 2022) or write essays (Herbold et al., 2023), but it is unclear if higher grades reflect an improvement in human capital formation.

feedback was provided by a human instructor, or two treatment groups where feedback was provided by either ChatGPT 3.5 or ChatGPT 4, respectively. Two courses were chosen for the experiment, where students had to complete up to three assessments. These courses where selected due to the design of their assessments, whereby feedback to one assessment could be used to improve the subsequent one. For all arms of the experiment, grades were generated by an instructor with no knowledge of the treatment allocation. In addition, students did not know whether they were receiving feedback from a human, ChatGPT 3.5 or ChatGPT 4. The double-blind experimental design, combined with the scaffolded assessment design, allows us to compare the effect of the feedback sources on subsequent grades.

In a nutshell, our results show that i) Chat GPT 3.5 feedback reduces subsequent grades by up to 0.25 of a standard deviation relative to human-generated feedback; ii) no such penalty was observed for students receiving feedback from Chat GPT4; iii) the penalty is heterogeneous by ethnicity; iv) the source of feedback had no differential effect on student engagement with the feedback or the course material.

While all submissions were graded by the course coordinator, we subsequently used both LLMs to grade them using the same grading criteria that were applied by the human grader. We then compared the AI-generated grades to the one provided by humans for the same submission. Each submission was graded five times by each LLM to assess its grades. We did this grading exercise, because, while Gen-AI might not reproduce the grade distribution of a human grader, it could still be used as a substitute for grading if it preserves the rank of students. Due to evidence of bias in grading across observable characteristics (Breda and Ly, 2015; Feld, Salamanca and Hamermesh, 2016; Hanna and Linden, 2012; Jansson and Tyrefors, 2022) we also assess whether some group of students would benefit from being graded by Gen-AI rather than humans. In a nutshell, we find that both LLMs provide more extreme grades and grade

more inconsistently, resulting in ranks substantially different from those generated by a human grader. However, neither LLM appears biased against any socio-demographic group.

II. Background on Gen-AI and experimental set-up

2.1 Generative AI

OpenAI launched ChatGPT in November 2022. It is a Gen-AI chatbot based on a Large Language Model (LLM) and capable of having "human-like" interactions. By January 2023 it accumulated over 100 million users, making it the "fastest growing app in history" (Reuters, 2023). Like all LLMs, ChatGPT uses a large database of text to predict the next most plausible sequence of letters in a response to a prompt. The process is repeated until a complete answer emerges.[5] Some amount of randomisation means that responses to the same prompt would differ. In our case, this means that students would receive slightly different feedback to very similar input. In March 2023, OpenAI released a more advanced ChatGPT 4, offering more accurate answers, as well as more sophisticated interactions.[6] However, ChatGPT 4 came with a subscription fee, compared to the free ChatGPT 3.5 which remain accessible. Due to their performance differences, as well as the difference in access costs at the time we conducted the experiment, we consider the two versions of Chat GPT as alternative treatments capable of generating different learning outcomes. Other LLMs exist, but ChatGPT is by far the most popular (see Figure A1), which justified our choice of LLM for this work.

The ability of Gen-AI to process text means that, with an appropriate prompt, ChatGPT could produce individualised feedback on student assessment. However, the free version only allows users to produce feedback using a dialog box, while the paid version could take advantage of an Application Programming Interface (API) that can automate feedback to many

[5] A more detailed description of the working of LLM is available in Ray (2023).
[6] Please see Kalyan (2024) for a broad overview and performance characteristics across a variety of generative large language models, and Koubaa (2023) for comparisons of the capabilities of GPT3.5 and GPT4 specifically.

student essays at scale. To not conflate the performance of the two versions with differences in the input provided we did not rely on the API when using ChatGPT 4 and always manually pasted each assessment, the associated prompt and marking criteria, into a dialog box for both chatbots.[7]

At the time of the experiment, ChatGPT 3.5 was unable to read some types of content, such as graphs or photos. Any part of the assessment that could not be copied in a ChatGPT dialog box was therefore ignored by the LLM. As some assessments required data analysis leading to output in regression tables, ChatGPT 3.5 would not process all output, particularly when it was laid out as an image.[8] Both versions of ChatGPT were able to comment on the associated description of a table, but only if the description was sufficiently detailed. For example, if a student commented on the value of a coefficient, its standard error or the associated p-value, both versions of ChatGPT assumed that the table contained those attributes and reflected this in their feedback. As students typically commented on their tables, the inability of ChatGPT3.5 to process images and tables was somewhat mitigated. In what follows, we provide details on how feedback was generated and used to improve student work.

2.2 Experimental Design: Can AI provide feedback like a human?

The experiment received ethical approvals from a London-based university and was conducted at the Economics Department among students enrolled in two undergraduate courses in separate cohorts of the program, ensuring that students could only be enrolled once in the

[7] All prompts and grading criteria are given in the Appendix.
[8] We experimented with attaching images as separate files, which ChatGPT 4 could read. However, this affected its performance. Feedback became more generic and did not refer to unique aspects of the analysed essay. At the same time, grading became unstable, with the same essay receiving grades varying, in some cases, by as much as 80 points!

experiment. The courses were selected due to their scaffolded assessment structure, i.e., feedback on one assignment is useful in a subsequent assignment. Therefore, the feedback on assessment a_{-1}, if read by the student, might affect the outcomes in subsequent assessment a.

One of the courses was run in the second year of a 3-year undergraduate degree programme. The course was either mandatory or elective, depending on the degree programme that students were registered on. First, students had to submit a mandatory, formative, research proposal and received feedback on it. Second, based on their proposal and its feedback, they submitted a completed research project. Both the proposal and final submissions were graded, but only the latter counted towards the final grade for this course and was worth 20% of the overall grade for this course. To provide additional incentives to submit the proposal, students were informed that failure to submit a proposal will result in a cap to the grade of their final essay.

The experiment was also conducted among students registered on a 3rd-year elective course. Its assessment structure included a large empirical project split into three linked assessments, each worth 20% of the final grade (literature review, empirical analysis and policy brief). Students would receive feedback and grades separately at each stage. For more detailed descriptions of each assignment and specific marking criteria, please see the Appendix.[9]

Participants were recruited in the first few lectures and seminars, and through an email campaign among students registered on the two courses.[10] Students were informed about the purpose of the experiment. Participation in the experiment was voluntary: No incentives were given for participation and no information was provided to the course leader who graded student scripts about consent or treatment arm allocation of any individual student. Formal consent was given by 110 students: 46 from the Year-2 cohort and 64 from the Year-3 cohort,

[9] Fig. A2 and Fig. A3 in the Appendix provide a timeline of the experiment in each course.
[10] Please see Appendix C: Ethics for the invitation email and the participant information sheet.

representing 57% and 65% of the total number of registered students on the two courses, respectively. Students who did not submit the consent form were assumed to be unwilling to participate and were excluded from the experiment. Naturally, they received human-generated feedback on their assessments but were not considered in the analysis. The data on their observables was only used to compare the characteristics of participants and non-participants and assess any selection into the experimental set-up. Students were informed that their consent was dynamic, and of their right to opt out at any time. None of the participants dropped out.

Participants were then randomised to one of three groups: a control group, who received human feedback, or one of two treatment groups, receiving feedback by either ChatGPT 3.5 or ChatGPT 4. The allocation into the groups was fixed throughout the experiment, i.e., students receiving feedback from a given source kept receiving feedback from the same source for all subsequent assessments. To prevent harm to the individual participants and embed safeguarding mechanisms, a human checked all outgoing feedback. Eventually, no feedback was judged inappropriate or harmful. Alternative sources of feedback (peers, course instructor) remained open to all students. At no point were students informed about the source of feedback they received. To guarantee a double-blind design, two conditions were met. First, all scripts were graded by a member of staff who did not know the allocated treatment arm or whether the student had provided consent. After human grades and feedback were generated for all students, scripts were deanonymized. Then, Gen-AI feedback was generated for scripts allocated to one of the two treatment arms. Once Gen-AI feedback was generated, it replaced the human feedback in the assignment submission inbox before releasing it to the student. Scripts provided to ChatGPT for feedback did not contain identifiable information about the author to ensure that Gen-AI generated feedback based on the same anonymous submission as the human did. Finally, students did not know whether they were receiving feedback produced by a human or Gen-AI.

The main outcome of interest in the RCT is the grade at assessment a, and how it depends on the source of feedback on the previous assessment, a_{-1}. Due to the randomisation of treatment with feedback, any average differences in subsequent grade improvement could be attributed to the difference in the quality of feedback, assuming no systematic observed or unobserved differences in individual characteristics across the treatment groups.[11] The estimate originating from this RCT should be interpreted as an Intention To Treat (ITT) since, while we can guarantee that the students received the expected treatment, students remain untreated if they do not access the provided feedback.

To understand the mechanisms by which the source of feedback might affect grades, we collected additional data on the number of logins to the course Learning Management System (LMS) used in these courses and, specifically, on whether the feedback on any specific assessment was accessed. Using the information on access to feedback allows estimating the average treatment effect on the treated (ATT). Since students are uninformed about their treatment arm allocation, the decision to access feedback should be independent of the treatment, and the ATT is thus an unbiased estimate of the effect of the source of feedback on subsequent performance, conditional on having accessed the feedback.

III Data

3.1 Selection and Balancing

The selected courses included a total of 178 students, of which 110 agreed to take part in the experiment. Using individual student identification number, the agreement record was matched to administrative data to assess any differences in observable characteristics between students who consented or declined to participate. The observables include gender, ethnicity and nationality, and have been dichotomized. Nine students (5 non-participants and 4

[11] We discuss those differences in observable characteristics below.

participants) had incomplete demographic characteristics; their missing observations has been imputed as the median value for their corresponding cohort, and indicator for imputed value has been created. We also control for program registration as this related to minimum entry requirements and whether the 2nd year course is elective or compulsory. We control for prior academic performance using grades at a first-year statistics course, which each student had to complete and might be the most relevant to the applied assessments that are considered here. The material covered in this statistics course was closely related to the assessments used in our experiments: formulating an econometric model, estimating it with statistical software and analyzing its output. When the first-year statistics grade was unavailable, which was the case with students who joined after their first year of study, we used their grades from a second-year Econometrics course. The last analyzed observable characteristic is the academic performance at the first assignment for this course, which we use as the baseline grade of students. Grades for all assessments were measured on a scale between 0 and 100. Non-submissions received a grade of zero, in line with formal administrative requirements. Non-submissions were included in the estimations, as they might be a treatment outcome and a result of optimization at the student level. Grades were normalized at either cohort level (for year-1 grade) or assessment level (for assessment-specific grades).

Table 1, Panel A indicates that there are no substantial differences in observable student characteristics between those who decided to take part and those who did not. Non-White students and those not-registered on the straight economics programs were slightly less likely to take part in the experiment but this is only statistically significant at the 10% level. Overall, there is no selection into the experiment which guarantees the external validity of our conclusion to the full set of students who could have been engaged in the experiment.

Table 1: Selection into the RCT and Randomisation of Treatment

Panel A: Selection into the RCT			
	Non-Participants	Participants	t-test (p-value)
Female	0.368 (0.486)	0.282 (0.452)	0.242
Non-White	0.750 (0.436)	0.627 (0.486)	0.083
Non-UK	0.309 (0.465)	0.400 (0.492)	0.216
BSc Economics	0.412 (0.496)	0.545 (0.500)	0.084
Year 1 Grade	-0.161 (0.942)	0.083 (0.993)	0.103
Assessment 1 Grade	-0.153 (1.048)	0.095 (0.957)	0.115
Non-submission for Assessment 1	0.118 (0.325)	0.073 (0.261)	0.337
Observations	68	110	

Panel B: Randomisation across Treatment Arms		
	ChatGPT 3.5	ChatGPT 4
Female	-0.147 (0.104)	0.016 (0.114)
Non-White	-0.033 (0.119)	-0.042 (0.119)
Non-UK	0.055 (0.115)	0.160 (0.119)
BSc Economics	0.065 (0.118)	0.103 (0.122)
Year 1 Grade	-0.040 (0.234)	-0.086 (0.246)
Assessment 1 Grade	0.042 (0.101)	0.145 (0.118)
Non-submission for Assessment 1	0.009 (0.028)	0.029 (0.030)

Notes: Panel A displays means of the observable characteristics between students who agreed to participate in the experiment and those that did not, along with their standard error (in parenthesis) and a two sided t-test of mean differences. Since Panel A compares participants to non-participants, the grades were normalized based on all students from the dataset. Panel B displays output from separate regressions of observable characteristics as a dependent variable and state of treatment along with controls for all other individual level observable characteristics as explanatory variables, for all 110 participants. Since Panel B compares three groups of participants, the grades were normalized based only on consenting students. Robust standard errors presented in the parentheses. Symbols: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Panel B focuses on whether the randomisation of treatment works and displays output from separate regressions of the observable characteristics on indicators of treatment, while controlling for all other individual-level observable characteristics. Estimates of the parameters on treatment arms show the comparison of analysed observable characteristic to the control group. The differences in student characteristics between treatment arms are small and statistically insignificant. Therefore, we conclude that treatment allocation was as good as random.

3.2 Can AI Grade Like a Human?

In addition to randomizing feedback treatment to evaluate its effectiveness on student outcomes, we assessed the performance of LLMs in grading student work. Each script was independently graded by a human – the grade released to the student – and then subsequently by each LLMs. As LLMs could be internally inconsistent in their answers – i.e., they can provide different answers for the same prompt – we repeated the grading procedure five times for each assessment and LLM. These grades were never used to assess students, nor released to students, and were only created for comparison purposes. This also ensured that no student was disadvantaged by biases or hallucinations embedded in the LLM grading algorithm. For official grading purposes, we always used the human grade. As there was no experiment per se, this part includes all student scripts. Still, this analysis excludes non-submitted assignments which would have been graded as 0 for each grader, potentially increasing the correlation between the human and AI grades.

We used the same assessments as in the experimental set-up and, independently of the RCT treatment allocation, prompted ChatGPT 3.5 and ChatGPT 4 to provide a grade for all assessments. Like for the provision of feedback, the prompt included the full set of criteria that students had been informed about and specified the maximum points for each criterion. The

prompts can be found in Appendix B. After receiving the prompt, the LLMs provided a grade for each criterion. Summing them up produced an assessment grade. If the Gen-AI's response was apparently using a wrong set of criteria or incorrect weights for the stated criteria, that grade was excluded and the request repeated until the grading was consistent with the prompt. In the case of ChatGPT 3.5, approximately 4% of assessments had to be regraded due to omitted or extra criterion used by the LLM and another 6% due to inconsistency with the provided prompt for at least one criterion. The similar figures for ChatGPT 4 are 4% and 1% of grades, respectively.

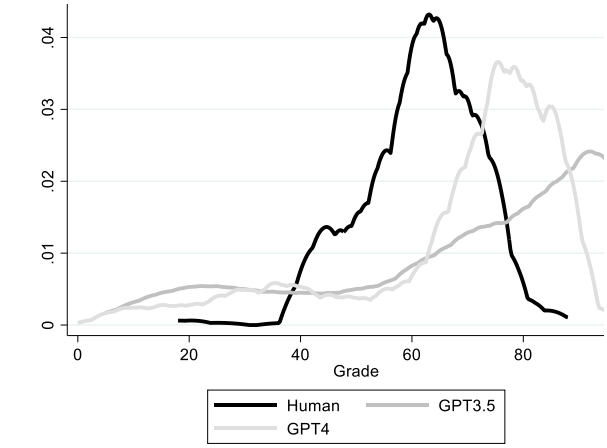
In addition, the university has a stepped grading policy for all written submissions. This requires human graders to assign grades ending in either 2, 5 or 8, unless the grade is 0 or 100. We adjusted the AI-generated grades to conform with the policy by rounding down the LLM grade to the nearest grade allowed by the policy. We chose to round the LLM grade down, because we noticed that the LLM grades were systematically higher than human grades. The AI generated grades are then compared to human-generated grades. To minimise human biases, human grading was always done on anonymous student scripts, without knowledge of the treatment arm allocation and prior to the assessment being graded by Gen-AI. Both LLMs were unaware of the human grade as well. Figure 1 illustrates the resulting raw grade distributions.

Figure 1A reports the distribution of raw grades on all assessments by grading source. Human grades have an almost normal distribution with a small tail of low achievers. Using ChatGPT4 results in a similarly shaped distribution but shifted to the right – indicating more generous grading – and a longer tail of low achievers. ChatGPT3.5 delivers a very different distribution, with a high fraction of very high grades and an almost uniform distribution of grades between 10 and 50. Figure 1B reveals the large standard deviation of grades within the same submission and highlights the lack of consistence in grading, especially for ChatGPT 3.5. The average submission-level standard deviation are 7.6 and 4.4 points for ChatGPT 3.5 and

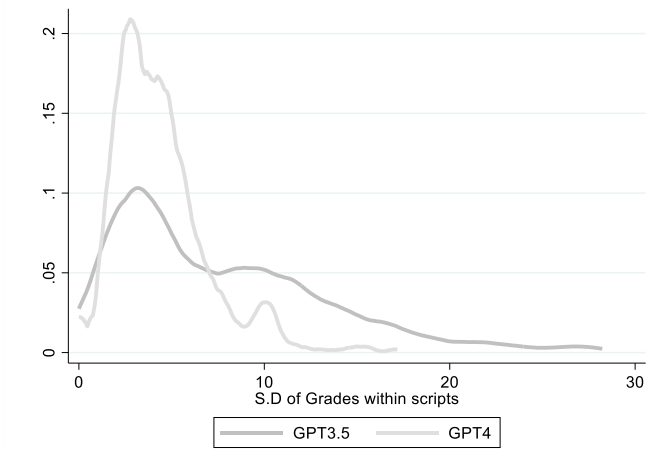
ChatGPT 4, respectively, with 30% of submissions having a standard deviation greater than 10 points for the former.

Figure 1: Distribution of Grades and Standard Deviation

A: Density of Grades by Type of Grader



B: Density of Standard Deviation of grades within Submissions

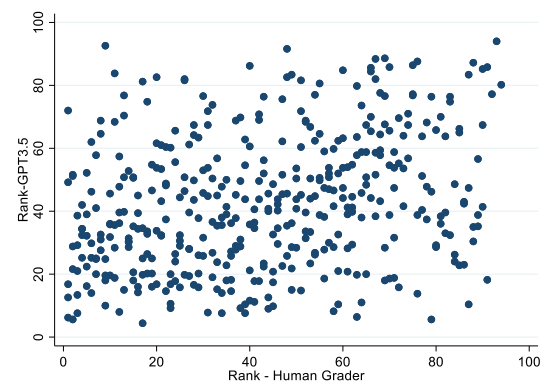


We then normalise the AI grades for each type of the assessment separately, using the mean and standard deviation of the human grades for the same assessment. The normalization uses all grades for submitted scripts and the moments from the human grade distribution, as we want to benchmark the AI grades against those of human graders. We called those grades "Normalized grades compared to human".

Even if differing in their distribution, AI-generated grades would still hold valuable information if they preserved the ranking of students. To be able to compare between assessments and courses of different sizes, the ranks (r) were transformed into percentiles of the distribution $\left(p = \frac{r_{max}-r+1}{r_{max}} * 100\right)$, so that the best student received a percentile rank of 100 and the worst one a percentile rank of $1/r_{max}$. The step marking results in bunching of grades, we break ties by randomising rank among students with the same grade. We then compute the absolute differences of percentile ranks given by Gen-AI and those of the human, which was called "Absolute difference in Percentile Rank compared to human".

Figure 2: Human Rank vs AI-Generated Rank, in percentiles

A) GPT 3.5



B) GPT 4

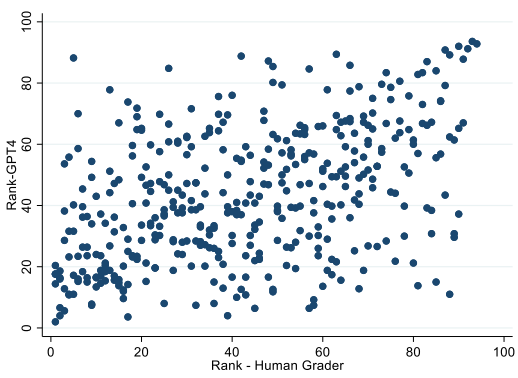


Figure 2 compares the average AI-generated rank to the human rank for each LLM separately. It clearly highlights the lack of significant correlation in rank between human and AI grading, with the correlation ranging from 0.30 to 0.45 for ChatGPT 3.5 and ChatGPT 4, respectively. Note that the correlation in ranks between the two LLMs is also only 0.49.

Table 2: Normalised Grades and Rank, by Grader Type and Assessments

	Normalized Grade compared to human		Absolute difference in Percentile Rank compared to human	
Year and Assignment	GPT3.5	GPT4.0	GPT3.5	GPT4.0
Year 2, Ass. 1	-2.921	-2.381	25.05	23.55
	(1.737)	(1.726)	(19.73)	(18.68)
Year 2, Ass. 2	0.638	0.583	27.76	24.39
	(1.567)	(0.940)	(21.53)	(19.82)
Year 3, Ass. 1	2.777	2.504	31.06	25.54
	(1.905)	(1.011)	(21.51)	(20.87)
Year 3, Ass. 2	1.772	1.062	28.27	23.26
	(1.626)	(1.309)	(23.62)	(18.88)
Year 3, Ass. 3	3.236	1.864	30.38	27.62
	(2.183)	(1.401)	(21.37)	(21.35)
Total	1.315	0.882	28.70	24.94
	(2.790)	(2.075)	(21.77)	(20.05)

Notes: The table presents the mean of normalized grades and the mean of absolute differences of percentile ranks between AI and human along with their standard deviations (in parentheses) for each assessment separately, as well as all assessments in total. Grades are normalised by assessment using the mean and standard deviation of human grades.

Table 2 reports relative grades and rank for both ChatGPT 3.5 and ChatGPT 4. With the exception of the first assignment in Year 2, AI-generated grades are considerably more generous than human grades, differing by 1.315 and 0.882 standard deviations, respectively. The initial Year 2 assignment was a proposal for an empirical project, rather than a fully developed project, which resulted in harsh marking from both LLMs and indicated an inability of either LLM to evaluate a proposal for its feasibility. The variations in absolute ranks are also considerable, ranging from 28.70 or 24.94 percentile ranks, with little variations between assignments. Overall, LLMs appear inconsistent in marking and are probably still unsuitable to substitute humans in grading.

IV Econometric specifications

To identify the impact of AI-generated feedback we estimate the effect for student *i* of receiving feedback $s = \{h, GPT35, GPT4\}$ from a human, ChatGPT 3.5, or ChatGPT 4, respectively, on its normalized grades (g) in assessment *a* as:

$$g_{ia} = \beta_0 + \sum_{s=1}^{2} \beta_s (D_{is}) + \gamma\, g_{ia=1} + X_i'\beta + \sum_{s=1}^{2} \lambda_s (D_{is} * X_i') + \varepsilon_{ia}, \quad (1)$$

where D_{is} indicates the treatment assigned to student *i*. Since the treatment only starts after the students completed their first assignment, the regression is estimated on subsequent assignments only, i.e., one assignment for Year-2 students, and two assignments for Year-3 students, and we control for first assignment grade, $g_{ia=1}$. Since the allocation to treatment is fixed over the course of the experiment, it is not possible to include student fixed effects. Instead, we include a set of control variables (X_i') featuring gender, ethnicity, nationality, program of study, 1st-year grades to approximate for student ability, and assignment sequence. In some specifications, we test for heterogeneity in the effect of the treatment by including

interactions between feedback source and student's characteristics; for student ability, this is defined as an indicator for scoring above the median at the first year's statistics course. The standard errors are clustered at the level of randomization, i.e., the individual student.

The estimate on the treatment allocation can be interpreted as an Intention-To-Treat (ITT) effect. However, using the information from our Learning Management System (LMS), we know if the student accessed the provided feedback. While deciding to access feedback is endogenous, this decision is unlikely to be correlated to the treatment allocation, since students are unable to identify which treatment they were allocated to (see formal test below) and could not guess it without accessing the feedback in the first place (at least for the first assignment). We thus also compute the Average Treatment on the Treated (ATT) effect by conditioning on having checked the feedback left on the previous assignment $(a-1)$. This might be considered the parameter of interest, since the source of feedback is unlikely to have any effect on students who do not access the feedback. We report both ITT and ATT in subsequent tables.

A similar model was used for the other outcomes of interest, which helps reveal the potential mechanisms by which feedback has generated differences in grades. Specifically, we created an indicator on whether the feedback provided at assessment *a* (F_{ia}) was accessed and, more generally, engagement with the learning management system (L_{ia}) prior to submitting assessment *a*.

Separately, to assess the relative grading performance of Gen-AI, we estimate the following model:

$$g_{ia} = \delta_0 + \delta_1 GPT3.5_{ia} + \delta_2 GPT4_{ia} + \gamma_{ia} + \varepsilon_{ia} \quad , \qquad (2)$$

where δ_j is the difference in grades (g) between LLM_j and a human grader for student i at assessment a. Equation (2) includes a fixed effect at the level of the submitted script γ_{ia} for each student; i.e. for each submission. Each script is graded 11 times – 1 from a human, 5 from ChatGPT 3.5 and 5 from ChatGPT 4 – we thus cluster the standard errors at the submission

level. Finally, we repeated the estimations using absolute difference in percentile rank between human and AI within an assignment rather than the normalized grade as the dependent variable.

V. Results

5.1 Impact of AI-generated feedback on student grades

By randomising the source of assessment feedback, we identify its effect on grades in subsequent assessments. These estimates are reported in Table 3a and Table 3b, separately for the full sample (ITT, Table 3a) and for the sub-sample of students who consulted the feedback for the previous assessment (ATT, Table 3b).

The estimated ITTs in column 1, reveal no differences in subsequent grades from providing AI-generated feedback. The grades of students receiving ChatGPT 4 feedback were undistinguishable from the ones using human-generated feedback. Students who received feedback from ChatGPT 3.5 performed 0.2 standard deviations worse at subsequent assessments, but while large this effect is not statistically significant. In the subsequent columns, we report estimates allowing for heterogenous effect by observable characteristics, and assess the effect of feedback source by ability, gender, nationality and ethnicity separately. The estimates indicate that Gen-AI feedback might have different effect by nationality and ethnicity– but those differences are only significant at the 10% level. Non-nationals appear to benefit from feedback provided by Chat-GPT 3.5 while non-white students perform worse at subsequent assessments when gaining AI-generated feedback. The ITT results suggest that it could be possible to switch to AI-generated feedback at no penalty to student learning, especially when relying on newer LLMs like ChatGPT 4. However, the switch may not be costless to learning in an ethnically diverse classroom.

The feedback from the previous assignment was not accessed for 17% of the scripts, which might bias the estimates towards zero, as these students did not receive any treatment. Table 3b reports ATT estimates, where we only include students who have consulted the

feedback on the previous assessment, and thus could have been treated. In Table 4, we report that accessing feedback is uncorrelated to treatment assignment, as expected due to the randomisation. Therefore, the reported ATT estimates can be interpreted as the causal effect of feedback source on subsequent grades for students who saw the feedback.

When we focus on grade comparisons for students who were treated, the precision of estimates improves and a penalty for gaining feedback from ChatGPT 3.5 becomes statistically significant at the 10% level. This is in line with Banihashem et al. (2024) and Steiss et al. (2024), who report lower-quality feedback by ChatGPT 3.5, compared to that of a human. In our case, students who got feedback from the previous generation LLM scored 0.26 standard deviations lower than the control group, a large effect for an educational intervention (Kraft, 2020). However, improvement in generative AI is rapid and, like Escalante et al. (2023), we find that in economics assessments Chat GPT 4 produces feedback that is not statistically different in quality to that of a human.

ATT results support most of the findings about the heterogeneity in the effect from the ITT model, with again non-white students being penalised when getting feedback from AI-generated feedback and receiving grades at the subsequent assignment 0.5 to 0.6 of a standard deviation lower. Overall, conditional on accessing the feedback, students' performance at subsequent assessments are similar when the feedback was generated by a human or ChatGPT 4 but inferior in the case of ChatGPT 3.5 generated feedback.

Table 3: Estimated Effect of Feedback Treatment on Subsequent Grades

Panel A: ITT	ITT	ITT Heterogeneous: Year 1 Grade	ITT Heterogeneous: Gender	ITT Heterogeneous: Nationality	ITT Heterogeneous: Ethnicity
ChatGPT 3.5	-0.200	-0.217	-0.300	-0.532**	0.355
	(0.205)	(0.254)	(0.208)	(0.236)	(0.417)
ChatGPT 4.0	0.008	-0.216	-0.106	0.027	0.527
	(0.195)	(0.213)	(0.176)	(0.166)	(0.388)
Observable Characteristic=1		-0.461	-0.293	-0.452	0.397
		(0.389)	(0.450)	(0.420)	(0.399)
Chat GPT3.5 # Observable characteristic=1		0.058	0.385	0.917*	-0.932*
		(0.389)	(0.546)	(0.489)	(0.525)
Chat GPT 4.0 # Observable characteristic=1		0.441	0.352	0.102	-0.833*
		(0.380)	(0.486)	(0.462)	(0.429)
Observations	174	174	174	174	174
Control	Yes	Yes	Yes	Yes	Yes

Panel B: ATT	ATT	ATT Heterogeneous: Year 1 Grade	ATT Heterogeneous: Gender	ATT Heterogeneous: Nationality	ATT Heterogeneous: Ethnicity
ChatGPT 3.5	-0.255*	-0.065	-0.194	-0.389**	0.115
	(0.138)	(0.150)	(0.165)	(0.193)	(0.165)
ChatGPT 4.0	-0.081	-0.052	-0.010	0.052	0.264
	(0.115)	(0.151)	(0.141)	(0.150)	(0.162)
Observable Characteristic=1		-0.040	0.214	-0.002	0.315**
		(0.232)	(0.151)	(0.153)	(0.155)
Chat GPT3.5 # Observable characteristic=1		-0.327	-0.245	0.346	-0.604**
		(0.252)	(0.236)	(0.252)	(0.256)
Chat GPT 4.0 # Observable characteristic=1		-0.057	-0.219	-0.241	-0.544**
		(0.269)	(0.260)	(0.232)	(0.244)
Observations	145	145	145	145	145
Control	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates. Additional controls include grade at initial assessment, gender, nationality, ethnicity, year 1 grade and indicator for assessment. Observable characteristic=1 means having year 1 grade above median in column (2), being female in column (3), being an overseas student in column (4) and being a non-White student in column (5). Estimates for interaction terms between treatment arms and the respective observable characteristic are included in all models except (1). The standard errors, reported in parentheses, are clustered at the student level. Symbols: * p<0.1, ** p<0.05, *** p<0.01.

5.2 Student engagement with feedback and LMS

5.2.1 Accessing feedback

The differences in engaging with feedback across types of treatment could be a mechanism by which the source of feedback affects subsequent grades. Indeed, across all assignments only 70% of students access their feedback. The perceived quality of the initial feedback may affect the likelihood to access subsequent feedback. To test this, we use a linear probability model with specification similar to Eq. (1) to estimate whether the treatment arm allocation correlates with the probability of engaging with feedback on the following assignment. The main difference between Eq. (1) and the model used in this part is that now we control for the grade received since a grade is the most immediate feedback that students receive. Additionally the ATT sample is defined on the sub-sample of students who access feedback at the first assessment and thus, whose subsequent decisions to access feedback is informed by their perception of the initial feedback received.

In Table 4 we report these estimates for various sub-samples. First, using the full sample, we estimate that students allocated to either type of AI-generated feedback are 11 to 12 percentage points less likely to engage with subsequent feedback, however these effects are not statistically significant. We then limited the sample by excluding the 10 non-submissions since mechanically the students did not have any feedback to access, but this did not alter the conclusions substantially. In the next two columns we estimate the same model but restrict the sample to students who viewed the feedback on their first assignment, and thus whose behaviour regarding accessing feedback for subsequent assessment could have been affected by differences in the perceived quality of the feedback, depending on its source. While the point estimates increase to 13.5 and 12.3 percentage points, respectively, we still cannot reject the hypothesis that the source of feedback has no impact on the probability of engaging with

it[12]. The feedback source thus cannot be a mechanism explaining the lower grades obtained by students allocated to ChatGPT 3.5. So any difference in subsequent grade is likely to be due to differences in feedback content between ChatGPT 3.5 and a human grader.

Table 4: Effect of Feedback Source on Accessing Feedback

	ITT	ITT if submitted	ATT	ATT if submitted
ChatGPT 3.5	-0.121	-0.128	-0.131	-0.135
	(0.084)	(0.090)	(0.092)	(0.092)
ChatGPT 4.0	-0.114	-0.119	-0.125	-0.123
	(0.090)	(0.095)	(0.095)	(0.095)
Normalised Grade	0.189***	0.234***	0.199***	0.220***
	(0.021)	(0.057)	(0.042)	(0.061)
Observations	174	164	153	151
Control	Yes	Yes	Yes	Yes

Note: OLS estimates. Additional controls include grade at initial assessment in the treated course, gender, nationality, ethnicity, year 1 grade status and indicator for assessment. The standard errors, reported in parentheses, are clustered at the student level. "ITT if submitted" restricts the sample to students who submitted the assessment of interest; "ATT" restricts the sample to students who checked feedback on their first assessment; "ATT if submitted" restricts the sample to students who checked feedback on their first assessment and submitted the assessment of interest.
Symbols: * p<0.1, ** p<0.05, *** p<0.01

5.2.2 Engaging with Course Material

The quality of feedback may also affect the need to access additional material and engage with the course Learning Management System. We can analyse the level of engagement with the LMS using login activity. To quantify this engagement, we tracked login activity for each week of teaching and each student. We then calculated the share of weeks between the release of feedback for assignment $a-1$ and the submission deadline of assignment a when a student logged into the LMS. For example, if that period was three weeks and a student logged on the LMS in two of them, then the share of "active weeks" was recorded as 2/3. We tracked the share of active weeks for each assignment and each student separately. The average share

[12] We also restricted the sample to observations for which students accessed the feedback for their previous assessment, rather than the first assessment, but the conclusions remain similar (Table A2 in Appendix).

of "active weeks" was found to be 0.85. As an alternative measure of engagement with the course material, we calculated the average number of logins over the weeks between the release of feedback for assignment $a-1$ and the submission deadline of assignment a. The average number of logins over the week was found to be 19.68.

Table 5: Effect of Feedback Source on Accessing Course Material

	ITT: Share of active weeks	ITT: Logins per week	ATT: Share of active weeks	ATT: Logins per week
ChatGPT 3.5	0.060	-1.607	0.019	-1.899
	(0.052)	(2.828)	(0.034)	(3.197)
ChatGPT 4.0	0.046	-1.118	0.045	-1.597
	(0.045)	(2.637)	(0.029)	(2.740)
Share of active weeks before Assignment 1	0.635***		0.248**	
	(0.156)		(0.111)	
Number of logins before Assignment 1		0.719***		0.674***
		(0.127)		(0.133)
Observations	174	174	145	145
Controls	Yes	Yes	Yes	Yes

Notes: OLS estimates. Additional controls include gender, nationality, ethnicity, year 1 grade and indicator for assessment. "ATT" restricts the sample to students who check feedback on the previous assessment; The standard errors, reported in parentheses, are clustered at the student level. Symbols: * p<0.1, ** p<0.05, *** p<0.01

We then estimated if the type of feedback influenced these measures of engagement using a specification identical to Eq. (1), controlling for the initial activity – i.e. prior to the student receiving any information about the treatment. This latter term is used to account for students unobserved propensity to engage with course material. In Table 5 we report these estimates which indicate marginal and insignificant impact of Gen-AI on the level of engagement with the LMS, regardless of how we measure engagement or students' sample. Gen-AI feedback caused an increase in the share of active weeks of 2.2% or 5.3% of the mean and decrease in the logins per week of 8.11% or 9.65% of the mean, however all estimated coefficients are insignificant.

VI. Can AI Grade Like a Human?

The second task where Gen-AI could support course leaders is grading. To assess its effectiveness in grading, we do not need an experiment but simply grade the same submission using human graders or ChatGPT, and compare the resulting grades. As explained above, each submission is marked five times by each LLM and normalized using the human distribution at the same assessment. Similarly, we use the absolute difference in ranking compared to the human grader rank.

6.1. Grading and Ranking

Table 6 reports comparisons between grades (Panel A) and ranks (Panel B) assigned by either Gen-AI or a human, overall and separately for each of the five analysed assessments. The results confirm that Gen-AI is considerably more generous in grading than a human grader, with grades being on average 1.315 and 0.882 standard deviations larger for ChatGPT 3.5 and ChatGPT 4, respectively. Moreover, this hides considerable heterogeneities between assessments. For example, the first assessment in Year 2 was graded considerably harsher by Gen-AI. This was to be expected, as the assignment was a proposal for an empirical project, rather than a fully developed one. It appears that both versions of ChatGPT were unable to infer the quality of the final product based on its rough draft, while the human grader could judge the potential to finalise a project, and their grade was a reflection of that potential. Unlike humans, Gen-AI was grading these assignments as if they were already finished essays, of poor quality. This is because Gen-AI seems to operate within a certain technological frontier, where most complex text generation tasks deliver performance advantages (Dell'Acqua et al., 2023), while other seemingly simple tasks – e.g., judging the potential of a student project – seem outside of its current frontier. For all other assessments, and independently of their specificity, Gen-AI grades were always larger than human graders. The largest difference between the two AI-generated grades is in Year 3, assessment 3 where ChatGPT 3.5 struggled to assess

empirical output, as it was unable to process tables of results that had been included as images, while ChatGPT 4 could.

While being overly generous in grading, generative AI could still be used if its ranks conform with those of a human grader. If so, grading could be outsourced to LLM with the human only fitting the grades to a preferred distribution. However, as presented in Panel B, ranks based on ChatGPT grades were significantly different between LLMs and a human. The absolute differences between ChatGPT 4 and human ranks varied from 23.46 to 27.75 points depending on assignments. Differences between ChatGPT 3.5 and human ranks were even larger and varied from 25.26 to 30.66 points. Therefore, our analysis indicates that neither grades nor ranks generated by ChatGPT can be used as a substitute for human marking for the time being.

Table 6. AI vs Humans: Grades and Ranks

Panel A: Normalized Grades compared to human						
	All	Year 2 Assess 1	Year 2 Assess 2	Year 3 Assess 1	Year 3 Assess 2	Year 3 Assess 3
ChatGPT 3.5	1.315***	-2.921***	0.638***	2.777***	1.772***	3.236***
	(0.136)	(0.167)	(0.147)	(0.181)	(0.160)	(0.237)
ChatGPT 4	0.882***	-2.381***	0.583***	2.504***	1.062***	1.864***
	(0.101)	(0.169)	(0.124)	(0.108)	(0.096)	(0.149)
Observations	4587	781	781	1001	1034	990
Panel B: Absolute difference in Percentile Rank compared to human						
	All	Year 2 Assess 1	Year 2 Assess 2	Year 3 Assess 1	Year 3 Assess 2	Year 3 Assess 3
ChatGPT 3.5	28.378***	25.255***	27.006***	30.664***	27.895***	30.116***
	(0.900)	(1.971)	(1.765)	(1.881)	(2.168)	(2.061)
ChatGPT 4	24.946***	23.829***	23.456***	25.557***	23.635***	27.752***
	(0.829)	(2.029)	(1.883)	(1.931)	(1.516)	(1.917)
Observations	4587	781	781	1001	1034	990

Notes: OLS estimates include fixed effects at the submission level. The standard errors, reported in parentheses, are clustered at the submission level. Symbols: * p<0.1, ** p<0.05, *** p<0.01

6.2. Grading and Ranking: Robustness Check

As highlighted in the data section, Gen-AI generated grades vary substantially within submission. While the previous results report the average performance of an LLM, we now assess whether its performance could be improved by selecting either the minimum or maximum grade provided by Gen-AI for a given submission. We then recalculate ranks for the distribution of minimum (maximum) grades.

Table 7. AI vs Human: Grades and Ranks, Accounting for Variation in AI Grading

Panel A: Normalized Grades compared to human		
	Minimum grade	Maximum grade
ChatGPT 3.5	0.303*	2.332***
	(0.169)	(0.152)
ChatGPT 4	0.271**	1.490***
	(0.122)	(0.116)
Observations	1251	1251
Panel B: Absolute difference in Percentile Rank compared to human		
	Rank based on minimum grade	Rank based on maximum grade
ChatGPT 3.5	28.626***	26.690***
	(1.343)	(1.273)
ChatGPT 4	23.792***	24.410***
	(1.182)	(1.180)
Observations	1251	1251

Notes: OLS estimates include fixed effects at the submission level. The standard errors, reported in parentheses, are clustered at the submission level. Symbols: * p<0.1, ** p<0.05, *** p<0.01

Table 7, Panel A report estimates when using minimum or maximum grade. It is worth noting that the minimum AI-generated grades are still 0.3 standard deviation higher than the human grades. Comparing the effect for minimum and maximum grade highlight the high variation in AI grades with the maximum grade being up to 2 standard deviations greater than the minimum grade for the same submission! This difference is reduced but still large (1.2 standard deviation) when using ChatGPT 4. Such large variance in the grades that LLM provides for the same submission makes it difficult to use it to replace human graders, unless

it is used repeatedly and the average grade used, but such a multiple grading strategy would negate the time saving benefit of using Chat-GPT for marking.

Panel B presents estimates from the same model when the dependent variable is the absolute difference in the percentile rank between Gen-AI and human. Ranks were recalculated based only on the minimum/maximum grades. The ranks assigned by both ChatGPT 3.5 and ChatGPT 4 remain significantly different from human ranks, regardless of the selection of AI-generated grades. In all analysed cases the difference was bigger than 23 percentile points. Compared with the ranks produced by ChatGPT 4, the ranks assigned by ChatGPT 3.5 came out further away from the human ones.

6.3. Are Gen-AI Grades and Ranks Biased?

Our results hinge on the implicit assumption that neither feedback nor grades produced by Gen-AI and humans embed conscious or unconscious biases. However, Gen-AI is known to produce biased content (Thomson & Thomas, 2023; Ferrer et al., 2021; Roselli et al, 2019), and so are human graders along a variety of observable characteristics like gender (Arceo-Gomez and Campos-Vazquez, 2022; Jansson and Tyrefors, 2022; Di Liberto et al., 2021; Contreras, 2023), ethnicity (van Ewijk, 2011) and teaching mode (Ayllón, 2022). Therefore, we test whether the source of grading is associated with differences in grades for various characteristics, using the following model separately for each of our three grader types:

$$g_{ia} = \delta_0 + \delta_1 Gender_i + \delta_2 Ethnicity_i + \delta_3 National_i + \delta_4 High\ Ability_i + \lambda_a + \varepsilon_{ia}\ , (3)$$

where δ_j is the impact of observable characteristic j on the normalized grades or ranks g of student i in assessment a, and λ_a is the assignment fixed effect. The variable *High Ability* takes the value of 1 if a student's Year-1 grade was above median in their cohort, and 0 otherwise.

For each assessment *a*, grades were normalized using only the grades at the same assessment and marking source, e.g. ChatGPT 3.5 grades were normalized with the mean and standard deviation of ChatGPT 3.5 grades at assessment *a*. Table 8 compares grading (Panel A) and ranking (Panel B) between human, ChatGPT 3.5 and ChatGPT 4 by gender, nationality, ethnicity and prior academic achievements.

Table 8. Heterogeneity in Grading

Panel A: Normalized Grades			
	Human	ChatGPT 3.5	ChatGPT 4
Female	0.071	0.091	0.162*
	(0.107)	(0.089)	(0.095)
Non-UK	-0.062	0.007	-0.023
	(0.104)	(0.089)	(0.094)
Non-White	-0.039	-0.030	0.118
	(0.103)	(0.086)	(0.095)
Year 1 grade above median	0.330***	0.003	0.231**
	(0.102)	(0.086)	(0.092)
Observations	417	2085	2085
Panel B: Percentile Rank			
	Human	ChatGPT 3.5	ChatGPT 4
Female	1.817	2.176	3.346
	(3.138)	(2.520)	(2.605)
Non-UK	-1.740	0.722	-1.185
	(3.079)	(2.479)	(2.580)
Non-White	-0.483	-0.529	4.294
	(3.068)	(2.455)	(2.616)
Year 1 Grade above median	10.387***	1.865	7.696***
	(2.933)	(2.374)	(2.480)
Observations	417	2085	2085

Note: OLS estimates. The standard errors, reported in parentheses, are clustered at the submission level.
Symbols: * p<0.1, ** p<0.05, *** p<0.01

Students with higher abilities, as measured by their Year-1 grades, tend to receive better grades, but not if graded by ChatGPT 3.5! This is surprising, since grades are a function of

ability and effort, which should be correlated over time, and we should expect grades in year 1 to be positively correlated with subsequent grades. Male students, international as well as non-White students tend to receive lower grades from the human grader – but those differences are not statistically significant. ChatGPT 3.5 tends to reduce these biases, while ChatGPT 4 makes them larger, to the point that female students have grades 0.16 of a standard deviation higher than male students, an effect significant at the 10% level. However, any biases by observed characteristics appear to have little effect on the rank of students. This is true for all graders, and the maximum rank difference is only 4 percentage points in favour of non-white students when being graded by ChatGPT 4. Altogether, the results suggest that any biases observed in Gen-AI grades appear insignificantly different from those of a human. Although ChatGPT replicated existing human biases, those were not significantly larger than the biases inherent in human grading.

VII. Conclusion

This work assessed the potential use of Gen-AI in performing two of the most time-consuming activities in higher education – generating feedback and grading student work. The study involved an RCT with 2nd and 3rd-year undergraduate students receiving feedback from a human grader, ChatGPT 3.5 or ChatGPT4. Our findings indicate that there is no statistically significant difference in subsequent grades between students receiving feedback from ChatGPT 4 and a human grader. However, our ATT estimates point to a possible limitation of using GPT3.5: Students receiving feedback from this LLM scored 0.26 of a standard deviation lower at the subsequent assignment. While the effect was only marginally significant, it is sufficiently large to recommend not using ChatGPT 3.5 for student feedback as it may harm student learning.

In addition, the source of the feedback does not interfere with the student engagement with the course material and willingness to access the feedback. Thus, any decrease in performance is likely to stem from differences in the quality of the feedback. Indeed, previous work has found a large feedback quality gap between ChatGPT3.5 and ChatGPT4.0 (Banihashem et al., 2024; Steiss et al., 2024; Escalante et al. 2023), which explains the effects we identify in the first part of the paper.

In the second part of the paper, we assess the ability of LLMs to grade. Both versions of ChatGPT grade student work more generously than a human and produce ranks that are significantly different than the ones reported by a human grader. Moreover, their grading is unstable, delivering large variability in grades for the same submission and implying reliability issues in grading.

Overall, our results indicate that ChatGPT 4 can produce feedback comparable to a human grader, but was not, at the time of the experiment, able to provide consistent grading. The rapid improvement in LLMs gives hope that it might be possible to substitute human graders with Gen-AI soon, but not yet. This could be achieved at relatively low price. While at the time of the experiment ChatGPT 4 was only available for a fee, the current free version ChatGPT4Ω should offer at least similar level of performance, making it potentially a cost-effective alternative to human feedback provision.

References

Arceo-Gomez, E. O., & Campos-Vazquez, R. M. (2022). Gender bias in evaluation processes. *Economics of Education Review, 89*, Article 102272. https://doi.org/10.1016/j.econedurev.2022.102272

Ayllón, S. (2022). Online teaching and gender bias. *Economics of Education Review, 89*, Article 102280. https://doi.org/10.1016/j.econedurev.2022.102280

Bandiera, O., Larcinese, V., & Rasul, I. (2015). Blissful ignorance? A natural experiment on the effect of feedback on students' performance. *Labour Economics*, *34*, *13-25*. https://doi.org/10.1016/j.labeco.2015.02.002

Banihashem, S. K., Kerman, N. T., Noroozi, O., & others. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education, 21*(23). https://doi.org/10.1186/s41239-024-00455-4

Bansal, G., Chamola, V., Hussain, A., et al. (2024). Transforming conversations with AI—A comprehensive study of ChatGPT. *Cognitive Computation, 16*, 2487-2510. https://doi.org/10.1007/s12559-023-10236-2

Breda, T and Ly, S. T. (2015). "Professors in Core Science Fields Are Not Always Biased against Women: Evidence from France." *American Economic Journal: Applied Economics, 7 (4), 53–75*. https://doi/10.1257/app.20140022

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027

Chevalier, A., Dolton, P., & Lührmann, M. (2018). 'Making it count': incentives, student effort and performance. *Journal of the Royal Statistical Society: Series A: (Statistics in Society*, *)*, *181*(2), 323-349. https://doi.org/10.1111/rssa.12278

Choi, J. H., & Schwarcz, D. (2023). AI assistance in legal analysis: An empirical study. *Journal of Legal Education*, 73 (forthcoming, 2024). http://dx.doi.org/10.2139/ssrn.4539836

Contreras, D. (2023). Gender differences in grading: Teacher bias or student behaviour? *Education Economics*, *1–24*. https://doi.org/10.1080/09645292.2023.2252620

Dai, W., et al. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, *323-325*. IEEE.

Dancer, W. T., & Dancer, J. (1992). Peer rating in higher education. *Journal of Education for Business, 61*, *306-309.*

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged*

technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality (Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, The Wharton School Research Paper). SSRN. https://doi.org/10.2139/ssrn.4573321

Di Liberto, A., Casula, L., & Pau, S. (2021). Grading practices, gender bias and educational outcomes: Evidence from Italy. *Education Economics, 30*(5), *481–508.* https://doi.org/10.1080/09645292.2021.2004999

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(57). https://doi.org/10.1186/s41239-023-00425-2

Feld, J., Salamanca, N., Hamermesh, D.S. (2016). Endophilia or exophobia: beyond dis-crimination. *The Economic Journal, 126, 1503–1527.*

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: A cross-disciplinary perspective. *IEEE Technology and Society Magazine, 40*(2), 72-80. https://doi.org/10.1109/MTS.2021.3056293

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology, 26*(2), 5–24. http://doi.org/10125/73474

Hanna, R. N., and Linden, L. L. (2012). "Discrimination in Grading." *American Economic Journal: Economic Policy, 4 (4): 146–68.* https://doi:10.1257/pol.4.4.146

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77(1), 81-112.* https://doi.org/10.3102/003465430298487

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays. *[Preprint in ArXiv], abs/2304.14276*. https://doi.org/10.48550/arXiv.2304.14276

Hovardas, T., Tsivitanidou, O. E., & Zacharia, Z. C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education, 71*, *133-152*. https://doi.org/10.1016/j.compedu.2013.09.019

Ipinnaiye, O., & Risquez, A. (2024). Exploring adaptive learning, learner-content interaction and student performance in undergraduate economics classes. *Computers & Education, 215*, 105047. https://doi.org/10.1016/j.compedu.2024.105047

Jansson, J., & Tyrefors, B. (2022). Grading bias and the leaky pipeline in economics: Evidence from Stockholm University. *Labour Economics, 78*, Article 102212. https://doi.org/10.1016/j.labeco.2022.102212

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence, 2*, 100017. https://doi.org/10.1016/j.caeai.2021.100017

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal, 6*(1), 100048. https://doi.org/10.1016/j.nlp.2023.100048

Keuning, H., Jeuring, J., & Heeren, B. (2018). A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education, 19*(1). https://doi.org/10.1145/3231711

Koubaa, A. (2023). GPT-4 vs. GPT-3.5: A concise showdown. *Preprints*,

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational researcher*, *49*(4), *241-253.*

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2019). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education, 45(2), 193–211*. https://doi.org/10.1080/02602938.2019.1620679

Nakavachara, V., Potipiti, T., & Chaiwat, T. (2024). Experimenting with generative AI: Does ChatGPT really increase everyone's productivity? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4746770

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations, and future scope. *Internet of Things and Cyber-Physical Systems, 3*, *121-154*. https://doi.org/10.1016/j.iotcps.2023.04.003

Reuters (2023) "ChatGPT sets record for fastest-growing user base - analyst note", https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Roselli, D., Matthews, J., & Talagala, N. (2019). Managing bias in AI. In *Companion Proceedings of the 2019 World Wide Web Conference*, *539-544*. Association for Computing Machinery. https://doi.org/10.1145/3308560.3317590

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Thomson, T., & Thomas, R. (2023). Ageism, sexism, classism, and more: 7 examples of bias in AI-generated images. *The Conversation*. https://theconversation.com/ageism-sexism-classism-and-more-7-examples-of-bias-in-ai-generated-images-208748

Valero Haro, A., Noroozi, O., Biemans, H. J. A., & Mulder, M. (2018). The effects of an online learning environment with worked examples and peer feedback on students' argumentative essay writing and domain-specific knowledge acquisition in the field of biotechnology. *Journal of Biological Education, 53*(4), *390–398.* https://doi.org/10.1080/00219266.2018.1472132

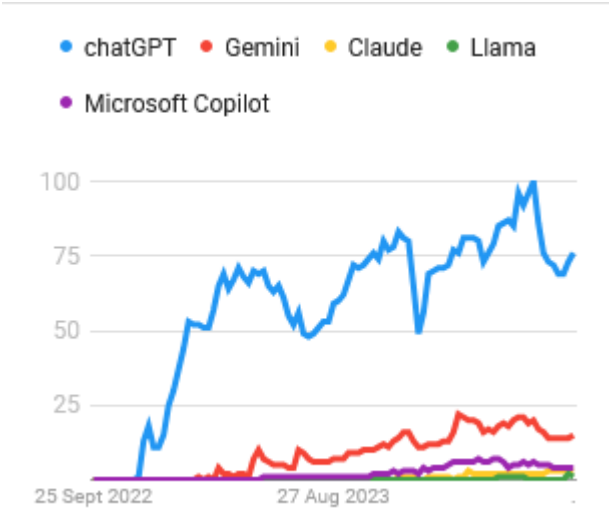
van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review, 30*(5), *1045–1058.* https://doi.org/10.1016/j.econedurev.2011.05.008

Wang, S., & Zhang, D. (2020). Perceived teacher feedback and academic performance: The mediating effect of learning engagement and moderating effect of assessment characteristics. *Assessment & Evaluation in Higher Education, 45*(7), *973-987.* https://doi.org/10.1080/02602938.2020.1718599

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Appendix A

Figure A1: Google Search of LLMs Worldwide from September 2022 to August 2024



Source: Google Trend (06/08/2024)

Figure A3A: Experimental Structure: Year 2 Course

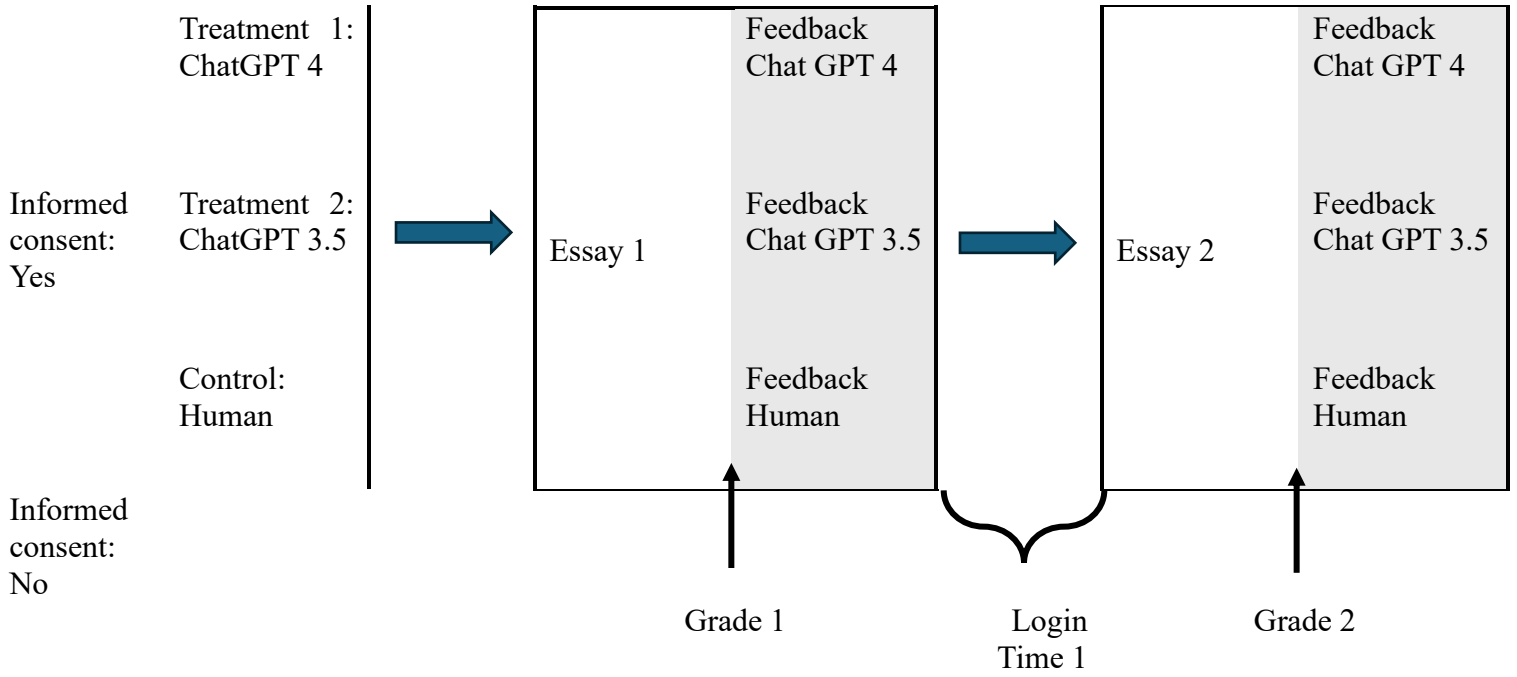


Figure A3B: Experimental Structure: Year 3 Course

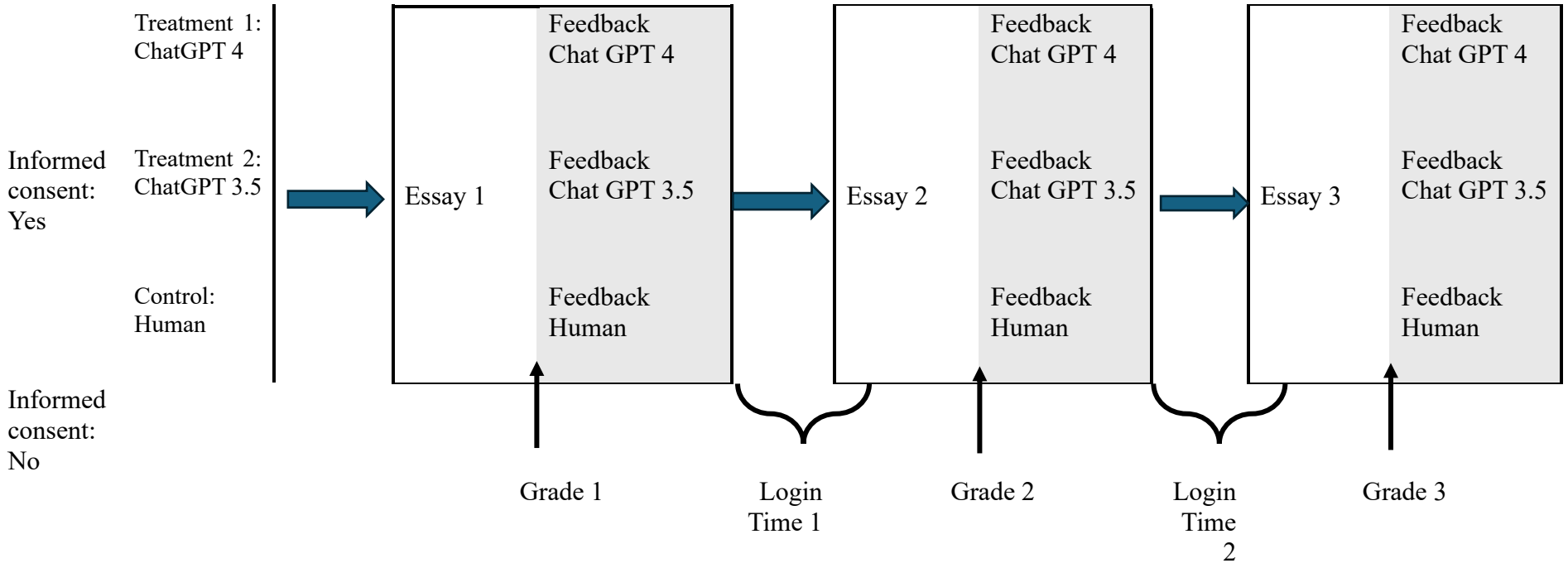


Table A4: Effect of Feedback Source on Accessing Feedback (robustness check)

	ITT	ITT if submitted	ATT	ATT if submitted
ChatGPT 3.5	-0.121	-0.128	-0.136	-0.138
	(0.084)	(0.090)	(0.094)	(0.094)
ChatGPT 4.0	-0.114	-0.119	-0.087	-0.087
	(0.090)	(0.095)	(0.095)	(0.095)
Normalised Grade	0.189***	0.234***	0.197***	0.204***
	(0.021)	(0.057)	(0.051)	(0.061)
Observations	174	164	145	144
Control	Yes	Yes	Yes	Yes

Note: OLS estimates. Additional controls include gender, nationality, ethnicity, year 1 grade and indicator for assessment. The standard errors, reported in parentheses, are clustered at the student level. Symbols: * p<0.1, ** p<0.05, *** p<0.01

"ITT if submitted" restrict the sample to students who submitted the assessment of interest; "ATT" restricts the sample to students who check feedback on the previous assessment; "ATT if submitted" restricts the sample to students who check feedback previous first assessment and submitted the assessment of interest"

Appendix B: Prompts

B1. Prompt used to generate feedback for each assignment:

Prompt used to generate feedback for each assignment (except the Draft essay for Year 2 Course)

"Give detailed critical feedback for the following essay. Treat it like it was submitted to Oxford University, i.e., be very strict.:
[Pasted essay]

Create the feedback based on the following criteria:
[Pasted marking criteria]"

B2. Prompt for Draft essay in Year 2 course

Please evaluate the strengths and weaknesses of the below draft essay based on the following necessary elements. Based on the weaknesses of the draft essay, please recommend improvements in the draft before the final submission.

1. **Introduction**. State your research question and write down why you think we need to know its answer. If you think it is an interesting question, who else might care? Why would answering this question make a difference to anyone: not only to you but also to society? In other words, motivate your study.
2. **Literature review**. Review how others have been studying the same or very similar questions before and place your ideas within a research field: Compare and contrast the views in the literature discussing your research question or similar questions.
3. **Methodology**.
    a. Restate your research question as a hypothesis. Based on your hypothesis (which is a restatement of your research question, ready for testing), write down your empirical model, justify its appropriateness, and run the model using appropriate data. The model will be a very simple version of one of the empirical models you have seen in the literature you are citing. Don't forget to spell out the level at which you observe the variation in the data (the _i or the _it subindices), and explain what the variables in your model mean.
    b. State why you think this is a good model to test your hypothesis.
    c. Describe the data in a separate subsection. Present summary statistics. Make sure you have used firm-level data such as the World Bank Enterprise Performance survey data.
    d. Test your hypothesis.
4. **Results**. Summarize your results in a regression table. Both the summary statistics table (the one from section 3c. above) and the regression results table will need to be professionally formatted. You will be penalized for copying and pasting R/Stata/Excel

code or results directly. Do not do this. Please imitate a table with regression results from one of the papers you are citing. We will further discuss what this means during the lectures. Then, discuss the results within the context of the previous literature. This means that you need to compare your results to the ones in the literature. Importantly, you also need to try to explain where the differences come from (Why do you think you see the differences between your results and the ones published earlier?). Read previous papers to see how they are doing this bit in their results presentation.

5. **Conclusion**. Conclude your research project with a summary of what you did, how you did it, and who cares (the So-What? question).
6. **References**. You must use APA or Harvard or MLA referencing format when citing your sources. Do not mix the formats when citing – if you pick one format, stick to it.

Please include the following text **verbatim** at the end of your feedback: "*If your final essay does not use firm-level data as instructed on Day 1, you would be misinterpreting the assignment, which will have a detrimental impact on your mark. This is because you would not plan to use the Enterprise Surveys data, which is a key requirement in this assignment. When you do get familiar with the ES data, please focus on something that is feasible. Make sure you log the monetary (sales) and nominal numbers (e.g. number of employees) for ease of comparison. You could interact the size of the firm with a Policy variable to see any differential impact across firms of different size. See Stankov and Vasilev (2019) Reform outcomes paper for further insights. Note the formatting requirements for regression output, which explicitly ban the direct copy-paste of R/Stata/Excel output. Note the other modelling and table formatting requirements as well. You could also include country dummies, but do not run those regressions separately for the countries of choice - best to include all available countries! Robust standard errors, please. Check the approaches in the literature for more info on how to estimate the Policy effect. If you need more advice, please come see me in my office hours for more feedback.*"

B3. Prompt used to generate grades for each assignment:

Prompt used to generate grades for each assignment (except the draft essay for Year 2 Course)

"Grade the following essay on a scale 0-100 and give detailed feedback. Grade this essay like it was uploaded for Oxford University, so be very strict while grading. Give a numerical grade for each individual criterion:
[Pasted essay]

To grade it use the following instructions:
[Pasted marking criteria]"

B4. Prompt used to generate grades for draft essay for Year 2 Course:

"The following text is the draft of the paper:
[Pasted draft essay]

Based on this draft place an aggregate score from 0-100 grading the likelihood that the student will finish a good empirical project that will meet the following criteria:
[Pasted marking criteria for the next assignment]

Grade this draft essay like it was uploaded for Oxford University, so be very strict while grading. Give a numerical grade for each individual criterion"

B5. Grading criteria for Year 2 Assignment:

"(10 marks) Introduction. State your research question and write down why you think we need to know its answer. If you think it is an interesting question, who else might care? Why would answering this question will make a difference to anyone: not only to you but also to society? In other words, motivate your study.
(20 marks) Literature review. Review how others have been studying the same or very similar questions before and place your ideas within a research field: Compare and contrast the views in the literature discussing your research question or similar questions.
(30 marks) Methodology. Restate your research question as a hypothesis. Based on your hypothesis (which is a restatement of your research question, ready for testing), write down your empirical model, justify its appropriateness, and run the model using appropriate data. The model will be a very simple version of one of the empirical models you have seen in the literature you are citing. Don't forget to spell out the level at which you observe the variation in the data (the _i or the _it subindices), and explain what the variables in your model mean. State why you think this is a good model to test your hypothesis. Describe the data in a separate subsection. Present summary statistics. Make sure you have used firm-level data such as the World Bank Enterprise Performance survey data. Test your hypothesis.
(30 marks) Results. Summarize your results in a regression table. Both the summary statistics table (the one from section 3c. above) and the regression results table will need to be professionally formatted. You will be penalized for copying and pasting R/Stata/Excel code or results directly. Do not do this. Please imitate a table with regression results from one of the papers you are citing. We will further discuss what this means during the lectures. Then, discuss the results within the context of the previous literature. This means that you need to compare and contrast your results to the ones in the literature. Importantly, you also need to try to explain where the differences come from (Why you think you see the differences between your results and the ones published earlier?). Read previous papers to see how they are doing this bit in their results presentation.
(10 marks) Conclusion. Conclude your research project with a summary of what you did, how you did it, and who cares (the So-What? question)."

B6. Grading criteria for Year 3 Assignment 1:

"(~150 words, 20 marks) Introduction: Summarize the major issue your field is trying to address: what is the common research question of those papers (they should have a similar research question)? State why this question is important to address – what is the policy problem or a real-life problem that is solved by answering this question? Finally, clearly state your hypothesis related to the research question: a statement which can be tested by using models and data, and which can be answered by either 'Yes' or 'No'. Then, go back to the literature.

(~850 words, 80 marks) Critical Literature Review:

a. Most of the debates in any field have numerous viewpoints, or at least nuances of the same viewpoint. Summarize the viewpoints on either side of the debate. Do not forget to reference properly by using a consistent referencing format across your work.

b. Clearly state what the literature is missing. This is a crucial part for your subsequent work. This is the room for your contribution – your potential innovation in the field. Specifically, criticize the literature on the basis of missing a recent trend or the need for a novel viewpoint on the same issue, or new data availability that can change the way we look at an old problem. In addition, you could criticize the literature based on purely methodological grounds – using old data or outdated methodology, or both.

3. References: provide a list of references you used. You must use APA or Harvard or MLA referencing format when citing your sources. Do not mix the formats when citing – if you pick one format, stick to it. Penalties of up to 10 marks (out of 100) apply for mis-specified citations and references."

B7. Grading criteria for Year 3 Assignment 2:

"(10 marks) The first several sentences reiterate the hypothesis or, at the minimum, the research question. They are followed by a clear description of methods used, a specification of the empirical model used. The marker evaluates if the model is appropriate to study the research question at hand.

(10 marks) The description of the data is given in a separate subsection or paragraph and contains sufficient assurance that the model can be used to generate results based on the data presented

(10 marks) The variation in the data matches the variation in the model

(10 marks) Summary statistics table and a separate table of regression results are presented

(10 marks) The regression table contains a title, regression coefficients and standard errors for each coefficient, significance stars, and notes detailing the variables in the table. The table is presented in a professional or near-professional format. Snapshots of R/Stata/Excel output receive 0 marks.

(10 marks) The discussion of the results contains a verbal confirmation that the main hypothesis is rejected / not being rejected, and how that maps to the research question.

(20 marks) The discussion presents an analysis of the reasons the author sees the magnitude and significance of the presented estimates.

(20 marks) The paper concludes with a summary of what the main goals of paper are, how the methodology achieves those goals, are the hypotheses confirmed, why, and what the implications are from the presented work."

B8. Grading criteria for the Year 3 Assignment 3:

"(10 marks) Summarises both homework 1 and homework 2 in a coherent way, i.e. as one big idea that has been completed. If your HW2 tackles different problem from the one discussed in HW1, give priority to the work done in HW2 as it has the potential for policy impact.

(10 marks) Uses minimal professional jargon, and only when unavoidable.

(10 marks) Compares results and conclusions to published references from theoretical and/or empirical articles.

(30 marks) Includes a reflection on how your results could change policies, political processes, or societal/community practices.

(10 marks) The reflection in stems from your own research work, not the work of others.

(10 marks) Uses correct grammar and punctuation.

(5 marks) Articulates the point within the suggested word limit (500-700 words).

(10 marks) Expresses clearly own opinion throughout.

(5 marks) References the work done by others, as well as own work if and when appropriate."

Appendix C: Ethics

C.1. Invitation email

Dear ,

[The Department] has designed a pioneering project on understanding the impact of artificial intelligence (AI) on student marks, quality of feedback and student satisfaction. This is an important question as it could potentially lead to large-scale efficiency gains across the higher-education sector, as well as a broader increase of the quality of marking and feedback, further eliminating conscious and subconscious biases in marking, and ultimately raising the value of your degree. You can become part of this project now.

Participation in the project is entirely voluntary and participants may opt out of it at any point by sending an email to the project lead, [anonymised], at [email address]. In addition, none of

your interim or final marks will depend on agreeing to take part. No extra work will be required from you. I will only ask you to fill out a 1-minute survey on how satisfied you are with the feedback given on assessments in your [name of course] course.

To learn more and give your explicit consent to take part, please click here.

If you have any questions, feel free to get in touch. I hope to share my findings with you after the end of the 2023/24 assessment cycle, as well as on high-level assessment and teaching-focused conferences.

Thanks in advance for your help, and best wishes in your new term,

[Project lead]

C.2. Participant Information Sheet

Prospective Research Participant Information Sheet

Economics Department

[Name of University]

Project Title: AI, Marking and Feedback

Principle Investigator (PI)'s name and email address: [anonymised]

Primary Researcher: [anonymised]

Project telephone number: [anonymised]

Introductory paragraph

You are being invited to take part in a research project. Before you decide to consent to take part it is important for you to understand why the research is taking place and what your participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Please consider carefully whether you wish to consent to take part.

Why is this research being conducted?

The primary purpose of the survey is for your instructor(s) and the researchers to evaluate the effectiveness of feedback provided by Gen-AI against the feedback provided by a human. The study is conducted for research purposes only. This is an important project for two reasons. First, it is designed to boost the quality of your feedback, eliminate subtle conscious or subconscious biases in teacher marking, and ultimately contribute to the equality, diversity and inclusion in higher education. Second, it could potentially lead to large-scale efficiency gains across the higher-education sector.

Why have I been invited to take part?

You are being invited, as you are a student to one of the project co-investigators: [name], who has developed interest in making teaching more effective. You are also invited because the project needs more students in our sample.

Do I have to take part?

No. It is up to you to decide whether you wish to take part or not. You can withdraw from the study at any time, without any consequences and without needing to give a reason, and you can withdraw your data until 31/08/2024 by contacting [name] at [email]. After this date the research will be submitted for assessment and withdrawing your data will no longer be possible. Note that your data will never be published or traced back to you – it will become part of an anonymised data set, where each student is given a number, but their answers are not matched to a name, email address or any other personally identifiable information. If you decide that you wish to withdraw your data from the project, please drop an email, without explaining your reasons to request a withdrawal, to [name] at [email].

What will my participation involve?

After giving your consent to take part, you will receive feedback from a human or a machine. You will not know whether you have received feedback from a human or a machine. After receiving the feedback, you will be asked to fill out a 1-minute survey on how satisfied you are with the feedback. No writing will be required. This will be done twice during the term. That's it, really.

What are the possible disadvantages and risks in taking part, and how might these be mitigated?

We do not anticipate any risks from participating in this research.

Are there any benefits in taking part?

The main benefit of participating in this research is the opportunity to contribute to future instructional innovation in economic education.

Payments No payments will be disbursed for taking part in this survey.

What information about me will be collected and why is the collection of this information relevant for achieving the research objectives?

Data on name, surname, email, student ID and additional observable data in College administrative records will be linked to your grades in this course and to whether you received feedback from a human or a machine. Identifiable data (including consent forms) will be stored on College servers until 31/08/2024 only. Once anonymised, the data will be stored indefinitely after publication for replication purposes.

The PI will have access to the full raw data. Once anonymised, the data may be shared with the global research community. The data will never leave College servers in a raw form (i.e. in a form that can identify you personally). Whenever it leaves College servers, it will do so in an anonymous form.

The anonymous research data may be transferred to, and stored at, a destination outside the UK and the European Economic Area. Identifiable data will be removed whenever possible and any data transfer will be undertaken with a similar level of data protection as required under UK law.

We would like your permission to use the anonymous data in future studies, and to share this with other researchers (e.g. in online databases). Your answers will never be traced back to you, once anonymised.

How will the results of my participation be used? Will the research be published? Could I be identified from any publications or other research outputs?

The findings from the research will/may be written up in academic publications and policy documents. It will never be possible to identify you personally in the data sets associated with those publications.

Who is funding the research?

[University name] has not received external funding for this research. The research is funded locally, by the [name of] Research Fund of [University name].

Who do I contact if I have a concern about the research or I wish to complain?

If you have a concern about any aspect of this study, please contact [name] or [University name]'s Research Ethics Committee via [email address]. If you wish to make a formal complaint, please email [email address].

Ethical Approval

This study has received ethics approval from [University name]'s Research Ethics Committee.

Data protection

This research commits to abide by the Data Protection Act (2018). For detailed information about what this means for research participants, please visit the Research Participant Privacy Notice: [Link]

General Data Protection Regulation Statement

Important General Data Protection Regulation information (GDPR). [University name] is the sponsor for this study and is based in the UK. We will be using information from you in order to undertake this study and will act as the data controller for this study. This means that we are responsible for looking after your information and using it properly. Any data you provide during the completion of the study will be stored securely on hosted on servers within the European Economic Area'. [University name] is designated as a public authority and in accordance with the [University name] College Act [year of incorporation] and the Statutes which govern the College, we conduct research for the public benefit and in the public interest. [University name] has put in place appropriate technical and organisational security measures to prevent your personal data from being accidentally lost, used or accessed in any

unauthorised way or altered or disclosed. [University name] has also put in place procedures to deal with any suspected personal data security breach and will notify you and any applicable regulator of a suspected breach where legally required to do so. To safeguard your rights, we will use the minimum personally-identifiable information possible (i.e., the email address you provide us). The lead researcher will keep your contact details confidential and will use this information only as required (i.e., to provide a summary of the study results if requested and/or for the prize draw). The lead researcher will keep information about you and data gathered from the study, the duration of which will depend on the study. Certain individuals from [University name] may look at your research records to check the accuracy of the research study. If the study is published in a relevant peer-reviewed journal, the anonymised data may be made available to third parties. The people who analyse the information will not be able to identify you. You can find out more about your rights under the GDPR and Data Protection Act 2018 by visiting [University name intranet website] and if you wish to exercise your rights, please contact [email address].

NB: You may retain this information sheet for reference and contact us with any queries.