Augment or Automate? An Early Field Experiment with Generative AI in Hiring *

Kobbina Awuah† Urša Krenk‡ David Yanagizawa-Drott§

November 11, 2026

Abstract

Can generative AI improve hiring? We experimentally embed GPT-4 into a teacher-recruitment screening process in Ghana, comparing three pipelines: human-only evaluation, human with AI assistance, and fully automated screening. Automation increases offer and hiring rates by 84% and 73%, respectively, over the human-only baseline. Human evaluators apply idiosyncratic standards, while the AI grades more consistently and more generously, advancing more applicants to the in-person assessment. In contrast, AI assistance yields no detectable improvement, as evaluators ignore most of the AI's recommendations. Taken together, our results suggest that, at least in this context, automation outperforms human-AI collaboration.

Keywords: Artificial intelligence, organizational economics, development economics, labor markets.

*We are grateful to Maria Korobeynikova, Minh Trinh, Andrin Pluess, and Alessandro Vanzo for excellent research assistance. We also thank seminar participants at Collegio Carlo Alberto, UC Berkeley Development Lunch, AI in Social Science Conference at the University of Chicago, ESA Helsinki, Workshop in AI + Economics at the University of Zurich, Stone Inequality Seminar at UBC, the Center for Behavioral Institutional Design Seminar at NYU Abu Dhabi, AFE Chicago, AI & Future of Human Capital Symposium at the World Bank and Georgetown University, the University of Gothenburg, Bocconi University, Paris Empirical Political Economy Seminar, HEC Lausanne, CERGE-EI, the Hoover Institution at Stanford, and Swiss Young Economists Meeting in Zurich. The experiment reported in this study can be found in the AEA RCT Registry (#0011651).

†University of Zurich: Email: kobbina.awuah@econ.uzh.ch

‡University of Zurich: Email: ursa.krenk@econ.uzh.ch

§University of Zurich. Email: david.yanagizawa-drott@econ.uzh.ch

1 Introduction

Firms screen large numbers of job applicants using noisy signals, and the quality of that screening shapes workforce composition and organizational outcomes. Algorithms could improve screening, for example by applying evaluation criteria uniformly and thereby reducing the noise that arises when different evaluators apply different standards. Yet, when using algorithms as assistants, human decision-makers routinely override algorithmic recommendations, and field evidence suggests these overrides often reduce rather than improve decision quality (Hoffman, Kahn, and Li, 2018; Vaccaro, Almaatouq, and Malone, 2024). Recent advances in generative AI have begun to transform candidate screening by enabling large-scale, structured evaluation of written and spoken materials. Whether such tools are best deployed as replacements or aids for human evaluators is an open question with direct implications for organizational design and labor market outcomes. In this paper, we contribute experimental evidence on this question by comparing full automation, AI assistance, and human-only evaluation. To do so, we embed GPT-4, one of the first generative AI models to achieve widespread commercial adoption, into a real-world teacher recruitment process in Ghana in the months following its March 2023 release.

We partnered with a nonprofit organization that recruits recent university graduates for teaching fellowships in rural schools in Ghana and embedded GPT-4 into its hiring process. At the initial screening phase, which consists of grading short essays written by applicants based on a fixed set of criteria defined by the organization, we randomly assigned applications to one of three pipelines relevant for policy-making, each differing in how applications are graded: (1) *Human-Only*: the current status quo, where evaluators relied solely on their own judgment, and only their grade determined whether a candidate advanced to the next stage of the hiring process; (2) *Human-with-AI-Assistance*: augmentation of human workers with GPT-4, where evaluators first recorded an initial grade, then reviewed a GPT-4-generated grade recommendation before finalizing their evaluation, with the final grade deciding advancement; and (3) *AI-Only*: where GPT-4 fully automated the evaluation, and its assessment alone determined which candidates advanced.[1] Prior work has compared augmentation and automation pipelines for traditional machine learning models in real-world settings (Agarwal et al., 2024), or has studied the productivity effects of generative AI assistance in workplace tasks (Noy and Zhang, 2023; Brynjolfsson, Li, and Raymond, 2025), but whether these designs improve actual hiring decisions, and how the three pipelines compare, has not been tested in a real-world recruitment process using generative AI. We conducted, to our knowledge, the first randomized controlled trial to test these three pipelines using generative AI in a real-world labor market, where grading decisions directly affected candidates' likelihood of receiving a job offer.[2] While the pipeline used for advancement was randomly assigned, we

[1] We prompted GPT-4 to not only provide the grading score but also the rationale for the score. In the assistance pipeline, we randomized whether the rationale was visible to the human user. We describe this in detail in Section 2.

[2] The distinction between "traditional" supervised machine learning and generative AI is important. The former

employed parallel grading: every application was evaluated independently by both a human (with or without AI assistance) and by the AI, enabling us to compare grades for the same sets of essays.

We document that full automation raises offer rates by 84% and hiring rates by 73% relative to the *Human-Only* baseline. By contrast, providing evaluators with the same algorithm as an assistant yields no significant improvement in offer or hiring rates, because the evaluators largely ignore the AI recommendations. Because AI grades are higher on average, the AI pipeline advances more candidates past the application cutoff, and a large part of the increase in the offer and hiring rates comes from this extensive margin. When we fix the number of candidates advanced in each pipeline to the number advanced in the *Human-Only* pipeline, AI-screened candidates remain 35% more likely to receive an offer, although this estimate is less precise.

Beyond assigning higher grades, the AI also grades more consistently. Human evaluators apply the rubric inconsistently: two independent graders award the exact same grade (on a discrete 1–5 scale) in only 44% of cases, and among similar essays the grade variance is 0.67.[3][4] AI grades are substantially more consistent: the grade variance among similar essays is 33% lower, and no human grader reaches the AI's level of consistency. With more consistent grading, the screening criteria set by the organization, rather than the luck of evaluator assignment, determine who advances.

In a standard measurement error framework, idiosyncratic grading attenuates the correlation between application grades and any external measure of quality. Given a rubric that is designed to capture the qualities the organization values, more consistent application of that rubric should produce more informative grades (Kahneman, Sibony, and Sunstein, 2021). We therefore examine how application grades relate to performance in the in-person assessment. The in-person assessment is conducted by independent evaluators who are blind to both application grades and treatment assignment, and evaluates candidates through problem-solving exercises, group activities, mock teaching, and interviews. The organization treats assessment performance as its primary measure of candidate quality: it determines final hiring decisions and informs placement of fellows to regions where strong performance is most needed. Among candidates who attended the assessment, the correlation between application grades and assessment performance is 0.34 in the *AI-Only* pipeline, compared to 0.07 and 0.12 in the *Human-Only* and *Human-with-AI-Assistance* pipelines. These correlations are computed on small and selected samples, so we see them as suggestive evidence.

uses predictive models trained on human-labeled data drawn from the same distribution as where it is deployed. By contrast, Generative AI is pre-trained on vast amounts of data from the internet and from human-labeled data across many domains, using a fundamentally different architecture. Open-source models can in principle be further fine-tuned on domain-specific data, but this is very rare in practice since the vast majority use closed-source models like ChatGPT, GPT-4, GPT-4o, Claude 4 Sonnet, Gemini 2.5 Pro, etc. (Some closed-source providers also allow for some fine-tuning, but if they do it is typically in very restrictive ways.) Here, we use the term in the sense of using an off-the-shelf generative AI model, GPT-4, without any fine-tuning on the specific task or domain.

[3] This implies a standard deviation of 0.82, nearly a full grade on a 1–5 scale.

[4] We measure essay similarity using cosine similarity of vector embeddings (Voyage AI `voyage-large-2-instruct`, 1,024 dimensions). For each essay, we identify its 20 nearest neighbors and compute the variance of their grades. Results are robust to $k \in \{50, 100\}$.

Why do evaluators in the *Human-with-AI-Assistance* pipeline fail to incorporate AI recommendations? Evaluators override the recommendation in over 80% of cases where their grade and the AI grade differ, and the pattern is strongly asymmetric: evaluators override the algorithm in almost 87% of cases when AI suggests a higher grade, and 45% of the time when it suggests a lower grade. This reluctance to award higher grades is consistent with a conservative grading bias. An additional factor associated with higher override is the context in which the experiment took place: for the first time, evaluators encountered large numbers of essays (about 45%) generated with the help of ChatGPT (we refer to these as "LLM essays").[5] Both AI and human evaluators award higher grades to LLM essays, which tend to be longer and more polished, but the premium is smaller for human evaluators, plausibly because evaluators recognize that part of the quality of the response reflects the tool rather than the candidate. When grading LLM essays with AI-assistance, evaluators are also roughly 27% less likely to revise toward the AI recommendation compared to non-LLM essays.

Several caveats apply. Our experiment took place in the early days of generative AI, when evaluators had little experience with the technology and LLM-generated essays were a novel phenomenon; both have likely shifted since. Our screening task is essay-based, which is less common than CV-based screening, and we study a single organization in Ghana using an early model (GPT-4) without task-specific fine-tuning. We do not observe downstream measures of teacher effectiveness such as teacher value-added. However, the in-person assessment, which the organization uses to make final hiring and placement decisions, provides a meaningful benchmark for candidate quality that is independent of the application stage. The finding that AI grades are more consistent may generalize beyond our setting, since algorithmic evaluation applies the same criteria to every candidate, while human grading noise is well documented across many contexts (Kahneman, Sibony, and Sunstein, 2021).

Our findings contribute to multiple strands of literature. First, we contribute to the economics of algorithmic hiring and selection. A large body of work documents that algorithms can improve hiring decisions or reduce bias in hiring (Agan et al., 2023; Avery, Leibbrandt, and Vecci, 2024; Avery et al., 2026; Jabarian and Henkel, 2026; Li, Raymond, and Bergman, 2026) relative to human judgment, and that managers who override algorithmic recommendations tend to select lower-quality candidates (Hoffman, Kahn, and Li, 2018; Cowgill, 2020; Chalfin et al., 2016). Our paper extends this literature to generative AI: rather than predictive models trained on firm-specific data, we study an off-the-shelf large language model applied to essay grading, a setting in which the algorithm has no task-specific training. Most closely related is Jabarian and Henkel (2026), who show that AI-powered voice interviews improve hiring outcomes by standardizing information collection.

[5] Using a state-of-the-art detection tool, we find that approximately 60% of applicants submitted at least one LLM-generated essay. We use Pangram Text (Pangram Labs), which achieves 99.85% accuracy and a 0.19% false positive rate (Emi and Spero, 2024). Jabarian and Imas (2025) evaluate multiple commercial and open-source LLM detectors and find that Pangram outperforms all others.

Our setting differs in a key respect: rather than using AI to collect information from candidates, we use AI to evaluate existing written materials, where what matters is not how information is gathered but how it is assessed. We directly compare *AI-Only* and *Human-with-AI-Assistance* pipelines against the *Human-Only* baseline, and find no evidence that human overrides improve selection, in line with Hoffman, Kahn, and Li (2018)'s finding that overrides reflect evaluator error rather than superior information. A related literature studies AI on the applicant side: Wiles, Munyikwa, and Horton (2025) show that algorithmic writing assistance on jobseekers' resumes increases hiring.

Second, we contribute to the growing literature on automation and augmentation using generative AI in the workplace. Prior work has examined traditional machine learning methods (Angelova, Dobbie, and Yang, 2026), compared augmentation and automation pipelines using conventional algorithms (Agarwal et al., 2024), or studied generative AI exclusively as an augmentation tool without a full-automation arm (Brynjolfsson, Li, and Raymond, 2025; Noy and Zhang, 2023). The broader question of whether AI should replace or assist workers, and under what conditions each is preferable, has significant economic implications given that generative AI is expected to affect a broad spectrum of occupations (Eloundou et al., 2024; Acemoglu, Autor, and Johnson, 2023; Acemoglu and Restrepo, 2019). Our study is, to our knowledge, the first randomized trial to compare all three policy-relevant pipelines using generative AI in a real-world economic setting with high stakes.

Third, while recent work documents that generative AI can boost productivity in tasks like coding, writing, and consulting (Brynjolfsson, Li, and Raymond, 2025; Bubeck et al., 2023; Dell'Acqua et al., 2026; Noy and Zhang, 2023; Peng et al., 2023; Kumar et al., 2025), we show that a highly capable model does not ensure improved outcomes unless users incorporate its output effectively. This relates to literature exploring why people systematically disagree with AI recommendations, due to factors such as algorithmic aversion (Dietvorst, Simmons, and Massey, 2015), bias against AI-generated content (Parshakov et al., 2025), priors that are far from algorithmic recommendations (Kim et al., 2024), cognitive constraints (Agarwal et al., 2024), or the tendency to overgeneralize AI performance across different domains (Vafa, Rambachan, and Mullainathan, 2024).

Finally, our research is among the first, alongside Otis et al. (2026), to study generative AI in a developing-country labor market. This context matters as organizations in these settings often have less standardized evaluation infrastructure and fewer resources for evaluator training, making human grading likely noisier and the potential gains from algorithmic evaluation larger. At the same time, decision-makers may have less prior exposure to algorithmic tools, which could amplify the override patterns we document. Most related research in these regions has focused on AI-powered educational tools designed to improve learning outcomes (De Simone et al., 2025).

The rest of the paper is structured as follows. In Section 2, we explain the setting and the experimental design. In Section 3, we cover a simple conceptual framework to structure the thinking around the signal extraction problem in the three policy pipelines. Section 4 covers the data and

estimation framework. Section 5 presents the main results. Section 6 compares the consistency of human and AI grades and explores reasons for the high algorithmic override in the pipeline with humans-in-the-loop. Section 7 concludes.

2 Background and Experimental Design

2.1 Background and the Organization's Recruitment Process

We collaborate with a Ghanaian educational non-profit organization. The organization recruits recent university graduates and places them in disadvantaged rural schools nationwide for a two-year teaching fellowship program. Prior teaching experience is not required, but candidates must hold at least a bachelor's degree before starting the program. The organization provides extensive pre-placement training and on-the-job support throughout the two-year fellowship. Candidates can apply for this position as either a regular job or as part of their compulsory "National Service".[6] Every year, a cohort of between 50 and 150 fellows is assigned to schools in rural areas. Fellows earn a stipend comparable to the average entry-level salary in Ghana.[7][8] The position is considered prestigious, and the candidate selection process is competitive, with only 15–20% of applicants being offered positions.[9] Importantly, the organization usually sets a number of slots to fill, and if they do not find enough high-quality candidates to fill those slots, the positions remain unfilled. After the program, the majority of fellows (about 60%) stay in the education sector (either working in educational non-profit organizations or as teachers in schools). Among those who leave the education sector, many work for other non-profit organizations or in the public sector.

Figure 1 illustrates the recruitment process for the partner non-profit organization, as well as the design of our policy experiment.[10] On the supply side, potential applicants need to enter the online application portal, register, and answer six essay questions before submitting the application. Once candidates submit their applications, the organization begins the evaluation phase. It is during this phase that our policy experiment, described in detail below, takes place. After the application essays are assessed, applicants who meet a predetermined cut-off score are invited to in-person interviews, after which fellowship offers are given.[11] Recruitment is cyclical and typically occurs

[6] In Ghana, all students who graduated from an accredited tertiary institution are required to complete a one-year civil service, usually in the public sector.

[7] Ghana has 16 regions in total, and the partner non-profit is present in 10 of them.

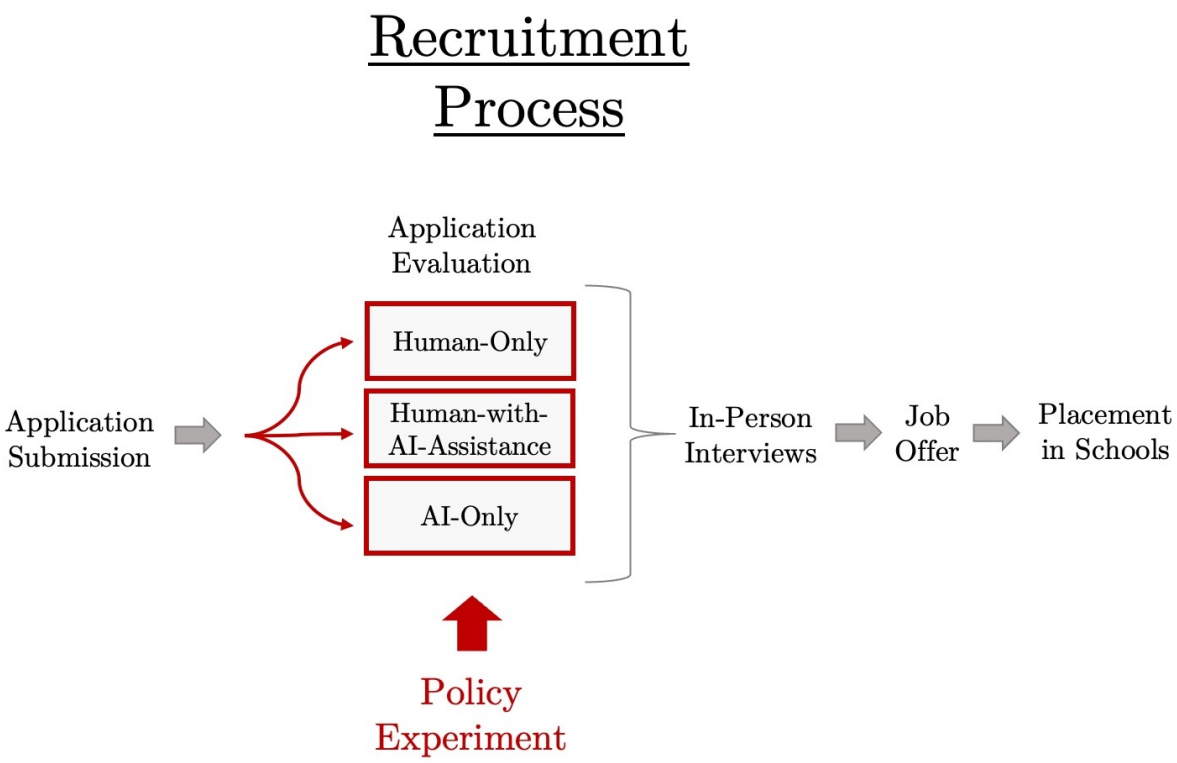
[8] The stipend received during the fellowship exceeds what the person would normally receive during National Service in the public sector

[9] Usually, about 50% of applicants are invited for an interview and between 50% and 75% of those who attend the interview are given offers. Since the in-person interviews are centrally organized in bulk, often scheduled at short notice and on fixed, non-flexible dates, many applicants are unable to attend.

[10] The organization received a total of about 1030 applications, but only a subset of those applications were included in our experiment – a total of 697. About 190 applications were graded outside our platform and about 143 of the received applications were not eligible and were therefore not graded.

[11] After the offer is given and prior to the posting, the candidates undergo a 3-week teaching and leadership training.

Figure 1: Recruitment Process



Notes: The figure shows the recruitment process for the partner non-profit organization. Potential applicants were required to enter the online application portal, register, and answer six essay questions before submitting the application. After application submission, the evaluation phase began on the organization's side, which is where our experiment took place. A total of 697 candidates submitted applications and were included in our policy experiment. These applicants were randomly assigned to one of three evaluation pipelines: Humans-Only, Humans-with-AI-Assistance, or AI-Only. Notably, each application was graded separately by humans (either with or without AI assistance) and independently by AI; afterward, randomization determined which grading method was ultimately used. Out of the 697 applicants, 494 were invited to in-person interviews, 247 attended the interviews, 191 received fellowship offers, and 129 accepted the offers.

between March and July. If a candidate accepts the offer, they begin their fellowship between October of that year and January of the following year.

Details of the Application Questions and Grading The application form consists of six open-ended essay-type questions designed to assess candidates' prior experiences, motivation and alignment with the organization's mission. Applications are assessed by evaluators who are either current non-profit organization employees or program alumni. Essay answers are graded on a scale

If their attendance or their performance at the sessions is not considered sufficient, the offer might still be rescinded, but this does not happen very often.

from 1 (lowest) to 5 (highest), based on clear grading criteria, unknown to the applicants. Applicants who achieve a total score of at least 18 (out of 30) are invited to participate in a subsequent in-person evaluation day. In a typical year, approximately half of the applicants advance to this stage. The grading process is blind; evaluators are unaware of applicants' demographic characteristics beyond those directly relevant to fellowship eligibility, such as education level, national service status, country of residence, and graduation year. Applications of ineligible candidates are not graded.[12]

We provide an overview of the questions and the corresponding grading criteria in Appendix Table A.1. Questions 1-4 are meant to assess how good the applicant's fit is to work for the organization (motivation, educational philosophy, alumni vision, value-alignment), question 5 is meant to be a proxy for "grit", and question 6 is meant to measure the applicant's ability to lead and influence others. The grading criteria for each question are exhaustive, and evaluators are trained to grade the essays strictly according to these criteria. For example, Question 2, which assesses applicants' educational philosophy asked: *"What is an excellent education to you, and how do you intend to provide that to your students?"*. The grading rubric for this question was as follows: *1. Does not define what an excellent education is and does not articulate how to provide that to their students. 2. Defines what an excellent education is but does not articulate how to provide that to their students. 3. Clearly defines what an excellent education is and shows a pathway to providing that to their students. 4. Rubric 3 plus: articulates factors that lead to academic achievement, mindset development, exposure to resources. 5. Rubric 4 plus: gives specific examples of actions they will take as a fellow and alumni to provide an excellent education to their students.*

In-Person Assessment Center The in-person assessment serves as a "fresh start", as the application grades no longer carry any weight. To avoid any grading biases arising from evaluators in the in-person assessment recalling applicants' essays, the evaluators for the in-person assessment are different from those who graded the essays as part of our experiment. Furthermore, neither the evaluators nor the candidates are aware of the treatment status assigned to each candidate's application (i.e., the evaluation was double-blind), and additionally, the evaluators did not know their application grades. The in-person assessment is typically organized about one month after the application portal closes and lasts an entire day. It consists of several components, each evaluated separately: a problem-solving exercise, a group activity, a mock teaching exercise, and an interview. Candidates are scored from 0 to 100 in each category, with equal weight assigned to each component. Those who achieve an average score of 50 (out of 100) are offered a fellowship position. Assessment performance is the organization's key metric for evaluating candidates: beyond determining offers, it guides the placement of fellows across regions, with the strongest performers assigned to schools where the need is greatest.

[12]This is in most cases due to applicants not holding at least a bachelor's degree or not graduating on time

Signal and Noise in Application Grading in the Pre-Experimental Sample A key premise of the recruitment process is that essay grades carry signal about candidate quality. In the pre-experimental 2022 cohort, they do not. The correlation between candidates' percentile rank in application essay grades and their percentile rank in in-person assessment grades is essentially zero ($\rho = -0.04$, Appendix Figure A.1). Noisy grading is one plausible contributing factor: if grades reflect which evaluator happens to read an essay as much as the essay's actual content, they are unlikely to correlate with independent assessments of candidate quality.

To characterize this noise, we measure how consistently similar essays receive similar grades.[13] If grading were consistent, essays that are similar in content should receive similar grades and the grade variance across similar essays should be low. For the pre-experimental 2022 cohort, the grade variance among the 20 essays most similar to any given essay is 0.890, implying a standard deviation of 0.94 grade points, nearly a full grade on the 1–5 scale (Table 2). If noise in grading is part of what drives the low predictive validity, then more consistent grading should improve the quality of candidates selected into the fellowship. In the experiment we conduct and describe below, we will test whether algorithmic grading can reduce this noise.

2.2 Experimental Procedures

Our policy experiment was conducted during the application evaluation phase, that is, after candidates applied and before the in-person assessment center. Figure 1 above illustrates the experiment design. We randomized applications to one of three policy pipelines: *Human-Only*, *Human-with-AI-Assistance*, and *AI-Only*. The pipeline assignment affected the final grade which determines whether candidates advance to in-person interviews. In the Human-Only pipeline, the grade is provided by human evaluators, without any AI input. In the AI-Only pipeline, the grade is provided exclusively by the AI algorithm (on which we provide details in Section 2.3 below). In the Human-with-AI-Assistance pipeline, the grade is provided by human evaluators who receive input from the AI. For half of the applications in this group, the evaluators receive only the AI-generated grade as input (*Human-with-AI-Grade*), while for the other half, the AI grade is accompanied by a rationale (also generated by the AI, *Human-with-AI-Grade-and-Rationale*), explaining why that particular grade was assigned to the response. This allows us to test whether providing a rationale for algorithmic decisions reduces the likelihood of human evaluators overriding the AI's recommendations. It is important to note that, despite the three policy pipelines, every application was actually graded by both humans (with or without AI assistance) and the AI. The randomization determined which of these grades (human, AI, or a combination) counted for advancement into the in-person interviews. Specifically, half of the applications graded by humans without any

[13] For each essay, we embed the response using the Voyage AI `voyage-large-2-instruct` model (1,024 dimensions) and compute cosine similarity to all other essays in the cohort. We identify the k nearest neighbors (excluding the essay itself) and compute the variance of their grades (the neighborhood grade variance). We report results for $k \in 20, 50, 100$; findings are robust across these choices.

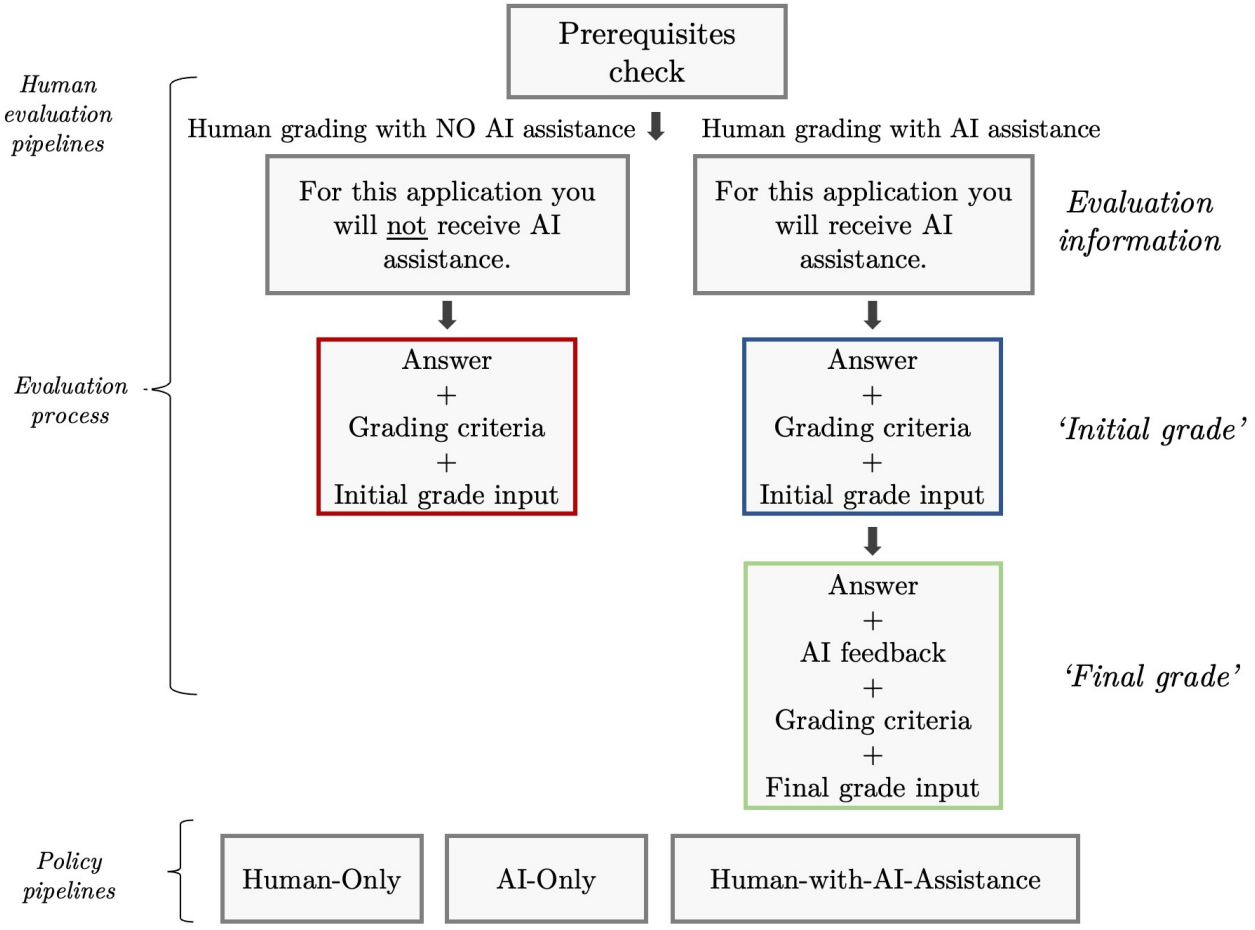
assistance were later randomized into the AI-Only pipeline, meaning the AI-grade was used for advancement. All applications graded by humans with AI-assistance were randomized into the Human-with-AI-Assistance pipeline. Evaluators were aware of this randomization process.

The goal of this design was twofold. First, by randomizing applications into three distinct policy-relevant pipelines, we can evaluate the causal impact of each grading approach on downstream outcomes such as job offer rates and hiring, shedding light on the relative effectiveness of using GPT-4 as an assistant and tool for automation relative to conventional grading. Second, the parallel grading, where each essay is graded by both humans (with or without AI assistance) and the AI, allows us to compare differences between human initial and AI grades, as well as between human final and AI grades, for the same set of essays.

Figure 2 illustrates the process the evaluators followed for grading.[14] The evaluation was blind: evaluators did not observe applicants' demographic characteristics beyond those directly relevant to fellowship eligibility. Evaluators were first shown information that determines applicants' eligibility for the program (i.e., "Prerequisites check"). If a participant failed to meet the eligibility criteria (most commonly, having a "Higher Education Diploma" rather than a Bachelor's degree as their highest level of education), their application was not assessed. Following the eligibility check, evaluators were informed whether they would receive AI assistance with grading. They were then presented with a screen displaying a question, the candidate's answer, and the grading criteria. At the end of this screen, evaluators were required to submit a grade, which we refer to as the "initial grade". After submitting the initial grade, the process differed depending on the random assignment of AI assistance. For applications assigned to receive no AI assistance, evaluators proceeded to the next question. For applications assigned to receive AI assistance, evaluators were shown an additional screen presenting the answer and the grading criteria again, along with the grade suggested by the algorithm. To identify potential mechanisms, in half the cases, evaluators were also provided with a justification for the algorithm's recommendation. At the end of that screen, evaluators were required to re-enter the grade for that question. We call that grade the "final grade". Additionally, we randomly selected around 15% of the applications and submitted them to a different human evaluator, without changing whether they were assigned to receive algorithmic assistance or not. The purpose of this was to check for consistency of grading across human evaluators, but the grades collected during this round were not relevant for the candidate selection process and we do not use them in our main analysis.

[14]For the experiment, all applications were evaluated on the survey platform Qualtrics, replacing the organization's standard evaluation platform. Qualtrics enabled us to track all the outcomes we were interested in, including time spent grading the applications.

Figure 2: Evaluation Process



Notes: Figure illustrates the process the evaluators followed for grading. Evaluators were first shown information determining applicants' eligibility for the program (i.e., "Prerequisites check"). If a participant failed to meet the eligibility criteria, their application was not assessed. After the eligibility check, evaluators were informed whether they would receive AI assistance for grading. They were then shown a screen displaying a question and its answer, along with the grading criteria, and were required to submit a grade (referred to as the "initial grade"). For applications assigned to receive no AI assistance, evaluators proceeded to the next question. For those assigned to receive AI assistance, evaluators were shown an additional screen after submitting their grade. On this screen, they were presented with the answer, grading criteria, and the algorithm's suggested grade. In half the cases, evaluators were also provided with a justification for the algorithm's recommendation. At the end of this screen, evaluators were required to re-enter the grade for that question (referred to as the "final grade").

2.3 The Generation of AI Grades and AI Rationales

To generate the AI grades and rationales used in a subset of applications in the Human with AI-Assistance group, we used OpenAI's GPT-4 model (gpt-4-0314 API). The model was only provided with the organization's grading criteria and asked to grade answers without any prior training on example answers.

GPT-4 Prompt Structure The input to GPT-4 consisted of two parts: a system prompt and a content prompt (a series of messages between "User" and the "Assistant"). Our system prompt (for details see Appendix Section B.1) adhered to best practices in prompting, by explicitly instructing the model to excel at the given task: *"You are an expert recruiter very attentive to detail."* Additionally, the prompt instructed the model to employ step-by-step reasoning to reach its decision, known to enhance model performance (Wei et al., 2022). Finally, it contained instructions on the desired structure for the rationale. We requested a concise explanation for the chosen grade, including reasons for not selecting the adjacent higher or lower grades.[15] The core of the content prompts (for details see Appendix Section B.2) consisted of instructions from the evaluator manual, including the grading criteria for each grade (1 to 5) and definitions for relevant terms (e.g., a specific definition of "resilience and adaptability"). The prompts had the following structure:

1. A brief description of the non-profit organization and the model's task. We clarified that we were assessing applications for a teaching fellowship program, and the task involved grading applicant responses based on provided criteria.

2. Relevant content from the organization's website. For example, we explicitly stated the non-profit organization's mission to the model in this section.

3. The question, its purpose, and its assessment focus. We provided the specific question the candidate had to answer, along with the intended assessment aspect according to the grading manual.

4. The grading criteria. The criteria from the training manual were "augmented"[16] with grade-specific factors. For instance, for question 2, grade 3, the augmented criterion read (the augmented part in italics): "Clearly defines an excellent education and outlines a path to offering it to students. *This includes a) sharing relevant personal experiences and background, b) demonstrating adaptability and flexibility, c) displaying passion and enthusiasm, d) demonstrating clear communication and organization, and e) exhibiting some problem-solving and critical thinking skills.*"

3 Conceptual Framework

The screening task is fundamentally a signal extraction problem: the organization seeks to identify candidates whose unobservable quality meets or exceeds a threshold, using grades as signals. We

[15] After about 200 applications were graded, we slightly modified the format in which the explanation was provided to the evaluator.

[16] The augmentation included incorporating implicit factors that were relevant for each grade, beyond those explicitly listed in the grading criteria. These factors were identified by providing GPT-4 with examples and prompting it to extract the relevant elements for each grade. This approach was designed to help GPT-4 correctly recognize the implicit factors, similar to how human graders received additional training on applying the criteria.

formalize this to clarify how the three experimental pipelines differ in the level and the noise of the grades they produce, and what this implies for screening quality.

The Screening Problem. We can think of teacher quality as an unobservable vector θ consisting of several dimensions $j = 1, \ldots, K$, each influencing student learning through an unknown production function $y = f(\theta)$. The organization views a sufficiently "good teacher" as someone above a quality threshold $\bar{\theta}$, and is willing to give offers to anyone deemed above that threshold.

As described in Section 2.1, the organization screens for quality in two stages. In the first stage, each applicant i submits a written answer X_{ij} to a question along each of $K = 6$ dimensions. The organization specifies grading criteria C_j for each dimension, and the evaluator's task is to assign a grade from 1 to 5 strictly following these criteria:

$$G^*_{ij} = f(X_{ij},\, C_j).$$

An applicant advances to the second stage if their total application grade, the sum of the six assigned grades $G_i = \sum_{j=1}^{6} G_{ij}$, is at least 18:

$$\text{AboveBar}_i \;=\; \mathbf{1}\{G_i \geq 18\}.$$

In the second stage, candidates are assessed through in-depth in-person interviews and exercises. Applicants whose average interview grade $\tilde{\theta}$ is at or above a numerical threshold (50 out of 100) receive an offer. There is no crowding-out: all applicants who surpass the quality bar are offered positions.

Grading Noise and Screening Quality. In practice, human evaluators do not apply the criteria identically. We can decompose the observed grade into a signal component and evaluator-specific noise:

$$G_{ij} = G^*_{ij} + \varepsilon_{ij}$$

where G^*_{ij} is the grade that would result from faithful application of the criteria and ε_{ij} captures evaluator-specific departures. This noise can arise from multiple sources: the grading task is cognitively demanding, requiring evaluators to assess several paragraphs of text against multi-dimensional criteria (Gabaix and Graeber, 2024; Gabaix, 2019); there are no explicit rewards for grading in a particular way and no verifiable way to conclude a grade is incorrect, weakening incentives to provide effort; and evaluators may hold views about what makes a "good teacher" that differ from the official criteria, a classical principal-agent problem. These sources of noise are well-documented across many different contexts (Kahneman, Sibony, and Sunstein, 2021). Relative to the organization's criteria, all such departures are noise. However, whether some of them also

contain information about candidate quality (like knowledge of the local context or a judgment about essays that appear to be written with the help of AI) is an empirical question.

Grading noise has direct consequences for screening quality. In a standard measurement error framework, where ε_{ij} is unrelated to candidate quality, idiosyncratic variation in ε_{ij} attenuates the correlation between application grades and any external measure of candidate quality, including performance in the in-person assessment. At the screening margin, noise increases misclassification: some candidates who would score well in the in-person assessment are screened out, while others who would not are advanced. Under this assumption, reducing the variance of ε should improve the informativeness of grades and, for a given number of applicants advanced, the quality of the resulting candidate pool.

Three Screening Pipelines. Our experiment introduces AI into this screening process through three policy pipelines. The final grade for applicant i on dimension j is determined by:

$$G_{H,ij} = f_H(X_{ij}, C_j), \qquad G_{AI,ij} = f_{AI}(X_{ij}, C_j), \qquad G_{HAI,ij} = f_{HAI}(X_{ij}, C_j, G_{AI,ij}). \tag{1}$$

Here G_H is the Human-Only grade (status quo), G_{AI} is the AI-Only grade (automation), and G_{HAI} is the final human grade after observing the AI recommendation (Human-with-AI-Assistance).

The three pipelines differ in their expected noise properties. In the Human-Only pipeline, grades reflect the idiosyncratic standards documented above. In the AI-Only pipeline, the algorithm receives the same inputs (X_{ij}, C_j) and applies a standardized grading procedure, although repeated evaluations can produce different grades. If the algorithm applies the criteria faithfully, $\text{var}(\varepsilon_{AI})$ should be lower than $\text{var}(\varepsilon_H)$, and AI grades should correlate more strongly with external quality measures. However, the algorithm may lack tacit knowledge about the educational context in rural Ghana that experienced evaluators possess, or it could introduce systematic biases, making the direction of the net effect an empirical question. The pipelines may also differ in the level of grades: a grader that assigns systematically higher grades advances more applicants past the cutoff of 18, which raises offers if the additional applicants meet the bar at the in-person assessment.

In the Human-with-AI-Assistance pipeline, the noise in G_{HAI} depends on whether evaluators incorporate the AI recommendation. If they fully update toward G_{AI}, the variance of ε_{HAI} should be the same as the variance of the AI-only pipeline. However, several forces may work against incorporation. The AI grading function f_{AI} is a black box: evaluators observe only its output, creating ambiguity about how G_{AI} should be interpreted. Human intuitions about AI performance across different tasks are often inaccurate (Vafa, Rambachan, and Mullainathan, 2024), and evaluators may distrust recommendations that conflict with their own assessment. If evaluators largely ignore the AI input, $G_{HAI} \approx G_H$ and the assistance pipeline produces similar noise to the status quo.

Testable Implications. This framework yields three predictions that we take to the data. First, if AI reduces grading noise, AI grades should correlate more strongly with the organization's independent assessment of candidates (the in-person assessment scores $\tilde{\theta}$). Second, for a given number of applicants advanced, the pipeline with lower grading noise should produce better screening outcomes, as more candidates who are above the organization's bar at the in-person assessment are correctly identified. A pipeline that assigns higher grades can also produce more offers by advancing more applicants, so the two channels have to be separated empirically. Third, if evaluators do not incorporate AI recommendations, the Human-with-AI-Assistance pipeline should resemble the Human-Only pipeline in both grade distributions and downstream outcomes.

4 Data, Outcomes, and Empirical Strategy

4.1 Data

Our experiment involved the evaluation of 697 eligible applications, corresponding to 4182 question answers. Within this set, 101 applications were independently graded by two distinct evaluators. Table A.2 presents baseline summary statistics of our sample (Panel A displays question-level summary statistics, and Panel B displays application-level summary statistics), and Table A.3 presents application-level baseline balance checks. The average length of an answer was 373 words (2238 words for the entire application). Among the applicants, 36% identify as female, 57% have completed their national service, 5% hold a Master's degree or higher, and 13% had previously applied to the program. 86% of the applicants come from five universities in Ghana (KNUST, University of Development Studies, University of Cape Coast, University of Education (Winneba), and University of Ghana). 39% of applicants originally come from one of Ghana's Northern regions, 14% from Volta region and the remainder from Ashanti (12%), Greater Accra (7%), and other regions in Southern and Central Ghana (27%).[17] Assignment of applications to policy treatment groups is largely balanced across observable characteristics. Columns 13 and 14 of Appendix Table A.3 report the joint F-statistic and the related p-value of a regression for each of the row variables on the set of three treatment indicators and strata fixed effects. We also fail to reject the null hypothesis of zero effect in a joint test of orthogonality of all variables in the table on assignment to any treatment status (p-val=0.52).

4.2 Outcome Variables

In this section, we describe our outcome variables in detail. First, we outline the question-level outcome variables used in our grading analysis. Next, we describe the application-level ("downstream") outcome variables used to analyze the effects of the three different policy pipelines.

[17] Due to a change in the non-profit organization's data privacy policy during the course of the experiment, we were able to obtain detailed background information for approximately 75% of applicants.

4.2.1 Question-Level Outcome Variables

Grades We have three types of grades in our data; "initial grades", "final grades" and "AI grades". "Initial grades" and "final grades" are grades recorded by human evaluators, while "AI grades" are grades provided by our algorithm. To analyze how evaluators respond to AI assistance, we use initial grades and final grades. As explained in Figure 2 and Section 2.2 above, initial grades are recorded after the evaluator has reviewed the answer for the first time, and final grades are recorded after the evaluator has seen the AI feedback page.[18]

We use human grades (initial and final) and AI grades to construct additional variables: a) *initial disagreement*: a dummy variable that equals one if the human initial grade is not equal to the AI grade ($G_H \neq G_{AI}$); b) *algorithmic override*: a dummy variable that equals one if the human final grade is not equal to the AI grade ($G_{HAI} \neq G_{AI}$) for the subset of applications given the AI assistance; c) *any revision*: a dummy variable that equals one if the evaluator revised their initial grade conditional on the initial and AI grades being different ($G_{HAI} \neq G_H \mid G_H \neq G_{AI}$) for the subset of applications for which AI assistance was given.

Grading Time We record the time evaluators spend grading application answers through our survey platform. We use these time variables to analyse how AI assistance affects grading time. The first measure, *time up to initial grade*, captures the time taken to assign the initial grade and allows us to examine potential anticipation effects of AI assistance, since evaluators were informed in advance whether they would receive such assistance. The second measure, *time up to final grade*, captures the time taken until the final grade is assigned and thus reflects the overall effect of AI assistance on grading time.[19]

4.2.2 Downstream Outcomes (Application-Level Outcome Variables)

The total application grade, which is used to determine which candidates advance to the next phase of the selection process, is calculated by summing the individual essay grades with equal weight given to each.[20] Based on this total grade, we construct *above-the-bar*, a dummy variable that equals one if the applicant achieved a total grade of 18 or above, and was invited to the in-person assessment center. For applicants who are "above-the-bar", we observe additional downstream outcomes, and create the following variables; a) *attended assessment center*: a dummy variable

[18] For applications assessed by humans without AI assistance, the initial grade is equal to the final grade by construction, since evaluators do not have the opportunity to revise their assessment. However, for applications for which AI-assistance was given, the initial grade might differ from the final grade, depending on whether the evaluators adjusted their grade.

[19] For question answers assessed without AI assistance, *time up to final grade* is equivalent to *time up to initial grade*. However, for question answers assessed with AI assistance, *time up to final grade* is the sum of *time up to initial grade* and the time spent on the AI feedback page.

[20] For applications assigned to the *Human-Only* pipeline, the sum of initial grades is used. For applications in the *Human-with-AI-Assistance* pipeline, the sum of final grades is used. Finally, for applications in the *AI-Only* pipeline, the sum of AI grades is used.

that equals one if the applicant attended the in-person assessment day; b) *assessment center grade*: the total grade the applicant got during that in-person assessment day, c) *offer*: a dummy variable that equals one if the applicant received a fellowship offer (i.e. achieved at least 50 out of 100 average grade in the in-person assessment day), and d) *accepted offer*: a dummy variable that equals one if the applicant accepted the offer, that is, was hired.

4.3 Main Empirical Specification

To estimate the effects of the three policy pipelines on downstream (application-level) outcomes, we estimate the following equation:

$$y_i = \alpha + \beta_1 \text{AI_Only}_i + \beta_2 \text{Human_with_AI_Assistance}_i + X_i'\lambda + \gamma_i + \epsilon_i \tag{2}$$

where AI_Only_i and $\text{Human_with_AI_Assistance}_i$ are indicator variables equal to one if the application was assigned to the *AI-Only* and *Human-with-AI-Assistance* pipeline, respectively, and γ_i is the stratification variable (randomization round). X_i' is a vector of control variables including evaluator fixed effects, the length of the application, the applicant's graduation year and an indicator variable for whether the applicant completed their national service.[21] We cluster the standard errors at the application level.

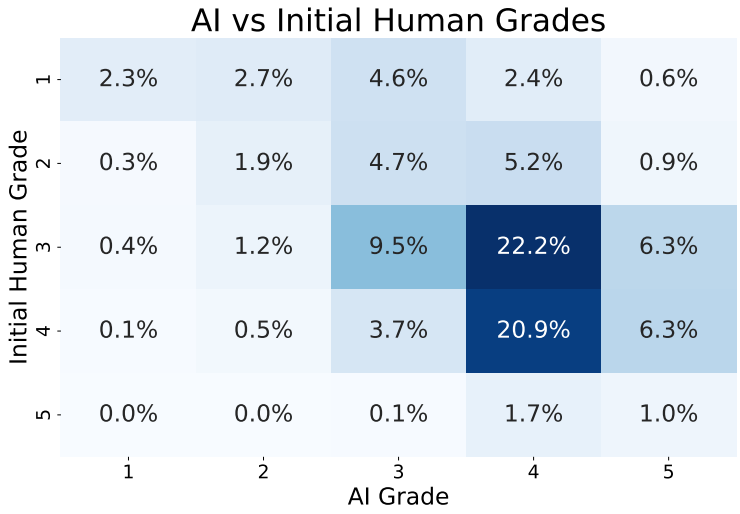
5 Main Results

In this section we present the effects of incorporating AI into the organization's recruitment process. We first document how AI and human grades compare and then present the downstream effects of the three policy pipelines on hiring outcomes.

5.1 How AI and Human Grades Compare

[21] As mentioned above, we have additional demographic variables for about 75% of the sample, but in order not to lose observations, we only use the variables available for everybody as controls.

Figure 3: Initial human grades vs. AI grades



	Number
Agreement rate (initial human & AI grades)	36 %
Grade difference, human minus AI (overall)	-0.71
Grade difference, human minus AI (conditional on disagreement)	-1.11

Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement percentages between initial human and AI grades (both ranging from 1 to 5). The diagonal (top-left to bottom-right) indicates complete agreement. Areas below (above) the diagonal represent cases where the initial human grade is higher (lower) than the AI grade. The table summarizes agreement rates (row 1), difference between initial human and AI grades (row 2), and grade difference between initial human and AI grades conditional on there being a grade disagreement (row 3).

AI grades are systematically higher than human grades ($G_H < G_{AI}$). Figure 3 compares initial human and AI grades for each essay answer across the 5-point scale. The two agree in only 36% of cases and when they disagree, the AI grade is higher in 87% of the cases.[22] The average grade difference (initial grade- AI grade) is -0.71 points (-1.1 conditional on disagreement).

In the assistance pipeline, where evaluators see the AI grade before finalizing their own, agreement rises only modestly, to 47.7%, with an average grade difference of -0.61 (Appendix Figures A.5, A.9, and A.11). These remaining large differences in AI and human grades are due to evaluators largely ignoring the AI recommendations. When the initial human grade differs from the AI grade, evaluators override the recommendation 80.6% of the time (Appendix Figure A.10, Panel a, dashed line). The override is strongly asymmetric. Evaluators reject the AI recommendation in 86.8% of cases when it suggests a higher grade, but only 44.9% of the time when it suggests a

[22]There is some heterogeneity across questions (Appendix Figure A.2), with agreement rates ranging from 26% (question 4, Value Alignment) to 47% (question 6, Leadership).

lower grade.[23] Evaluators thus disproportionately reject the AI recommendation when it suggests a higher grade, which is consistent with a conservative grading bias.

5.2 Downstream Effects of the Three Policy Pipelines

Compared to applicants assigned to the *Human-Only* pipeline, applicants in the *AI-Only* pipeline are 18.4 p.p. (65%) more likely to be interviewed (attend the assessment center), 17.4 p.p. (84%) more likely to receive an offer, and 10.9 p.p. (73%) more likely to be hired (Table 1 Panel A, columns (2), (4) and (6)). Since AI grades are higher than human grades on average, the AI advances 29.5 p.p. (50%) more candidates past the application cutoff (Appendix Table A.6), and a large part of the increase in offer rates comes from this extensive margin.[24] A thing worth noting is that the in-person assessment uses a fixed threshold (50 out of 100) to determine offers, and the organization leaves positions unfilled rather than lower the bar. Advancing more candidates therefore only increases the offer rate if those marginal candidates (who were advanced by the AI, but would not have been advanced by a human) are genuinely above the quality threshold.

To test whether the increase in the offer rates is only due to the increased "pass-through" rates of these "marginal" candidates, we shut down the extensive margin channel completely and fix the number of candidates advanced in each pipeline. We show the results in Table 1, Panel B, where we rank candidates by their application grades within each pipeline and compare offer rates for the top 115, where the top-115 cutoff corresponds to the number of candidates advanced in the *Human-Only* pipeline. We further narrow the pool to the top 50 and top 30 candidates. Among the top 115, candidates in the *AI-Only* pipeline are 35% more likely to receive an offer (column (2)). The point estimates are larger for the top 50 and top 30 (64% and 102%, columns (4) and (6)), but they rest on smaller samples and are less precise.[25] Within each pipeline, the correlation between application grades and in-person assessment performance is 0.34 in the *AI-Only* pipeline and 0.07 in the *Human-Only* pipeline (Figure 4). These correlations are descriptive, since each is computed among the candidates that pipeline advanced, so they do not by themselves establish that the AI ranks candidates more informatively. We return to grading consistency in Section 6.

[23] The revision rates are actually slightly higher than the algorithmic override statistics would suggest: conditional on disagreement, evaluators revise their grade in only 28.5% of cases, but when they do they mostly adjust by approximately one grade point and not all the way to the AI grade (Appendix Figure A.10, Panels b and c). There is a negligible number of revisions where human and AI grades are in agreement, a total of 10 cases representing 0.51% of the sample where AI assistance was given.

[24] A related concern could be that the higher advancement rate floods the in-person assessment stage with more candidates than the organization can handle. However, the organization specifically told us that they mainly care about not screening out capable candidates and that they do not consider a larger assessment center pool especially costly.

[25] The top-115, top-50, and top-30 candidates are also more likely to be hired: 16, 34, and 42%, respectively, but we cannot reject the null that these effects are zero. This is because among the best candidates of every pipeline, a large proportion actually rejects the offer, presumably because of better outside options. See Appendix table A.7

Table 1: Downstream Outcomes for Policy Pipelines

Panel A: Downstream Outcomes

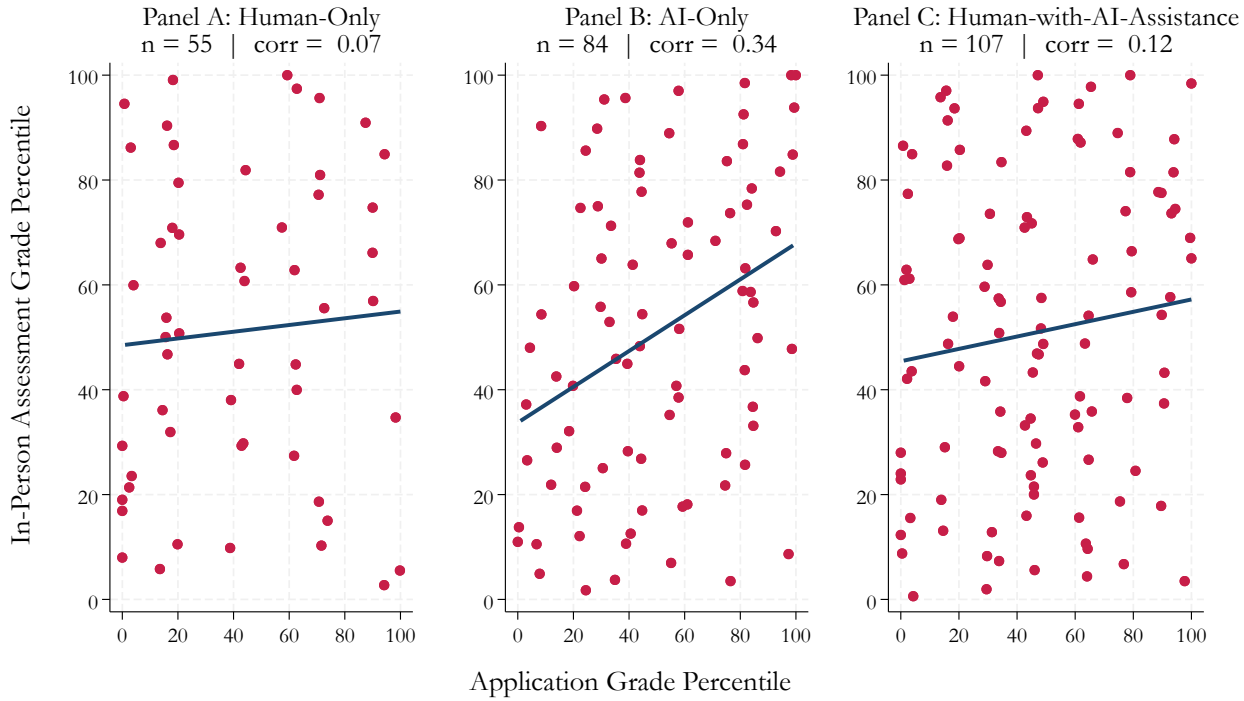
	Interviewed		Offer		Hired	
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Only	0.196***	0.184***	0.183***	0.174***	0.113***	0.109**
	(0.050)	(0.050)	(0.047)	(0.047)	(0.042)	(0.043)
Human-with-AI-Assistance	0.044	0.050	0.046	0.049	0.015	0.019
	(0.042)	(0.040)	(0.038)	(0.038)	(0.033)	(0.033)
Mean (Human-Only)	0.284	0.284	0.206	0.206	0.149	0.149
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes
N	697	697	697	697	697	697
p-values						
$\beta_AI{=}\beta_AIAssistance$	0.001	0.003	0.002	0.004	0.013	0.024

Panel B: Offer Rates, Holding the Number of Applicants Advanced Fixed

	Top-115		Top-50		Top-30	
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Only	0.112*	0.122*	0.174**	0.204**	0.254**	0.319***
	(0.063)	(0.064)	(0.083)	(0.089)	(0.114)	(0.119)
Human-with-AI-Assistance	0.067	0.060	0.082	0.111	0.114	0.224*
	(0.064)	(0.068)	(0.079)	(0.086)	(0.110)	(0.122)
Mean (Human-Only)	0.348	0.348	0.319	0.319	0.312	0.312
Sample	Top-115	Top-115	Top-50	Top-50	Top-30	Top-30
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes
N	365	365	221	221	125	125
p-values						
$\beta_AI{=}\beta_AIAssistance$	0.469	0.343	0.258	0.251	0.186	0.382

Notes: Panel A reports estimated coefficients from OLS regressions of indicator variables for whether the applicant was interviewed, i.e., attended the assessment center (Columns (1) and (2)), received a job offer (Columns (3) and (4)), and was hired, i.e., accepted the job offer (Columns (5) and (6)). All outcomes are unconditional: a value of zero is assigned if the applicant did not reach that stage. Panel B reports estimated coefficients from OLS regressions of an indicator variable for whether the applicant received an offer, for the top-n candidates based on application scores within each pipeline: top-115 (Columns (1) and (2)), top-50 (Columns (3) and (4)), and top-30 (Columns (5) and (6)). The top-115 cutoff corresponds to the number of candidates advanced in the *Human-Only* pipeline. Because some candidates share the same application grade, the number of candidates in each bin may exceed n. In both panels, all columns include stratum (week) fixed effects. Even columns additionally include controls for evaluator fixed effects, the length of the application, the applicant's graduation year, and an indicator variable for whether the applicant completed their national service. Standard errors are clustered at the application level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 4: Correlations Between Application Grades and In-Person Assessment Grades, by Pipeline



Notes: The figure shows scatter plots of pipeline-specific in-person assessment grades (percentiles) versus application grades for the three pipelines: Human-Only, AI-Only, and Human-with-AI-Assistance. Each subplot includes a fitted line indicating the within-pipeline correlation between application grades and in-person assessment grades.

By contrast, Table 1 Panel A shows that applicants in the *Human-with-AI-Assistance* pipeline do not have a statistically significantly higher likelihood of receiving an offer or being hired compared to applicants in the *Human-Only* baseline, and the correlation of the grades in this pipeline (0.12) is similar to the correlation in the *Human-Only* pipeline (0.07).[26] In summary, the *Human-with-AI-Assistance* pipeline therefore fails to close the gap to the *AI-Only* pipeline even though the evaluators are given access to the same algorithm. AI assistance also increases grading time: essays take 13–17% longer to grade, driven by disagreement cases where grading time increases by approximately 20% (Appendix Table A.8).

[26] Applicants assigned to the *Human-with-AI-Assistance* pipeline receive on average 0.736 points (4.2%) higher total grades than those in the *Human-Only* baseline (Appendix Table A.6). However, this increase is not large enough to statistically significantly affect the advancement rate to the next stage.

6 Mechanisms

The results in Section 5 raise two questions. First, beyond assigning higher grades, how does grading by the AI differ from grading by human evaluators? Second, why is the algorithmic override so severe in the *Human-with-AI-Assistance* pipeline? We address the first question in Subsection 6.1, where we compare the consistency of human and AI grades, and the second in Subsection 6.2, where we document the role of LLM-generated essays.

6.1 Signal and Noise in the Age of AI

The framework in Section 3 yields a directly testable prediction: if idiosyncratic variation in human grading (ε_{ij}) attenuates the correlation between application grades and candidate quality, and if the AI applies the rubric more consistently, then AI grades should be more informative. We examine this prediction in three steps.

First, we measure the variance of grades assigned to semantically similar essays. For each essay, we identify its k nearest neighbors using cosine similarity of vector embeddings and compute the variance of their grades (see Appendix Figure A.12 for a visual representation of these embeddings). Lower variance indicates that similar essays receive similar grades. Table 2 shows that human initial grades have a mean neighborhood variance of 0.667 at $k = 20$. AI scores are substantially more consistent, with a mean variance of 0.448, a 33% reduction relative to human initial grades.[27] Conditional on essay content, the AI assigns considerably more uniform grades than human evaluators and no individual human grader reaches the AI's level of consistency.[28]

Second, direct grade agreement rates corroborate this pattern. Among a random subset of applications graded independently by two different human evaluators, the two scores agreed in only 44% of cases, meaning most essays receive different grades depending on which evaluator reads them (Appendix Figure A.3).[29] By contrast, when GPT-4 re-grades the same essays with different random seeds, it disagrees with itself in only 21.5% of cases, indicating that the model applies the rubric in a more consistent manner.[30]

Third, the framework predicts that more consistent grades should correlate more strongly with independent assessments of candidate quality. As already mentioned above in Section 5, applica-

[27] The difference between AI and human initial grade variance is highly significant ($p < 0.001$, paired t-test on essay-level neighborhood variances). The same pattern holds for $k = 50$ and $k = 100$. Final human scores, available only for the Human-with-AI-Assistance arm, have a mean variance of 0.628, between the human initial and AI values. For comparison, the pre-experimental 2021/22 cohort, where only human grades are available, has a mean neighborhood variance of 0.890 (Appendix Table A.4), indicating that grading noise was even higher before the experiment.

[28] Within-grader neighborhood variance ranges from 0.522 to 0.823 at $k = 20$, with a roughly 60% gap between the most and least consistent grader.

[29] There is some heterogeneity across questions, with agreement rates ranging from 36% (question 1) to 53% (question 6) (Appendix Figure A.4).

[30] Within-model disagreement rates vary across models: Claude 3.5 Sonnet (5.7%), GPT-4o (15.3%), GPT-4 (21.5%), and Gemini 1.5 Pro (27.0%).

Table 2: Variance in Grades for Similar Essays

		Mean neighborhood grade variance		
Sample	N	$k = 20$	$k = 50$	$k = 100$
22/23 cohort				
Initial score	4182	0.667	0.698	0.719
Final score	1968	0.628	0.660	0.713
AI score	4182	0.448	0.464	0.476
p-value (AI − Initial)		0.000	0.000	0.000
p-value (AI − Final)		0.000	0.000	0.000
p-value (Initial − Final)		0.000	0.000	0.248
21/22 cohort				
Human score	2976	0.890	0.942	0.966

Notes: Each cell reports the mean neighborhood grade variance. For each essay, we compute cosine similarity to all other essays using Voyage AI embeddings (voyage-large-2-instruct model), identify the k nearest neighbors (excluding the essay itself), and compute the variance of their grades. Lower values indicate that similar essays receive similar grades (less noisy grading). p-value rows report two-sided paired t-tests on essay-level neighborhood variances. We use all available grades (parallel grading) for this analysis.

tion grades in the *AI-Only* pipeline correlate more strongly with in-person assessment performance (0.34) than grades in the *Human-Only* (0.07) or *Human-with-AI-Assistance* (0.12) pipelines (Figure 4). However, these correlations should be interpreted with caution. We can only compute them for candidates who attended the in-person assessment, which is why the samples are small. Moreover, the candidates who were advanced have a narrower range of application grades than the full applicant pool, which lowers the correlation, especially in the *Human-Only* pipeline where fewer candidates were advanced. Thus, we see this as suggestive evidence that the more consistent AI grades are also more closely related to performance in the in-person assessment.

The *Human-with-AI-Assistance* pipeline fails to reproduce these gains in consistency. The mean neighborhood variance of human grades with assistance is similar to the variance of initial human grades (0.628 vs. 0.667, a 6% decrease), compared to the 33% reduction achieved by full automation. Because the evaluators largely ignore the AI recommendations, the grading noise in the assistance pipeline remains close to that of the initial human grades.

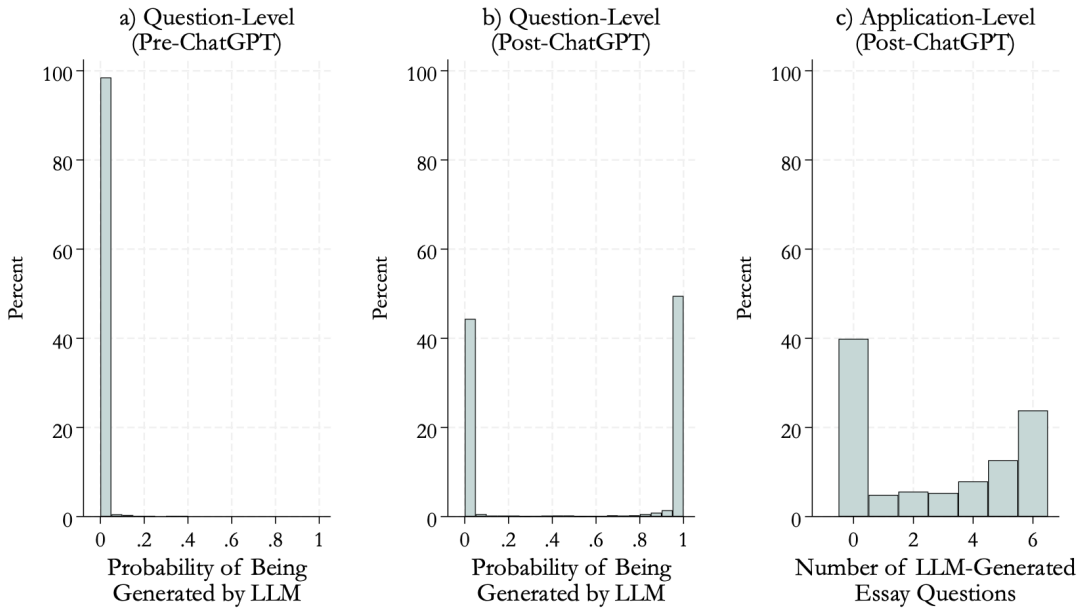
6.2 Investigating Potential Reasons for High Algorithmic Override in the *Human-with-AI-Assistance* Pipeline

As documented in Sections 5 and 6.1 above, the *Human-with-AI-Assistance* pipeline fails to close the gap between full automation and human only screening, because the evaluators mostly do

not follow AI-recommendations. We now examine a factor that is associated with even higher algorithmic override: the prevalence of LLM-generated application essays.

6.2.1 LLM-Generated Essays in the Applicant Pool

Figure 5: How common are AI-generated essays?



Notes: The figure displays the usage of LLMs in generating essay answers submitted with applications. Panel (a) shows the probability that each individual essay answer was generated by an LLM for applicants from the cohort that applied before ChatGPT became commercially available (Spring 2022). Panel (b) shows the probability that each individual essay answer was generated by an LLM for the cohort that applied after ChatGPT's release (Spring 2023). Panel (c) presents the distribution of the number of LLM-generated answers per application for the cohort that applied after ChatGPT's release.

After the experiment was completed, some evaluators mentioned that they noticed many essays appeared to be ChatGPT-generated and felt the AI grader did not take this into account. To assess the prevalence of LLM-generated content, we use Pangram Text, a transformer-based neural network with an overall accuracy of 99.85% and a 0.19% false positive rate (Emi and Spero, 2024). We classify an essay as LLM-generated if the estimated probability exceeds 0.99, and validate this threshold using pre-ChatGPT essays from the 2022 application cycle, where it produces a 0% false positive rate. LLM-generated essays are very common in our experimental setting. We classify approximately 45% of essay answers as LLM-generated (Figure 5).[31] Additionally, 60% of

[31] This percentage is based on a 0.99 cutoff; for a 0.9 cutoff, the corresponding percentage is 50%. The share varies by question, ranging from 39% to 48% (Appendix Figure A.6).

applications have at least one LLM-generated essay, and about 32% of applications are classified as fully LLM-generated. LLM-generated essays are longer, less likely to include specific information, more semantically compact, and occupy distinct regions of the semantic space compared to non-LLM essays (see Appendix Figures A.13, A.7 and A.8 for details).

6.2.2 How Does LLM Usage Affect Grading and Algorithmic Override?

Both human graders and the AI award higher grades to LLM-generated essays, but the premium differs: humans discount LLM essays relative to the algorithm, with the gap between human and AI grades approximately 25% larger for LLM essays (Appendix Table A.9). This discount grows over time: by the final third of graded applications, humans assign grades that are 35% lower for LLM essays compared to those assigned by AI. Humans are also about 4.5 percentage points (12%) less likely to agree with the AI grade for LLM essays than for non-LLM essays.[32]

Table 3: Human Graders Override the Algorithm More When Grading LLM-Written Essays: Sample of AI-Assisted Screening

	Algorithmic Override				Any Revision		Final Grade -AI grade	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
LLM-essay	0.079***	0.069***	0.091***	0.075***	-0.084***	-0.062**	-0.184***	-0.147***
	(0.023)	(0.024)	(0.022)	(0.022)	(0.026)	(0.025)	(0.040)	(0.042)
Mean (non-LLM)	0.497	0.497	0.772	0.772	0.314	0.314	-0.557	-0.557
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	1,968	1,968	1,265	1,265	1,265	1,265	1,968	1,968

Notes: Columns 1-8 report estimated coefficients from OLS regressions respectively of algorithmic override (final human grade differs from AI grade) overall (Columns (1) and (2)) and when there is initial grade disagreement (Columns (3) and (4)), any grade revision after seeing the AI grade when initial human and AI grades differ (Columns (5) and (6)) and the difference between final human and AI grades (Columns (7) and (8)). All columns include controls for evaluator fixed effect, the even columns additionally include controls for the week application was submitted, length of the application, the applicant's graduation year and an indicator variable for whether the applicant completed their national service. We use the sample of AI-Assisted screening in this analysis.

This relative discounting also occurs in the assistance pipeline. Table 3 shows that when grading LLM essays with algorithmic assistance, evaluators override the algorithm 16% more often (Column (1)), are 26.7% less likely to revise their grade (Column (5)), and produce final grades that diverge from the AI grade by 33.2% more (Column (7)).[33] As with unassisted grading, the differential override for LLM essays emerges over time rather than immediately (Appendix Table A.13). Our experimental design also allows us to examine whether providing a justification for the AI grade mitigates this pattern. When evaluators receive a rationale alongside the AI grade, they follow

[32]The results are robust to using different cutoffs for classifying essays as LLM-generated (see Appendix Table A.10).

[33]The results are very similar when we use alternative cutoffs for classifying the essays as being LLM-generated. See Appendix Tables A.11 and A.12.

the recommendation more frequently, but only for non-LLM essays (Appendix Figure A.14).[34] We speculate that this occurs because the rationale makes it clear the algorithm does not consider whether an essay was LLM-generated, so once evaluators recognize an essay is AI-written, they disregard the explanation altogether.

6.2.3 Signal and Noise in LLM-Generated Essays

The semantic compactness of LLM-generated essays documented above has a direct implication for grading consistency: because LLM essays are more similar to one another in content and structure, they should be easier to grade consistently. Table 4 confirms this. For both human and AI grades, neighborhood grade variance is substantially lower among LLM essays than among non-LLM essays. Human initial scores have a mean neighborhood variance of 0.765 for non-LLM essays, compared to 0.552 for LLM essays, a 28% reduction. AI scores show the same pattern (0.499 vs. 0.381), though at a lower overall level. Importantly, the AI's consistency advantage over human grading holds within both essay types: for non-LLM essays, the AI reduces neighborhood variance by 35% relative to human initial scores; for LLM essays, the reduction is 31%.

Table 4: Variance in Grades for Essays by LLM Status

	Non-LLM essays				LLM essays			
Sample	N	$k = 20$	$k = 50$	$k = 100$	N	$k = 20$	$k = 50$	$k = 100$
Initial score	2303	0.765	0.804	0.840	1879	0.552	0.583	0.633
Final score	1096	0.765	0.825	0.892	872	0.491	0.557	0.669
AI score	2303	0.499	0.523	0.544	1879	0.381	0.394	0.413
p-value (AI − Initial)		0.000	0.000	0.000		0.000	0.000	0.000
p-value (AI − Final)		0.000	0.000	0.000		0.000	0.000	0.000
p-value (Initial − Final)		0.901	0.001	0.000		0.000	0.103	0.000

Notes: Each cell reports the mean neighborhood grade variance. For each essay, we compute cosine similarity to all other essays within the same LLM-tag group using Voyage AI embeddings (voyage-large-2-instruct model), identify the k nearest neighbors (excluding the essay itself), and compute the variance of their grades. Lower values indicate that similar essays receive similar grades (less noisy grading). p-value rows report two-sided paired t-tests on essay-level neighborhood variances. We use all available grades (parallel grading) for this analysis.

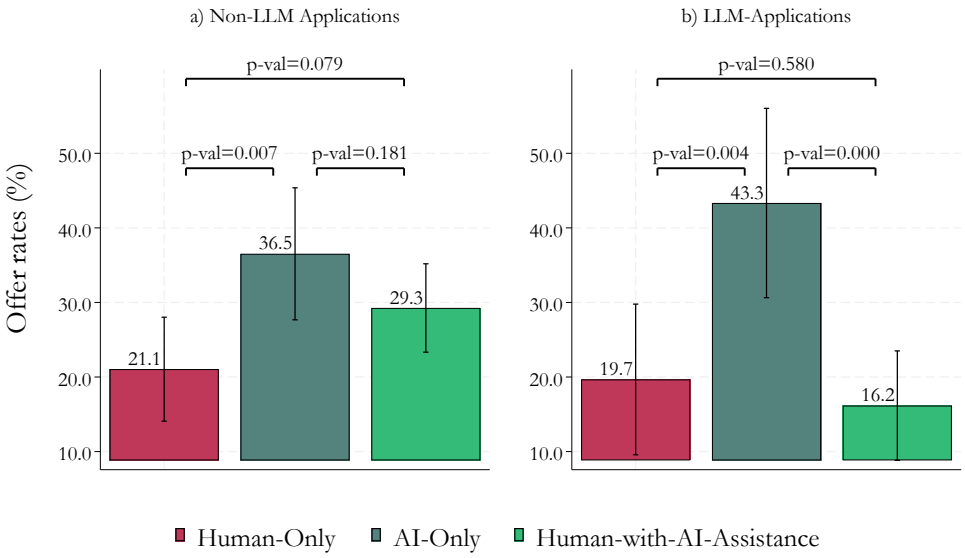
These results reveal two distinct forces. First, LLM-generated essays are inherently easier to grade consistently, likely because they are more formulaic and exhibit less variation in style and structure. Second, the AI maintains a large consistency advantage over human graders regardless of whether the essay is LLM-generated. Neighborhood grade variance is lowest for AI grades of LLM essays and highest for initial human grades of non-LLM essays.

[34] The results are similar when we use the two alternative classification cutoffs, 90% and 95% (see Appendix Figures A.15 and A.16).

6.2.4 Downstream Outcomes for LLM and Non-LLM Applications

In this analysis, we classify an application as an LLM application if it is fully LLM-generated, and as a non-LLM application otherwise. Non-LLM applications thus also include applications with some LLM-generated essays. If evaluators' discounting of LLM essays reflected superior information about candidate quality, we would expect the *Human-with-AI-Assistance* pipeline to produce better outcomes for LLM applications than the *AI-Only* pipeline. However, Figure 6 shows that the opposite occurs. Offer rates in the *Human-Only* pipeline are nearly identical for LLM and non-LLM applications (19.7 vs. 21.1%). However, in the *Human-with-AI-Assistance* pipeline, applicants with non-LLM applications receive offers at significantly higher rates, 13 percentage points (80%) higher than those with LLM applications. For non-LLM applications, we cannot reject that the *AI-Only* and *Human-with-AI-Assistance* coefficients are equal; for LLM applications, we can (Appendix Table A.14).[35] Thus, we do not find evidence that the evaluators' discounting of LLM essays reflects superior information about candidate quality: for LLM applications, the *Human-with-AI-Assistance* pipeline does not produce better outcomes than the *AI-Only* pipeline.

Figure 6: Offer Rates by Pipeline and LLM-Application



Notes: The figure shows regression coefficients from equation 2 without control variables run separately for a subsample of non-LLM- and LLM-applications, where the outcome variable is a binary indicator of whether a candidate received a job offer. The cranberry bar represents mean offer rates in the Human-Only pipeline (i.e. the constant term), and the emerald and mint bars represent the sum of the mean offer rates and the respective beta coefficients. Error bars indicate 95% confidence intervals based on standard errors of the relevant coefficients or linear combinations of the constant and the relevant coefficients. P-values come from t-tests evaluating whether the coefficients are statistically different from zero, and from testing for equality of the beta coefficients between AI-Only and Human-with-AI-Assistance pipelines.

[35]The results are qualitatively the same when using alternative cutoffs to classify applications as LLM-generated (see Appendix Tables A.16 and A.15 and Appendix Figures A.17 and A.18).

7 Conclusion

As generative AI technologies become widely adopted, an open question is whether such systems should carry out hiring tasks on their own or with humans in the loop. To inform this question, we embedded GPT-4 into the hiring process of an organization recruiting teachers in Ghana. Fully automating the screening process increases offer rates by 84% and hiring rates by 73% relative to the human-only baseline. We find that a large part of this gain comes from the extensive margin: the AI assigns higher grades and thus advances more marginal candidates to the in-person assessment. We also find that, compared to human evaluators, the AI grades more consistently: AI grades exhibit 33% lower variance among similar essays, and no individual human evaluator reaches the AI's level of consistency. Using GPT-4 as an assistant provides no significant improvement. Evaluators in the assistance pipeline largely ignore the AI recommendations, exhibiting a conservative grading bias that is compounded by distrust of the algorithm.

Several caveats are worth noting. Our setting is specific: a single nonprofit in Ghana, an essay-based screening task, and an early generative AI model (GPT-4) used without fine-tuning, at a time when both evaluators and applicants were encountering the technology for the first time. We also lack downstream measures of on-the-job performance such as student achievement or teacher value-added. That said, the in-person assessment that determines final hiring decisions was conducted by independent evaluators blind to treatment assignment and provides a credible benchmark for candidate quality. More broadly, the finding that algorithmic evaluation reduces grading noise by applying the same criteria to every candidate may not be specific to our context. Human grading noise is well-documented across educational assessment, peer review, and hiring (Kahneman, Sibony, and Sunstein, 2021), and the consistency advantage of algorithmic evaluation may generalize to other settings where human evaluation of written materials is noisy.

These results should not be viewed as static. Labor market conditions, generative AI capabilities, and perceptions of these technologies are all evolving. Once generative AI reaches widespread adoption, entire labor markets could shift (Galdin and Silbert, 2025), prompting changes in both applicant strategies and employer practices. Importantly, the algorithm we study was not designed to be an assistant: evaluators received a grade recommendation from a black-box model, for half of the applications together with a short rationale, and had no mechanism for dialogue. Designing AI systems that are genuinely complementary to human decision-makers, rather than simply presenting a take-it-or-leave-it recommendation (McLaughlin and Spiess, 2024), remains an important challenge for both research and practice.

References

Acemoglu, Daron, David Autor, and Simon Johnson. 2023. "Policy Insight 123: Can we Have Pro-Worker AI? Choosing a path of machines in service of minds." https://cepr.org/publications/policy-insight-123-can-we-have-pro-worker-ai-choosing-path-machines-service-minds.

Acemoglu, Daron, and Pascual Restrepo. 2019. "Automation and New Tasks: How Technology Displaces and Reinstates Labor." *Journal of Economic Perspectives* 33 (2): 3–30. 10.1257/jep.33.2.3.

Agan, Amanda Y., Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. 2023. "Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias." Working Paper 30981, National Bureau of Economic Research. 10.3386/w30981.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2024. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." https://papers.ssrn.com/abstract=4505053.

Angelova, Victoria, Will Dobbie, and Crystal Yang. 2026. "Algorithmic Recommendations and Human Discretion." *The Review of Economic Studies* 93 (4): 2250–2283. 10.1093/restud/rdaf084.

Avery, Mallory, Edwin Ip, Andreas Leibbrandt, and Joseph Vecci. 2026. "A Brave New World of Hiring: A Natural Field Experiment on How Asynchronous Interviews and AI Assessment Reshape Recruitment." 10.2139/ssrn.6180838.

Avery, Mallory, Andreas Leibbrandt, and Joseph Vecci. 2024. "Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech." Working Paper 10996, CESifo, https://www.ifo.de/DocDL/cesifo1_wp10996.pdf.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2025. "Generative AI at Work." *The Quarterly Journal of Economics* 140 (2): 889–942. 10.1093/qje/qjae044.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan et al. 2023. "Sparks of Artificial General Intelligence: Early experiments with GPT-4." 10.48550/arXiv.2303.12712, arXiv:2303.12712 [cs].

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review* 106 (5): 124–127. 10.1257/aer.p20161029.

Cowgill, Bo. 2020. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening." Working paper.

De Simone, Martin, Federico Tiberti, Maria Barron Rodriguez, Federico Manolio, Wuraola Mosuro, and Eliot Jolomi Dikoru. 2025. "From Chalkboards to Chatbots: Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria." Policy Research Working Paper 11125, World Bank. 10.1596/1813-9450-11125.

Dell'Acqua, Fabrizio, Edward McFowland, III, Ethan Mollick et al. 2026. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality." *Organization Science*. 10.1287/orsc.2025.21838, Published online March 11, 2026.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm aversion: People erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General* 144 (1): 114–126. 10.1037/xge0000033.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. "GPTs are GPTs: Labor market impact potential of LLMs." *Science* 384 (6702): 1306–1308. 10.1126/science.adj0998.

Emi, Bradley, and Max Spero. 2024. "Technical Report on the Pangram AI-Generated Text Classifier." 10.48550/arXiv.2402.14873, arXiv:2402.14873 [cs].

Flesch, Rudolph. 1948. "A new readability yardstick." *Journal of Applied Psychology* 32 (3): 221–233. 10.1037/h0057532.

Gabaix, Xavier. 2019. "Chapter 4 - Behavioral inattention." In *Handbook of Behavioral Economics: Applications and Foundations 1*, edited by Bernheim, B. Douglas, Stefano DellaVigna, and David Laibson Volume 2. of Handbook of Behavioral Economics - Foundations and Applications 2 261–343, North-Holland, . 10.1016/bs.hesbe.2018.11.001.

Gabaix, Xavier, and Thomas Graeber. 2024. "The Complexity of Economic Decisions." 10.3386/w33109.

Galdin, Anaïs, and Jesse Silbert. 2025. "Making Talk Cheap: Generative AI and Labor Market Signaling." 10.48550/arXiv.2511.08785.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li. 2018. "Discretion in Hiring." *The Quarterly Journal of Economics* 133 (2): 765–800. 10.1093/qje/qjx042.

Jabarian, Brian, and Luca Henkel. 2026. "Voice AI in Firms: A Natural Field Experiment on Automated Job Interviews." 10.2139/ssrn.5395709.

Jabarian, Brian, and Alex Imas. 2025. "Artificial Writing and Automated Detection." 10.2139/ssrn.5407424.

Kahneman, Daniel, Olivier Sibony, and Cass R Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Little, Brown Spark.

Kim, Hyunjin, Edward L. Glaeser, Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2024. "Decision authority and the returns to algorithms." *Strategic Management Journal* 45 (4): 619–648. 10.1002/smj.3569.

Kumar, Harsh, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2025. "Math Education with Large Language Models: Peril or Promise?" In *Artificial Intelligence in Education: AIED 2025*, Volume 15880. of Lecture Notes in Computer Science 60–75, Springer, . 10.1007/978-3-031-98459-4_5.

Li, Danielle, Lindsey Raymond, and Peter Bergman. 2026. "Hiring as Exploration." *The Review of Economic Studies* 93 (2): 1200–1240. 10.1093/restud/rdaf040.

McLaughlin, Bryce, and Jann Spiess. 2024. "Designing Algorithmic Recommendations to Achieve Human-AI Complementarity." 10.48550/arXiv.2405.01484, arXiv:2405.01484 [cs].

Noy, Shakked, and Whitney Zhang. 2023. "Experimental evidence on the productivity effects of generative artificial intelligence." *Science* 381 (6654): 187–192. 10.1126/science.adh2586.

Otis, Nicholas, Rowan Clarke, Solène Delecourt, David Holtz, and Rembrand Koning. 2026. "The Uneven Impact of Generative Artificial Intelligence on Entrepreneurial Performance: Evidence from a Field Experiment in Kenya." *Management Science*. 10.1287/mnsc.2024.06909, Published online July 10, 2026.

Parshakov, Petr, Iuliia Naidenova, Sofia Paklina, Nikita Matkin, and Cornel Nesseler. 2025. "Users Favor LLM-Generated Content – Until They Know It's AI." 10.48550/ARXIV.2503.16458.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot." 10.48550/arXiv.2302.06590, arXiv:2302.06590 [cs].

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. 2024. "When combinations of humans and AI are useful: A systematic review and meta-analysis." *Nature Human Behaviour* 8 (12): 2293–2303. 10.1038/s41562-024-02024-1.

Vafa, Keyon, Ashesh Rambachan, and Sendhil Mullainathan. 2024. "Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function." In

Proceedings of the 41st International Conference on Machine Learning, Volume 235. of Proceedings of Machine Learning Research 48919–48937, PMLR, , https://proceedings.mlr.press/v235/vafa24a.html.

Wei, Jason, Xuezhi Wang, Dale Schuurmans et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." In *Advances in Neural Information Processing Systems*, Volume 35. 24824–24837, https://papers.nips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Wiles, Emma, Zanele Munyikwa, and John Horton. 2025. "Algorithmic Writing Assistance on Jobseekers' Resumes Increases Hires." *Management Science* 71 (12): 10144–10164. 10.1287/mnsc.2024.04528.

Online Appendix

Table of Contents

A Figures and Tables **2**

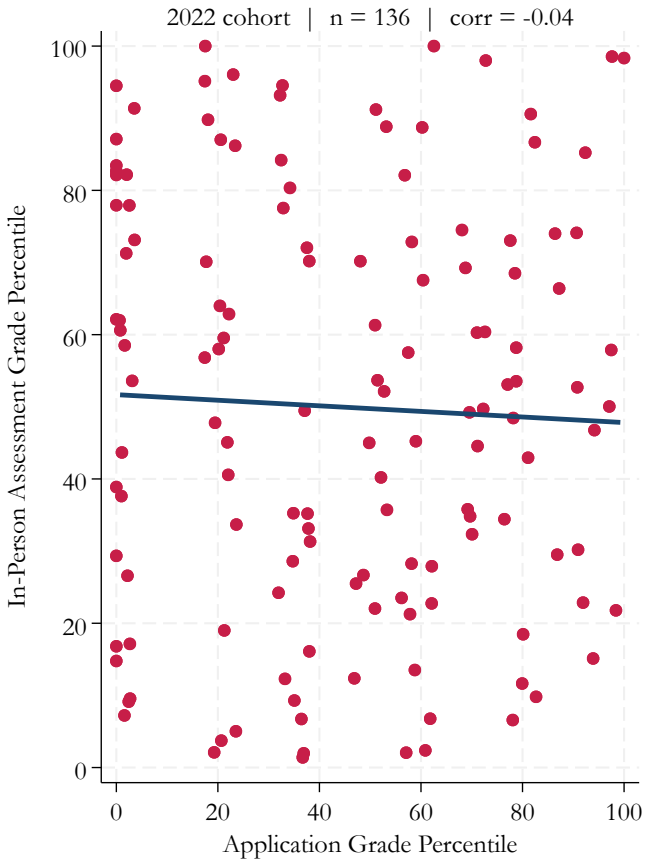
B Technical Appendix **32**

B.1 System Prompt 32

B.2 Content Prompts 32

A Figures and Tables

Figure A.1: Correlation Between Application and In-Person Assessment Grades (pre-experimental cohort)



Notes: Figure shows a scatter plot of in-person assessment grades (percentiles) versus application grades (percentiles). The solid line is a linear fit.

Table A.1: Questions and Grading Rubric for Fellowship Application

1. Why do you want to be a [*name of the NGO*] Fellow?
1. Does not give a reason for wanting to be an [*name of the NGO*] Fellow. 2. Gives a reason that is not linked to the [*name of the NGO*] vision or approach. 3. Gives a reason that is clearly linked to solving educational inequity in Ghana. 4. Can articulate elements of the Fellowship that they are most interested in for their own development. 5. Gives rationale for own desire to be a fellow and is able to talk about how past OR future activities connect to the [*name of the NGO*] vision.
2. What is an excellent education to you, and how do you intend to provide that to your students?
1. Does not define what an excellent education is and does not articulate how to provide that to their students. 2. Defines what an excellent education is but does not articulate how to provide that to their students. 3. Clearly defines what an excellent education is and shows a pathway to providing that to their students. 4. Rubric 3 plus: articulates factors that lead to academic achievement, mindset development, exposure to resources. 5. Rubric 4 plus: gives specific examples of actions they will take as a fellow and alumni to provide an excellent education to their students.
3. As a [*name of the NGO*] alumni, how do you envision yourself contributing to the [*name of the NGO*] alumni vision?
1. Does not demonstrate an understanding of the [*name of the NGO*] alumni vision. 2. Understands the [*name of the NGO*] alumni vision but does not articulate their role in achieving it. 3. Understands the [*name of the NGO*] alumni vision and can articulate their role in achieving the vision. 4. Rubric 3 plus: gives more than one example of how they're going to achieve the alumni vision. 5. Rubric 4 plus: mentions a specific sector/ job they have in mind and how they intend to leverage their position to achieve the [*name of the NGO*] alumni vision.
4. How do our core beliefs resonate with you?
1. Does not make reference to any of our core beliefs. 2. Makes some reference to our core beliefs but does not articulate how they resonate with them. 3. Makes reference to our core beliefs and articulates how they resonate with them. 4. Rubric 3 plus: shares an example of how at least one of our beliefs resonates with them. 5. Rubric 4 plus: shares an example of how all three core beliefs resonate with them.
5. Please describe a moment(s) when you overcame a challenge in order to achieve a non-academic goal.
1. Does not describe a challenge. 2. Describes a challenge(s) but does not share how they overcame the challenge(s). 3. Clearly defines a robust challenge and shares how they overcame the challenge. 4. Rubric 3 plus: shares more than one robust challenge and how they overcame them. 5. Rubric 4 plus: articulates what they would have done differently.
6. Please share with us two (2) instances when you were in a position of influence and motivated others (a team or group of people) to make a desired change and achieved a desired outcome.
1. Does not describe a clear position of influence and the people they motivated. 2. Describes some position of influence but does not articulate how they motivated others to take a desired action. 3. Clearly describes two robust positions of influence and shares examples of how they motivated others to take desired actions. 4. Rubric 3 plus: articulates the outcomes of the actions. 5. Rubric 4 plus: shares an exceptional position of influence (a position that affects a large group of people i.e more than 100 people) and clear

Notes: The Table presents an overview of the questions and the corresponding grading criteria. Questions 1-4 are meant to be proxies for how good the applicant's fit is to work for the organization, question 5 is meant to proxy "grit", and question 6 is meant to measure the applicant's ability to lead and influence others.

Table A.2: Summary Statistics

Panel A: Grading (Question-Level)

	All			Human Grading			Human Grading with AI Assistance		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Human initial grade	4,182	2.988	1.034	2,214	2.956	1.032	1,968	3.024	1.035
Human final grade	4,182	3.019	1.031	2,214	2.956	1.032	1,968	3.089	1.026
AI grade	4,182	3.701	0.912	2,214	3.698	0.899	1,968	3.704	0.927
Time to initial grade	4,182	165	183	2,214	170.2	186.5	1,968	159.9	179.1
Time to final grade	4,182	181	220	2,214	170.2	186.5	1,968	192.5	252.8
Initial agreement	4,182	0.357	0.479	2,214	0.357	0.479	1,968	0.357	0.479
Algorithmic Override	1,968	0.523	0.500	N/A	N/A	N/A	1,968	0.523	0.500
Revised grade	4,182	0.089	0.284	2,214	0.000	0.000	1,968	0.189	0.391
Human initial-AI grade	4,182	-0.713	1.011	2,214	-0.742	0.976	1,968	-0.680	1.049
Human final-AI grade	1,968	-0.615	0.889	N/A	N/A	N/A	1,968	-0.615	0.889
LLM-essay	4,182	0.449	0.497	2,214	0.455	0.498	1,968	0.443	0.497

Panel B: Policy Experiment (Application-Level)

	All			Human-Only			AI-Only			Human-with-AI-Assistance		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Total grade	697	19.228	4.295	194	17.691	4.096	175	22.234	3.380	328	18.534	4.070
Above-the-bar	697	0.709	0.455	194	0.593	0.493	175	0.926	0.263	328	0.662	0.474
Attend interviews	697	0.354	0.479	194	0.284	0.452	175	0.480	0.501	328	0.329	0.471
Offer received	697	0.274	0.446	194	0.206	0.406	175	0.389	0.489	328	0.253	0.435
Offer accepted	697	0.185	0.389	194	0.149	0.357	175	0.263	0.441	328	0.165	0.371
LLM-application	697	0.316	0.465	194	0.314	0.465	175	0.343	0.476	328	0.302	0.460
Number of LLM-essays	697	2.696	2.547	194	2.629	2.518	175	2.840	2.604	328	2.659	2.538

Notes: The Table displays summary statistics for the overall experimental sample. Panel A displays question-level summary statistics from our grading "experiment", and Panel B displays application-level summary statistics from our policy experiment. The outcome variables are defined in Section 4.2.

Table A.3: Balance

	All			Human Only			AI-only			AI-assistance			Joint	
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	F-stat	p-val
Application														
Length (words)	697	2,238	248	194	2,236	234	175	2,228	251	328	2,244	256	0.217	0.805
Demographics														
Female	515	0.357	0.484	145	0.359	0.481	131	0.374	0.486	239	0.347	0.486	0.154	0.858
National Service	697	0.572	0.495	194	0.582	0.494	175	0.577	0.495	328	0.564	0.497	0.067	0.935
University														
KNUST	515	0.177	0.382	145	0.207	0.406	131	0.153	0.361	239	0.172	0.378	0.677	0.508
UDS	515	0.198	0.399	145	0.207	0.406	131	0.191	0.394	239	0.197	0.398	0.080	0.923
UCC	515	0.169	0.375	145	0.159	0.367	131	0.122	0.329	239	0.201	0.401	1.940	0.145
UEW	515	0.167	0.373	145	0.152	0.360	131	0.206	0.406	239	0.155	0.362	0.939	0.392
UG	515	0.153	0.361	145	0.152	0.360	131	0.206	0.406	239	0.126	0.332	1.859	0.157
Other	515	0.136	0.343	145	0.124	0.331	131	0.122	0.329	239	0.151	0.358	0.454	0.636
Education														
Bachelor's	697	0.555	0.497	194	0.557	0.498	175	0.554	0.498	328	0.555	0.498	0.007	0.993
Final Year	697	0.397	0.490	194	0.392	0.489	175	0.400	0.491	328	0.399	0.491	0.019	0.981
Master's	697	0.047	0.213	194	0.052	0.222	175	0.046	0.209	328	0.046	0.209	0.050	0.951
Completion Year														
>2 years ago	697	0.204	0.403	194	0.201	0.402	175	0.194	0.397	328	0.210	0.408	0.116	0.890
<= 2 years	697	0.359	0.480	194	0.376	0.486	175	0.366	0.483	328	0.345	0.476	0.249	0.779
Yet to complete	697	0.438	0.496	194	0.423	0.495	175	0.440	0.498	328	0.445	0.498	0.110	0.896
GPA														
1.0-2.0	515	0.017	0.131	145	0.014	0.117	131	0.023	0.150	239	0.017	0.129	0.159	0.853
2.1-3.0	515	0.355	0.479	145	0.331	0.472	131	0.298	0.459	239	0.402	0.491	2.492	0.084
3.1-4.0	515	0.627	0.484	145	0.655	0.477	131	0.679	0.469	239	0.582	0.494	2.255	0.106
Current Region														
Ashanti	514	0.154	0.361	144	0.181	0.386	131	0.122	0.329	239	0.155	0.362	0.953	0.386
Greater Accra	514	0.331	0.471	144	0.312	0.465	131	0.321	0.469	239	0.347	0.477	0.316	0.729
Northern regions	514	0.300	0.459	144	0.299	0.459	131	0.305	0.462	239	0.297	0.458	0.010	0.990
Other South	514	0.177	0.382	144	0.160	0.368	131	0.206	0.406	239	0.172	0.378	0.617	0.540
Volta	514	0.039	0.194	144	0.049	0.216	131	0.046	0.210	239	0.029	0.169	0.647	0.524
Home Region														
Ashanti	514	0.123	0.328	144	0.111	0.315	131	0.153	0.361	239	0.113	0.317	0.657	0.519
Greater Accra	514	0.076	0.265	144	0.104	0.307	131	0.046	0.210	239	0.075	0.264	1.775	0.171
Northern regions	514	0.389	0.488	144	0.354	0.480	131	0.405	0.493	239	0.402	0.491	0.483	0.617
Other South	514	0.270	0.445	144	0.299	0.459	131	0.237	0.427	239	0.272	0.446	0.650	0.522
Volta	514	0.142	0.349	144	0.132	0.340	131	0.160	0.368	239	0.138	0.346	0.228	0.797
Mother tongue														
Twi	515	0.557	0.497	145	0.524	0.501	131	0.618	0.488	239	0.544	0.499	1.480	0.229
Ewe	515	0.070	0.255	145	0.097	0.296	131	0.053	0.226	239	0.063	0.243	0.995	0.370
Ga/Dangme	515	0.076	0.265	145	0.110	0.314	131	0.031	0.173	239	0.079	0.271	4.227	0.015
Northern lang.	515	0.297	0.457	145	0.269	0.445	131	0.298	0.459	239	0.314	0.465	0.449	0.638
Applied before	515	0.128	0.335	145	0.090	0.287	131	0.145	0.353	239	0.142	0.350	1.756	0.174

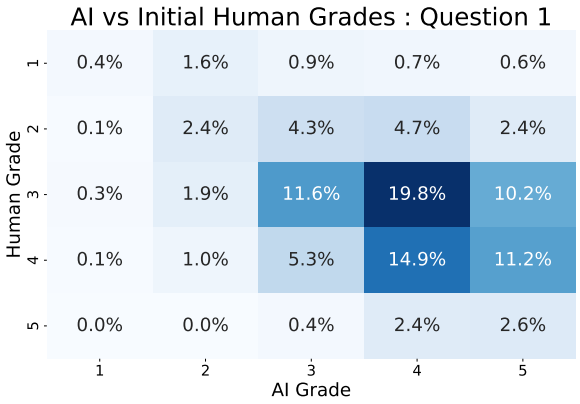
Notes: The figure shows the balance table for our policy experiment. Last two columns (under "Joint") report the F-statistic and the p-value from a joint test of significance of the set of treatment dummies in explaining each row variable in a regression with strata (week) fixed effects included and with standard errors clustered at the application level. Joint test of orthogonality of all variables in the table on any treatment group is from a multinomial logit: Chi-squared(26)=25, p-val=0.52.

Table A.4: k-NN Neighborhood Grade Variance by Treatment Arm, 2022/23 Cohort

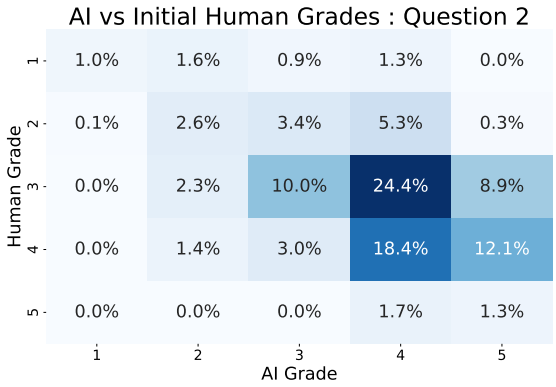
		Mean neighborhood grade variance		
Sample	N	$k = 20$	$k = 50$	$k = 100$
22/23 cohort				
Human Only	1164	0.706	0.792	0.870
Human + AI Assisted	1968	0.628	0.660	0.713
AI Only	1050	0.463	0.532	0.605
21/22 cohort				
Human Only	2976	0.890	0.942	0.966

Notes: Each cell reports the mean neighborhood grade variance. For each essay, we compute cosine similarity to all other essays *within the same arm* using answer-only Voyage AI embeddings (1,024 dimensions), identify the k nearest neighbors (excluding the essay itself), and compute the variance of their grades. Lower values indicate that similar essays receive similar grades (less noisy grading). The grade used for each arm is the arm-specific score: initial human score for Human Only, final (post-AI-assisted) human score for Human + AI Assisted, and AI score for AI Only. Significance tests are not reported because arms use different essay subsets and score columns, making paired comparisons across arms infeasible. The 21/22 cohort has human grades only; LLM-tag classification is not available.

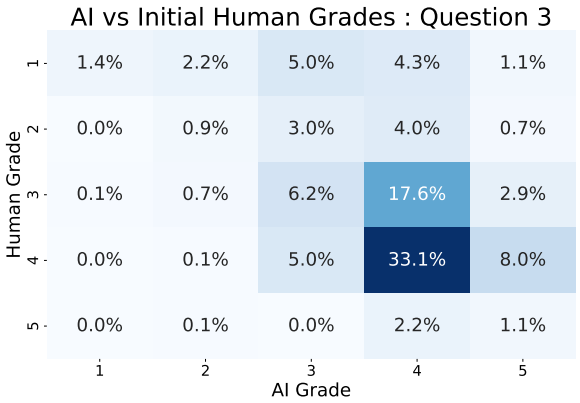
Figure A.2: Initial Human Grades vs. AI Grades by Question



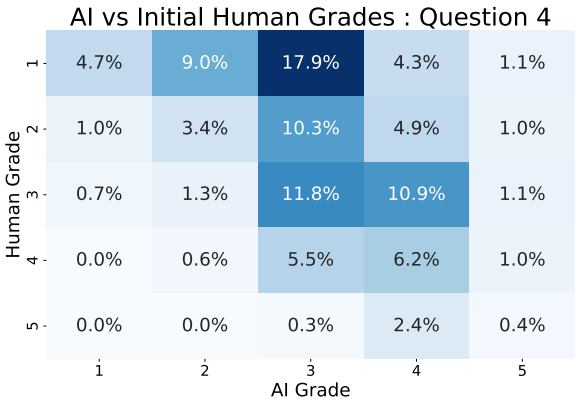
a) Question 1



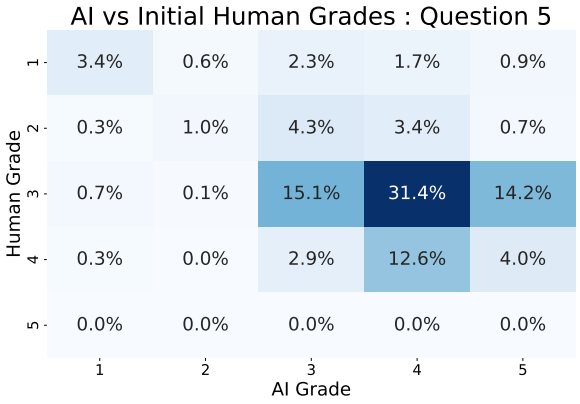
b) Question 2



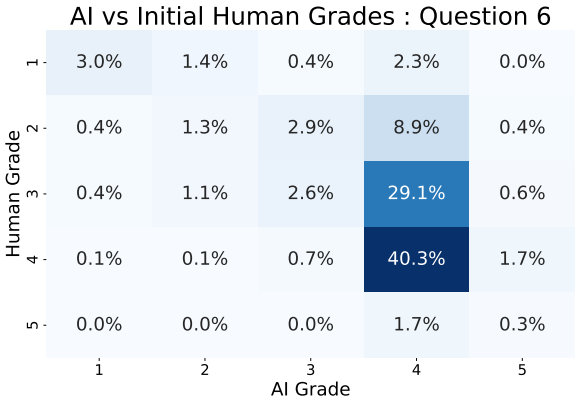
c) Question 3



d) Question 4



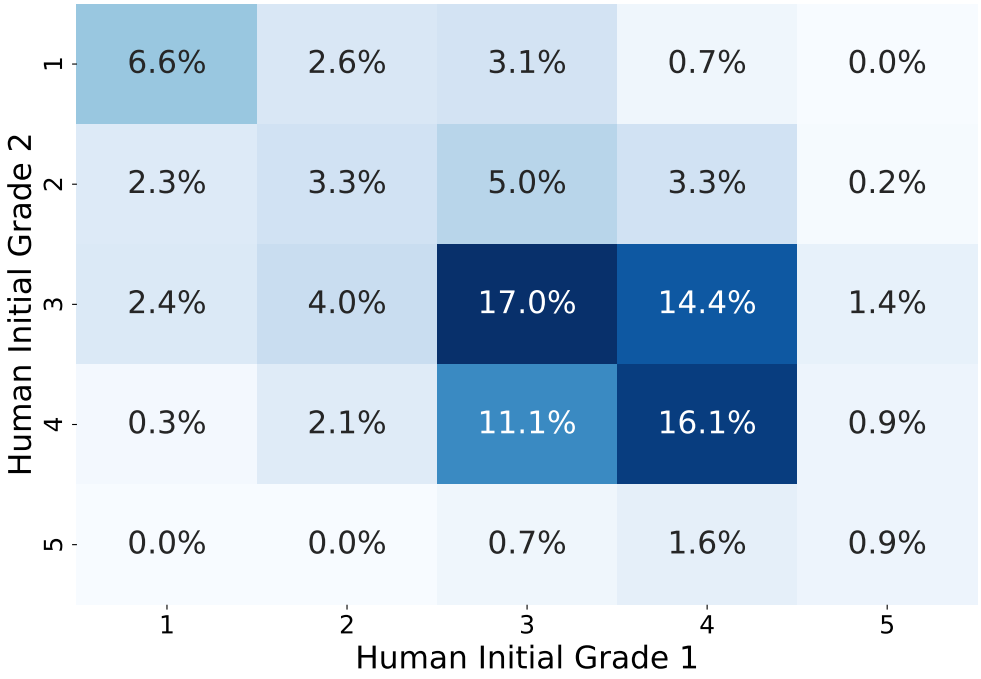
e) Question 5



f) Question 6

Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement frequencies between initial human and AI grades (both ranging from 1 to 5), separately for each question. The diagonal (top-left to bottom-right) indicates complete agreement. Areas below (above) the diagonal represent cases where the initial human grade is higher (lower) than the AI grade.

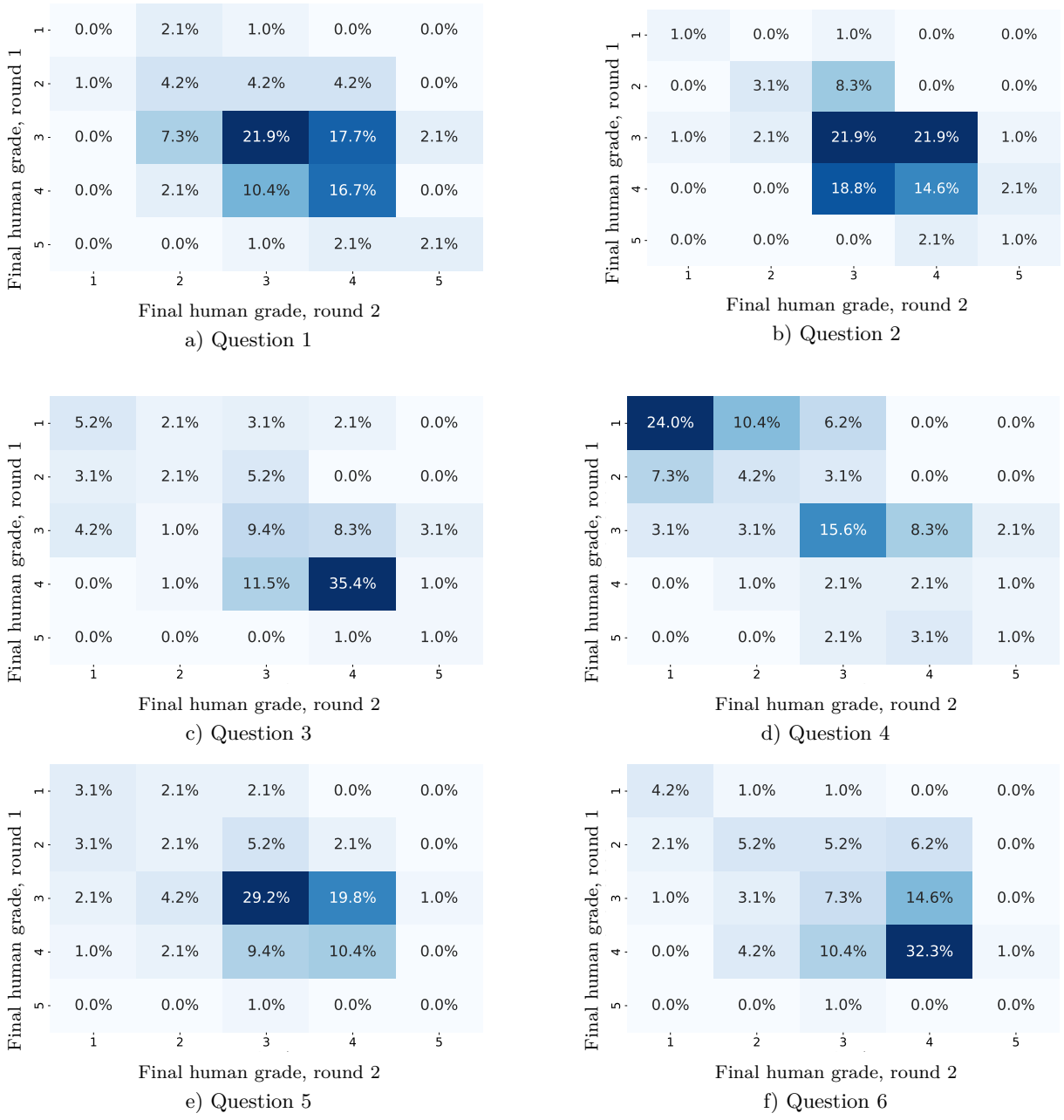
Figure A.3: Initial Human Grades Consistency



	Number	Percentage
Human grade 1!=Human grade 2	323	56
Human grade 1=Human grade 2	253	44

Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement frequencies between initial human grades (ranging from 1 to 5) for applications that were graded twice. The diagonal (top-left to bottom-right) indicates complete agreement. Areas above (below) the diagonal represent cases where the initial human grade in the first round was higher (lower) than the initial human grade in the second round. The table summarizes question counts off (row 1) and on (row 2) the diagonal.

Figure A.4: Final Human Grade Consistency by Question



a) Question 1 b) Question 2

c) Question 3 d) Question 4

e) Question 5 f) Question 6

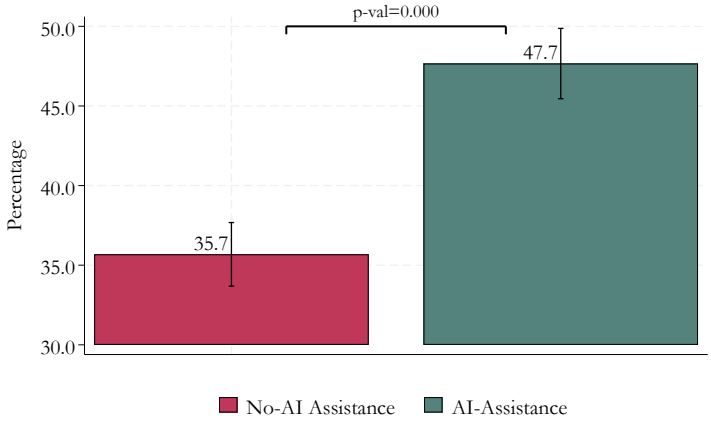
Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement frequencies between final human grades (ranging from 1 to 5) for applications graded twice, separately for each question. Rows show final human grades in the first grading round; columns show final human grades in the second grading round. The diagonal (top-left to bottom-right) indicates agreement in grades from the two grading rounds and areas off the diagonal indicate disagreement across the two grading rounds.

Table A.5: Downstream Outcomes (conditional)

	Interviewed	Offer	Hired
	(1)	(2)	(3)
AI-only	0.0377	0.0836	-0.0411
	(0.061)	(0.075)	(0.091)
Human-with-AI-Assistance	0.00881	0.0442	-0.0690
	(0.058)	(0.073)	(0.089)
Mean (Human-only)	0.478	0.727	0.725
Stratum FE	Yes	Yes	Yes
Controls	No	No	No
N	494	247	191
p-values			
$\beta_AI = \beta_AIAssistance$	0.578	0.508	0.721

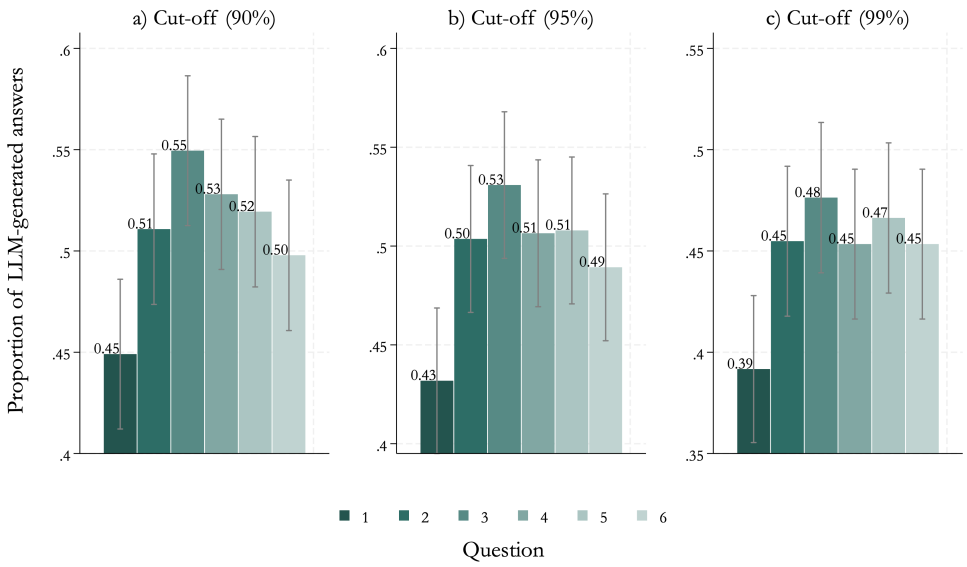
Notes: Columns 1-3 report estimated coefficients from OLS regressions respectively of an indicator variable for whether the applicant attended the assessment center, conditional on being advanced to the assessment center (column 1), received a job offer (column 2) and was hired, that is accepted the job offer (column 3). All columns include stratum (week) fixed effects. Note that the variables in columns 1-3 are conditional, meaning that they take a missing value if the person has not reached that stage. Standard errors are clustered at the application level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.5: Average Agreement in Final Grades



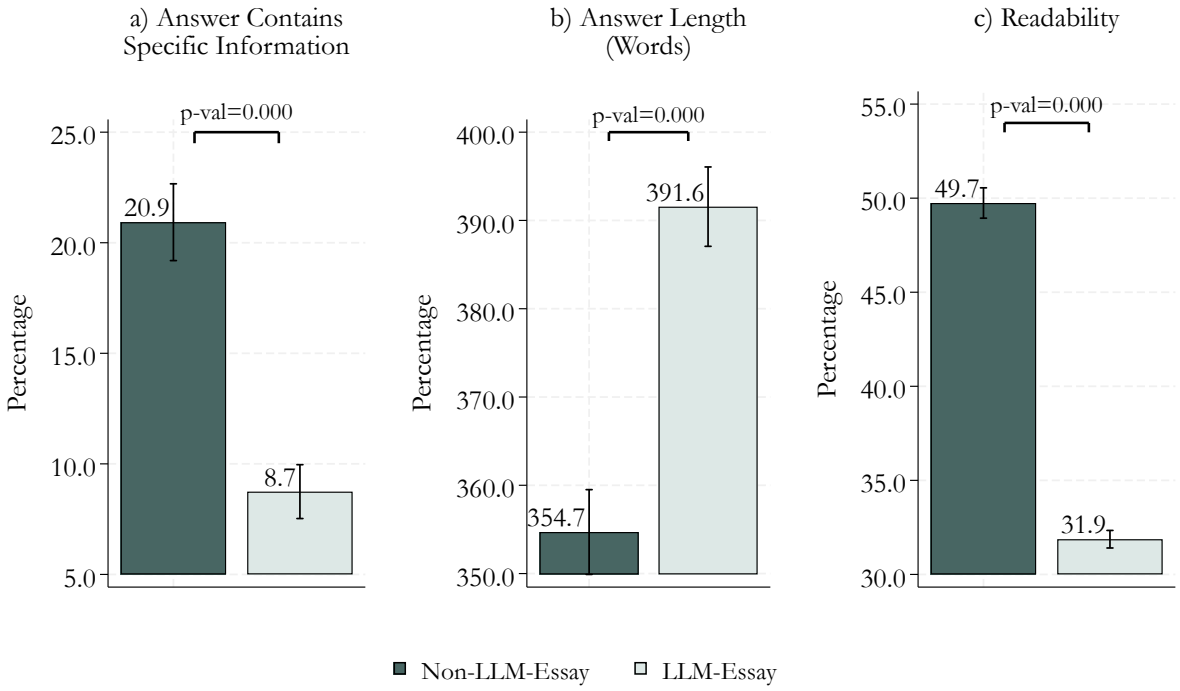
Notes: The figure shows the proportion of questions where the final human grade matched the AI grade, separately by whether the application was assigned AI-assistance. Error bars indicate the 95% confidence intervals. p-values are calculated from a t-test from a regression of a binary indicator for grade agreement on a dummy variable indicating whether the application was assigned to receive AI assistance.

Figure A.6: How common are LLM-Generated Essays (by question)?



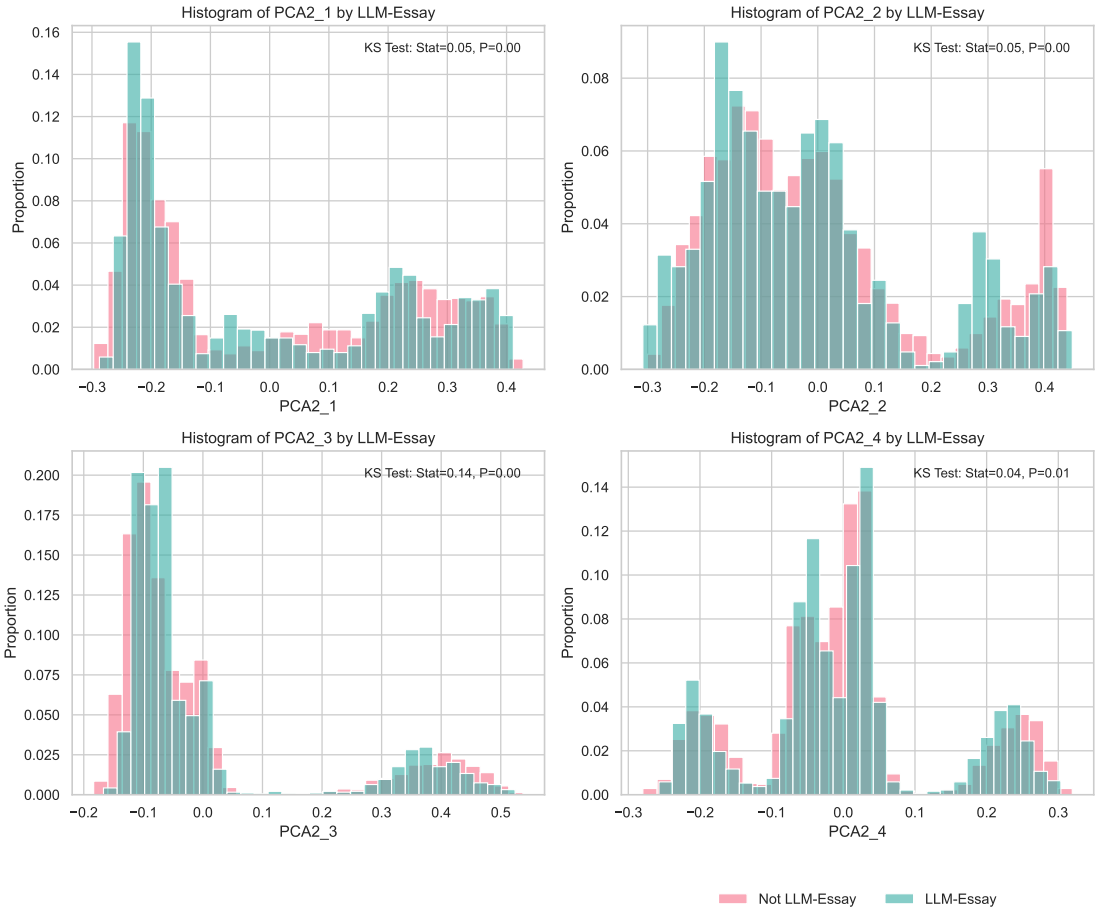
Notes: The figure shows the proportion of answers classified as LLM-generated for each question, using three different probability cut-offs based on the Pangram Text model's estimates.

Figure A.7: Characteristics of AI-generated answers



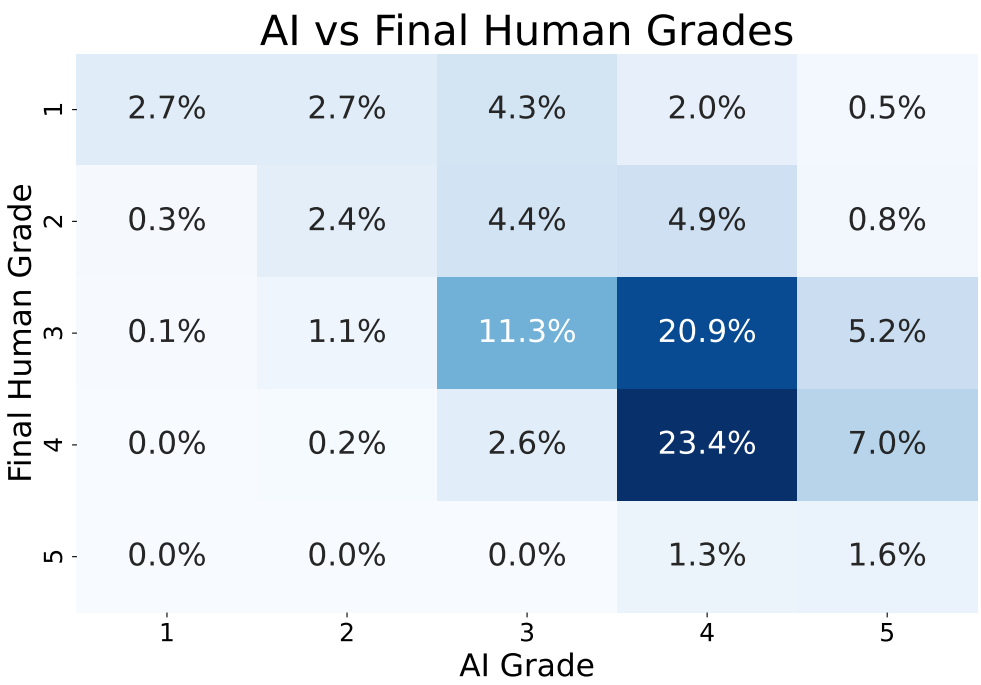
Notes: The figure depicts the characteristics of LLM- and non-LLM-essays. Panel a: Proportion of answers that contain specific information (for example on applicant's gender or university). Panel b: Answer length in words. Panel c: The complexity as measured by the Flesch reading ease (Flesch, 1948), a widely used metric that depends on sentence length and the number of syllables in words used in sentences. The exact formula is: Reading Ease $= 206.835 - 1.015\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) - 84.6\left(\frac{\text{Total Syllables}}{\text{Total Words}}\right)$. The Flesch reading ease score is a widely used metric for readability, and it is conveniently available in tools like Microsoft Word's editor. The readability measure scores usually range from 0 to 100, with higher scores indicating easier reading (for reference, "Time" averages around 50, while "the Harvard Law Review" sits at around 32). The original classifications are as follows: (0-30) Very difficult; (30-50) Difficult; (50-60) Fairly difficult; (60-70) Standard; (70-80) Fairly easy; (80-90) Easy; (90-100) Very easy.

Figure A.8: Is Semantic Content Different Across LLM and Non-LLM answers?



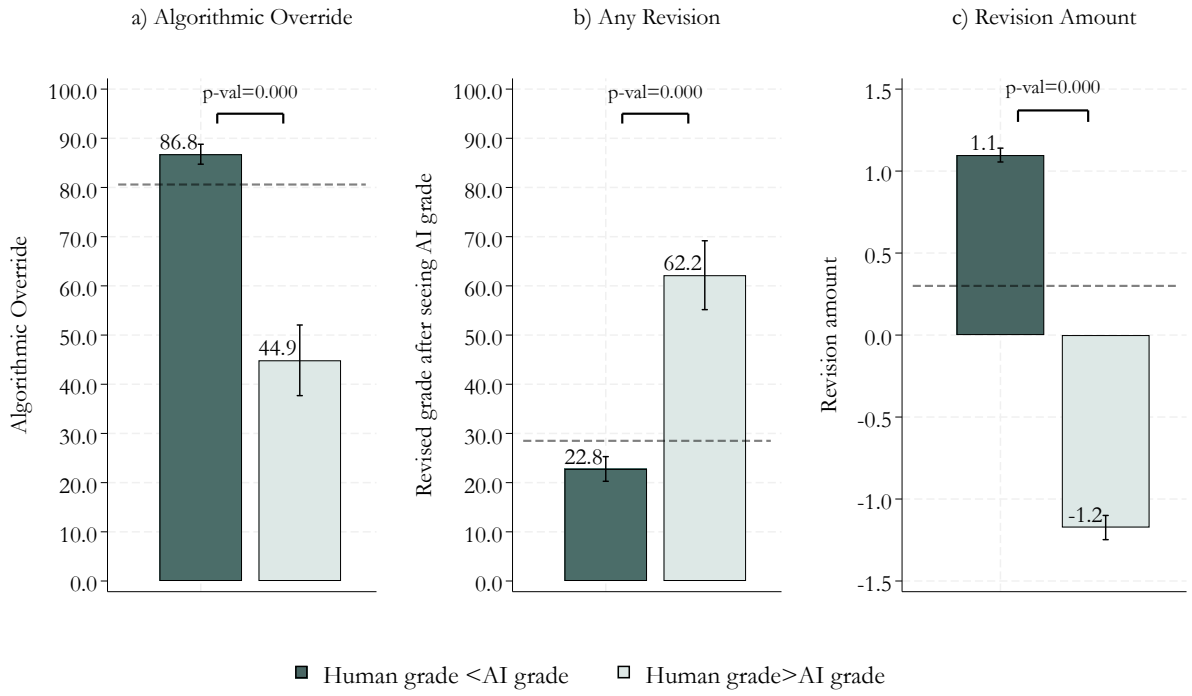
Notes: The figure depicts the distribution of the first four principal components (out of 10 that were generated) of the vector embeddings that were generated using "voyage-large-2-instruct" model from Voyage AI for LLM- and non-LLM-essays, and the test statistic and the p-value of the Kolmogorov-Smirnov test for equality of distributions.

Figure A.9: Final Human Grades vs. AI Grades



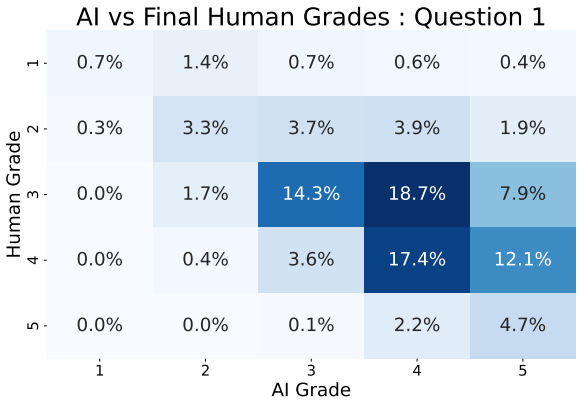
Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement frequencies between final human and AI grades (both ranging from 1 to 5). The diagonal (top-left to bottom-right) indicates complete agreement. Areas below (above) the diagonal represent cases where the final human grade is higher (lower) than the AI grade.

Figure A.10: Algorithmic Override and Grade Revisions by Initial Grade Disagreement

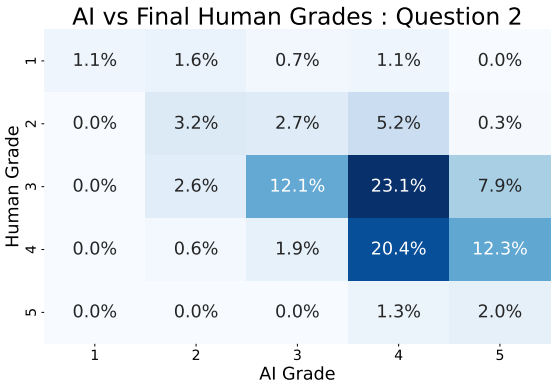


Notes: The figure shows the proportion of times the evaluators override the algorithm (Panel a), revise their initial grade (Panel b), and the average amounts they revise for (Panel c), categorized by initial human and AI grade disagreement. The dashed line represents the overall weighted average. Note that algorithmic override and revisions also occur when there is initial agreement between human and AI grades, but this happens in fewer than 1% of the cases and is thus omitted from the graph. Error bars indicate 95% confidence intervals around the means; *p-values* are calculated from a t-test, obtained from a regression of the outcome variable on an indicator variable denoting whether the human grade was higher than the AI grade, conditional on initial disagreement.

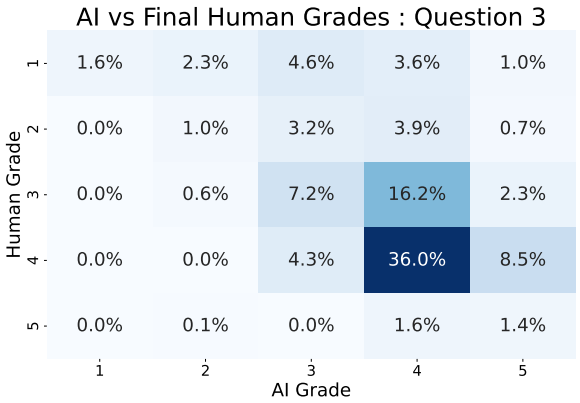
Figure A.11: Final Human Grades vs. AI Grades by Question



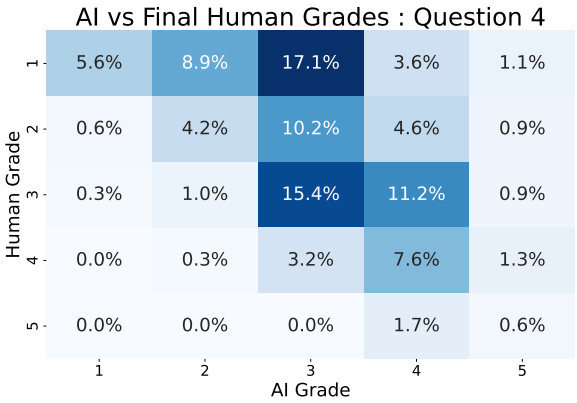
a) Question 1



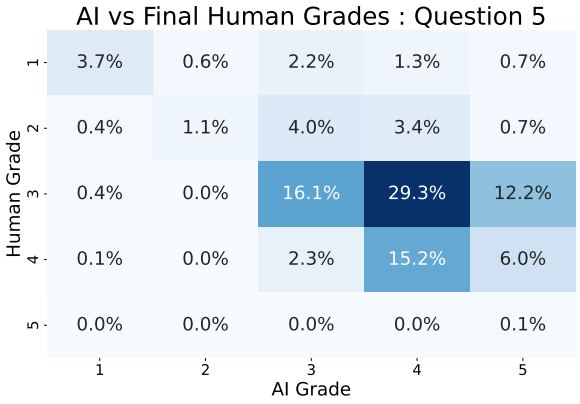
b) Question 2



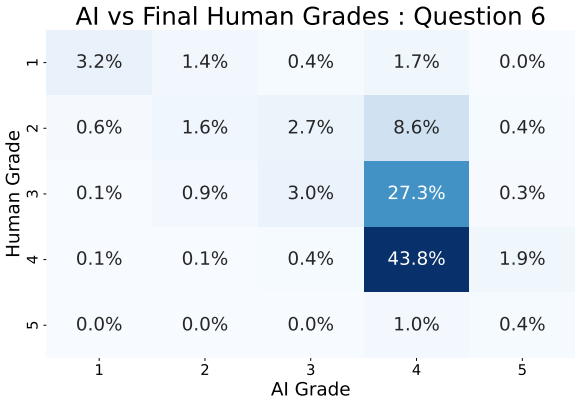
c) Question 3



d) Question 4



e) Question 5



f) Question 6

Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement frequencies between final human and AI grades (both ranging from 1 to 5), separately for each question. The diagonal (top-left to bottom-right) indicates complete agreement. Areas below (above) the diagonal represent cases where the final human grade is higher (lower) than the AI grade.

Table A.6: Grading Outcomes for Policy Pipelines

	Total Score		Above-the-bar	
	(1)	(2)	(3)	(4)
AI-Only	4.504***	4.213***	0.330***	0.295***
	(0.389)	(0.361)	(0.041)	(0.040)
Human-with-AI-Assistance	0.873**	0.736**	0.070	0.060
	(0.366)	(0.308)	(0.044)	(0.039)
Mean (Human-Only)	17.691	17.691	0.593	0.593
Stratum FE	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
N	697	697	697	697
p-values				
$\beta_AI{=}\beta_AIAssistance$	0.000	0.000	0.000	0.000

Notes: Table reports estimated coefficients from OLS regressions of total application score (Columns (1) and (2)) and an indicator variable for whether the applicant was advanced to the assessment center (Columns (3) and (4)). All columns include stratum (week) fixed effects. Even columns additionally include controls for evaluator fixed effects, the length of the application, the applicant's graduation year, and an indicator variable for whether the applicant completed their national service. Standard errors are clustered at the application level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: Grading Outcomes for Policy Pipelines: Hired

	Top-115		Top-50		Top-30	
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Only	0.040	0.040	0.084	0.104	0.106	0.142
	(0.057)	(0.059)	(0.077)	(0.082)	(0.109)	(0.117)
Human-with-AI-Assistance	0.014	0.004	0.059	0.071	0.073	0.150
	(0.057)	(0.062)	(0.073)	(0.082)	(0.104)	(0.121)
Mean (Human-Only)	0.252	0.252	0.246	0.246	0.250	0.250
Sample	Top-115	Top-115	Top-50	Top-50	Top-30	Top-30
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes
N	365	365	221	221	125	125
p-values						
$\beta_AI{=}\beta_AIAssistance$	0.651	0.539	0.741	0.670	0.751	0.934

Notes: Table reports estimated coefficients from OLS regressions of an indicator variable for whether the applicant accepted the offer (i.e. was hired), for the top-n candidates based on application scores within each pipeline: top-115 (Columns (1) and (2)), top-50 (Columns (3) and (4)), and top-30 (Columns (5) and (6)). The top-115 cutoff corresponds to the number of candidates advanced in the *Human-Only* pipeline. Because some candidates share the same application grade, the number of candidates in each bin may exceed n. All columns include stratum (week) fixed effects. Even columns additionally include controls for evaluator fixed effects, the length of the application, the applicant's graduation year, and an indicator variable for whether the applicant completed their national service. Standard errors are clustered at the application level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8: Time Spent on Application

	Time up to initial grade (log)			Time up to final grade (log)		
	(1)	(2)	(3)	(4)	(5)	(6)
AI assistance	-0.102*	-0.146***	-0.196***	0.167***	0.127***	0.024
	(0.060)	(0.045)	(0.063)	(0.055)	(0.041)	(0.058)
Disagreement in grade			0.085**			0.088**
			(0.042)			(0.042)
Disagreement x AI-Assistance			0.079			0.162***
			(0.064)			(0.060)
Mean (Human-Only) in seconds	170	170	170	170	170	170
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	No	Yes	Yes
N	4,182	4,182	4,182	4,182	4,182	4,182

Notes: Columns 1-6 report estimated coefficients from OLS regressions of log of time (in seconds) spent grading questions. Columns 1-3 represent time up to the initial grade, and columns 4-6 represent time up to the final grade. For the group without AI assistance, times to initial and final grades are equal. All columns include stratum (week) fixed effects; columns 2, 3, 5 and 6 additionally include controls for evaluator fixed effect, the length of the application, question number, the applicant's graduation year, and an indicator variable for whether the applicant completed their national service. Standard errors are clustered at the application level and reported in parentheses. Time is winsorized at 95th percentile on question-level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.9: Human Graders Discount LLM-Written Essays Relative to AI: All Applications

	Human grade - AI grade		Human grade= AI grade	
	(1)	(2)	(3)	(4)
LLM-essay	-0.184***	-0.168***	-0.030**	-0.045***
	(0.032)	(0.034)	(0.015)	(0.015)
Mean (non-LLM)	-0.665	-0.665	0.362	0.362
Controls	No	Yes	No	Yes
N	4,182	4,182	4,182	4,182

Notes: Columns 1-4 report estimated coefficients from OLS regressions respectively of difference between Human-Only grades and AI grades (Columns (1) and (2)) and an indicator variable for whether the Human-Only grade agreed with the AI grade (Columns (3) and (4)). All columns include controls for evaluator fixed effect, the even columns additionally include controls for the week application was submitted, length of the application, the applicant's graduation year and an indicator variable for whether the applicant completed their national service. We use the entire sample of grades in this analysis (what we call Human initial grade in Section 4).

Table A.10: Robustness Check: Human Graders Discount LLM-Written Essays Relative to AI (All Applications)

	Human grade - AI grade				Human grade= AI grade			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
LLM-essay	-0.175***	-0.158***	-0.171***	-0.153***	-0.031**	-0.045***	-0.025	-0.039**
	(0.033)	(0.035)	(0.033)	(0.035)	(0.016)	(0.016)	(0.016)	(0.016)
Mean (non-LLM)	-0.662	-0.662	-0.662	-0.662	0.364	0.364	0.361	0.361
Cutoff	95%	95%	90%	90%	95%	95%	90%	90%
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	4,182	4,182	4,182	4,182	4,182	4,182	4,182	4,182

Notes: The table presents robustness checks for two alternative probability cutoffs (90% and 95%) used to classify an essay as LLM-generated. Columns 1–8 report estimated coefficients from OLS regressions: Columns (1), (2), (3), and (4) show the difference between Human-Only grades and AI grades; Columns (5), (6), (7), and (8) report results for an indicator variable capturing whether the Human-Only grade agreed with the AI grade. Columns (1), (2), (5), and (6) use the 95% cutoff, while Columns (3), (4), (7), and (8) use the 90% cutoff. All regressions include evaluator fixed effects; the even-numbered columns additionally control for the week the application was submitted, the length of the application, the applicant's graduation year, and whether the applicant completed their national service. The analysis uses the full sample of grades (referred to as the Human initial grade in Section 4).

Table A.11: Robustness Check (95% Likelihood Cut-Off): Human Graders Override the Algorithm More When Grading LLM-Written Essays

	Algorithmic Override				Any Revision		Final Grade -AI grade	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
LLM-essay	0.061**	0.063***	0.062***	0.042*	-0.063**	-0.036	-0.163***	-0.153***
	(0.024)	(0.024)	(0.023)	(0.022)	(0.027)	(0.026)	(0.041)	(0.042)
Mean (non-LLM)	0.503	0.503	0.783	0.783	0.305	0.305	-0.561	-0.561
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	1,968	1,968	1,265	1,265	1,265	1,265	1,968	1,968

Notes: This table reports a robustness check for an alternative probability cut-off (95%) used to classify an essay as LLM-generated. Columns 1-8 report estimated coefficients from OLS regressions respectively of algorithmic override (final human grade differs from AI grade) overall (Columns (1) and (2)) and when there is initial grade disagreement (Columns (3) and (4)), any grade revision after seeing the AI grade when initial human and AI grades differ (Columns (5) and (6)) and the difference between final human and AI grades (Columns (7) and (8)). All columns include controls for evaluator fixed effect, the even columns additionally include controls for the week application was submitted, length of the application, the applicant's graduation year and an indicator variable for whether the applicant completed their national service. We use the sample of AI-Assisted screening in this analysis.

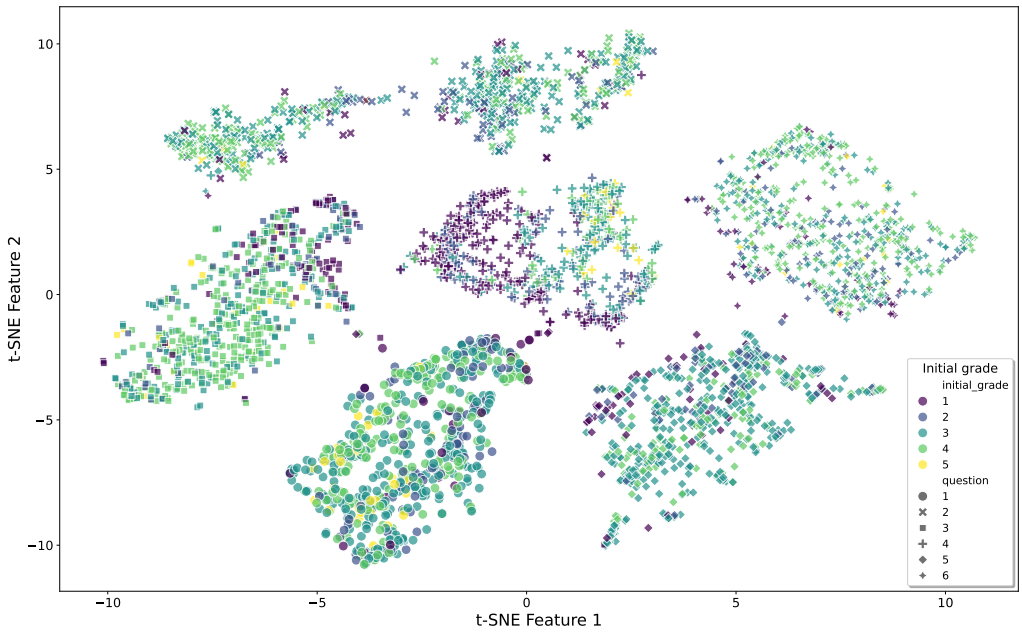
Table A.12: Robustness Check (90% Likelihood Cut-Off): Human Graders Override the Algorithm More When Grading LLM-Written Essays

	Algorithmic Override				Any Revision		Final Grade -AI grade	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
LLM-essay	0.060**	0.062**	0.061***	0.042*	-0.063**	-0.039	-0.163***	-0.153***
	(0.024)	(0.025)	(0.023)	(0.023)	(0.028)	(0.026)	(0.041)	(0.043)
Mean (non-LLM)	0.503	0.503	0.782	0.782	0.307	0.307	-0.559	-0.559
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	1,968	1,968	1,265	1,265	1,265	1,265	1,968	1,968

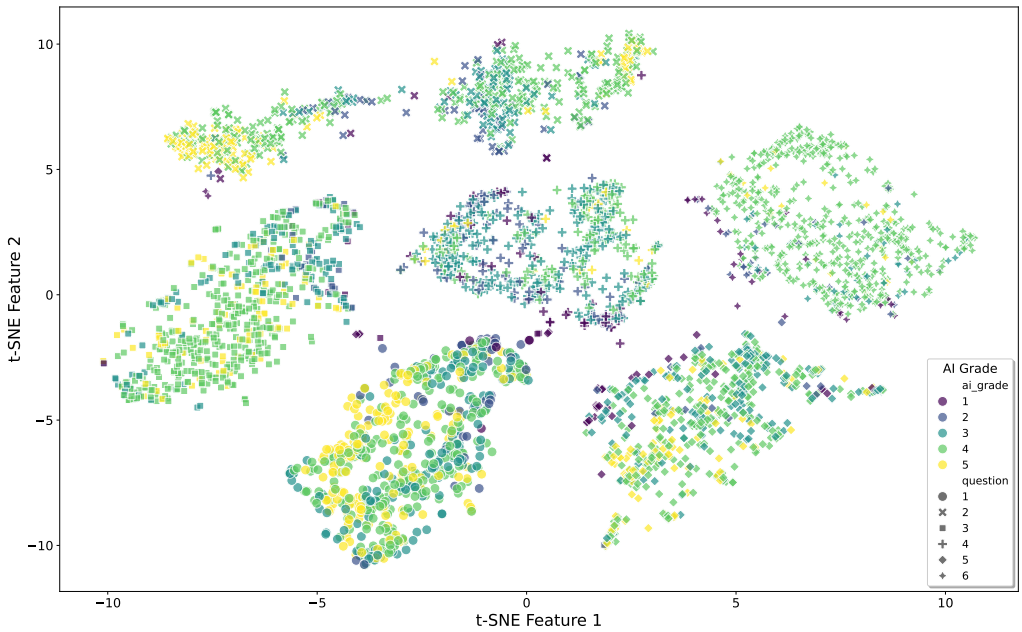
Notes: This table reports a robustness check for an alternative probability cut-off (90%) used to classify an essay as LLM-generated. Columns 1-8 report estimated coefficients from OLS regressions respectively of algorithmic override (final human grade differs from AI grade) overall (Columns (1) and (2)) and when there is initial grade disagreement (Columns (3) and (4)), any grade revision after seeing the AI grade when initial human and AI grades differ (Columns (5) and (6)) and the difference between final human and AI grades (Columns (7) and (8)). All columns include controls for evaluator fixed effect, the even columns additionally include controls for the week application was submitted, length of the application, the applicant's graduation year and an indicator variable for whether the applicant completed their national service. We use the sample of AI-Assisted screening in this analysis.

Figure A.12: t-SNE Clustering of Answer Embeddings by Essay Question and Grade

a) Human initial grades

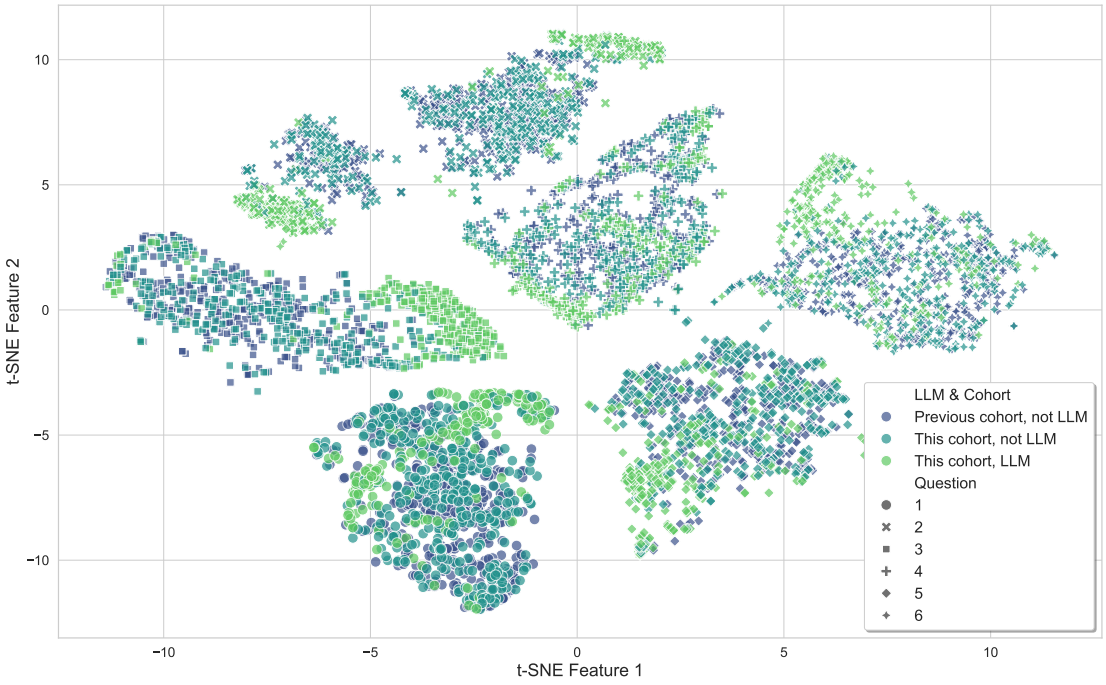


b) AI grades



Notes: The figure shows a two-dimensional t-SNE (t-distributed Stochastic Neighbour Embedding) visualisation of high-dimensional answer embeddings corresponding to responses from the six essay questions and by grade (1 to 5); Panel a shows the visualisation by Human-Only grade, Panel b shows the visualisation by AI-only grade. The embeddings were generated using the "voyage-large-2-instruct" model from Voyage AI, condensed to 50 principal components using PCA and ultimately to two components using t-SNE, a non-linear dimensionality reduction technique. Each point represents an individual answer's embedding.

Figure A.13: Is Semantic Content Different Across LLM and Non-LLM Answers?



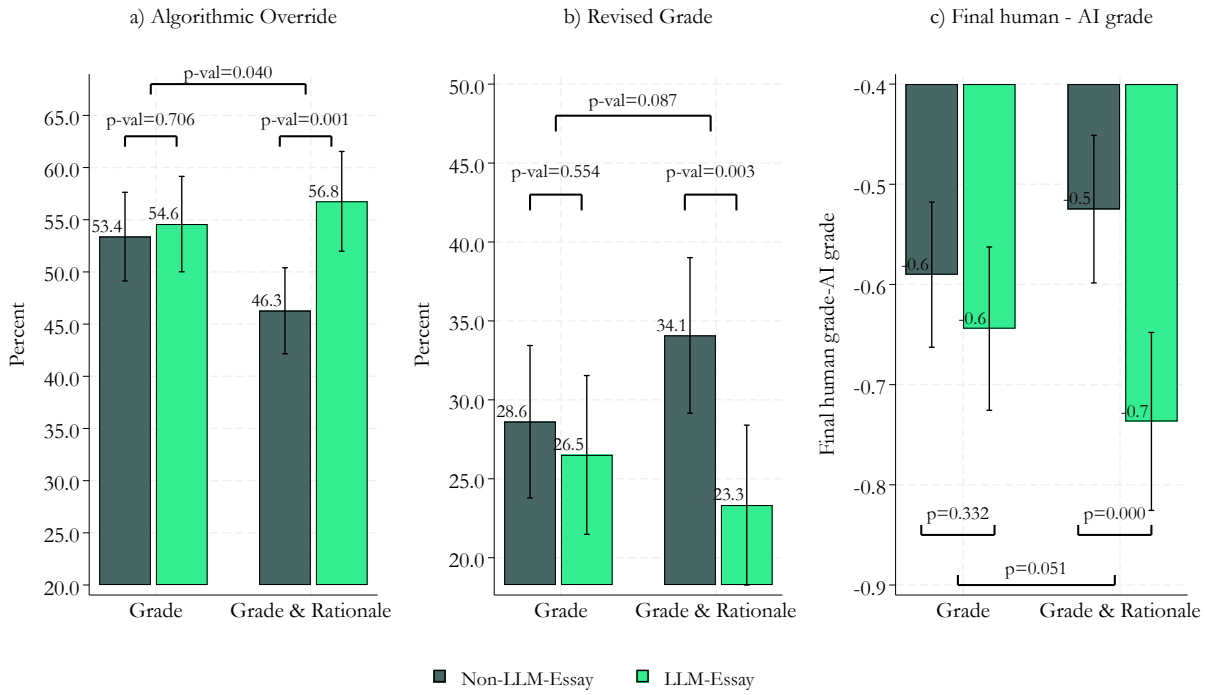
Notes: The figure shows a two-dimensional visualisation of high-dimensional embeddings of responses to the six essay questions. Each point represents a single response, with the marker indicating the question number and the colour representing LLM usage and applicant cohort. Embeddings were generated using the "voyage-large-2-instruct" model from Voyage AI, then reduced to 50 dimensions via PCA before being projected onto two dimensions using t-SNE, a non-linear dimensionality reduction technique. The distance between points reflects the relative semantic similarity of the original high-dimensional embeddings: points that are closer together correspond to answers that are more similar in meaning.

Table A.13: Human Graders Override the Algorithm More When Grading LLM-Written Essays As They Gain More Experience

	Algorithmic Override				Any Revision		Final Grade -AI grade	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
LLM-essay	0.032	0.042	0.061	0.055	-0.010	0.005	-0.122**	-0.108*
	(0.036)	(0.037)	(0.040)	(0.041)	(0.043)	(0.045)	(0.059)	(0.059)
Middle	0.016	0.017	0.050	0.055	-0.041	-0.010	-0.118*	-0.052
	(0.035)	(0.041)	(0.038)	(0.046)	(0.042)	(0.053)	(0.061)	(0.074)
End	0.054	0.065	0.116***	0.139**	-0.128***	-0.086	-0.030	0.090
	(0.039)	(0.052)	(0.039)	(0.058)	(0.045)	(0.066)	(0.065)	(0.100)
LLM-essay x Middle	0.040	0.031	0.084	0.074	-0.123**	-0.117*	0.074	0.075
	(0.052)	(0.052)	(0.052)	(0.053)	(0.058)	(0.061)	(0.086)	(0.086)
LLM-essay x End	0.094*	0.092*	-0.011	-0.018	-0.072	-0.072	-0.243**	-0.258***
	(0.055)	(0.054)	(0.054)	(0.053)	(0.060)	(0.060)	(0.096)	(0.095)
Mean: non-LLM, Start	0.469	0.469	0.709	0.709	0.371	0.371	-0.510	-0.510
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	1,968	1,968	1,265	1,265	1,265	1,265	1,968	1,968

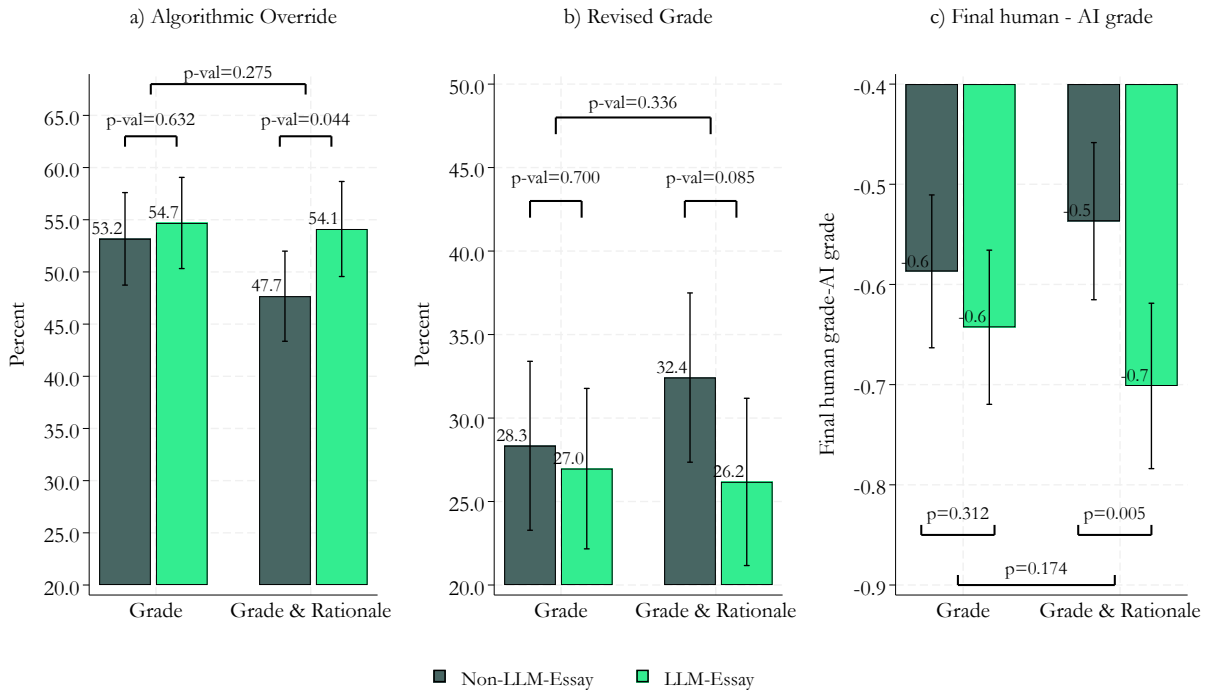
Notes: Columns 1-8 report estimated coefficients from OLS regressions respectively of algorithmic override (final human grade differs from AI grade) overall (Columns (1) and (2)) and when there is initial grade disagreement (Columns (3) and (4)), any grade revision after seeing the AI grade when initial human and AI grades differ (Columns (5) and (6)), and the difference between final human and AI grades (Columns (7) and (8)). All columns include controls for the week application was submitted, evaluator fixed effect, the even columns additionally include controls for length of the application, the applicant's graduation year and an indicator variable for whether the applicant completed their national service. Start, Middle, End refer to the first, second and third tercile of evaluator-level order of applications.

Figure A.14: Algorithmic Override, Grade Revision, and Differences in Final Human and AI Grades by the Type of AI Assistance



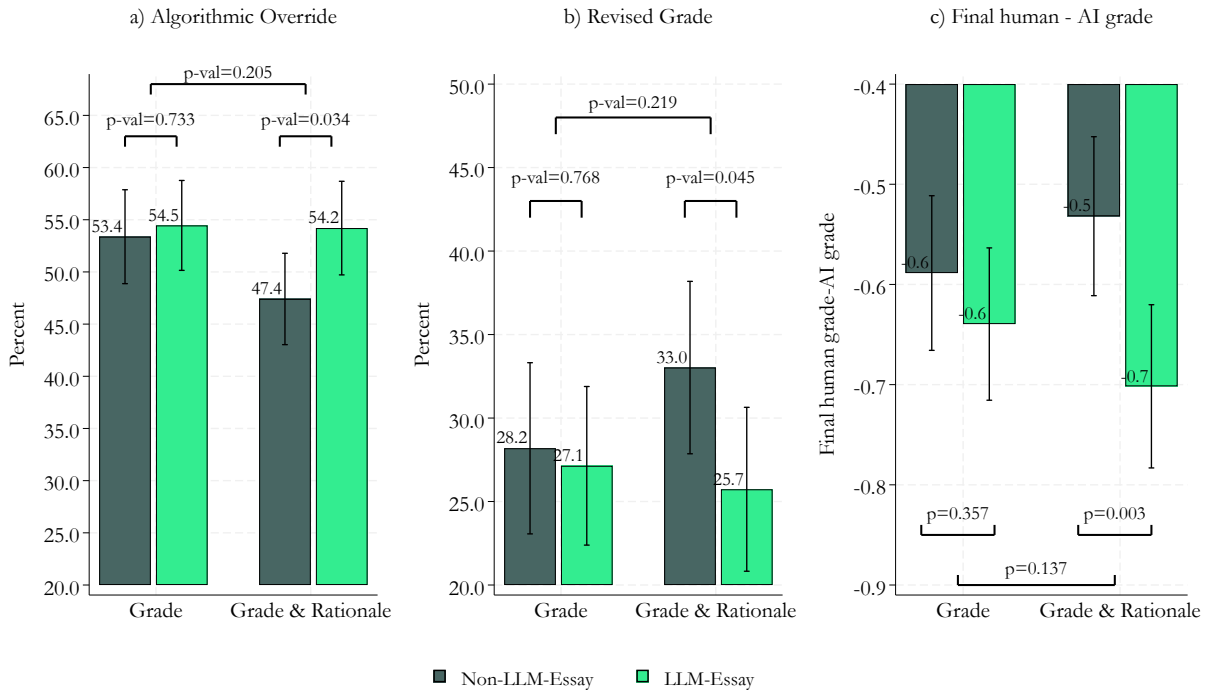
Notes: The figure depicts differences in algorithmic override (Panel a), initial grade revision (Panel b), and the difference between final human and AI grade (Panel c) for the two different types of AI assistance–AI grade & AI grade with rationale. p-values come from t-tests of equality of means.

Figure A.15: Robustness Check (95% cutoff): Algorithmic Override, Grade Revision, and Differences in Final Human and AI Grades by the Type of AI Assistance



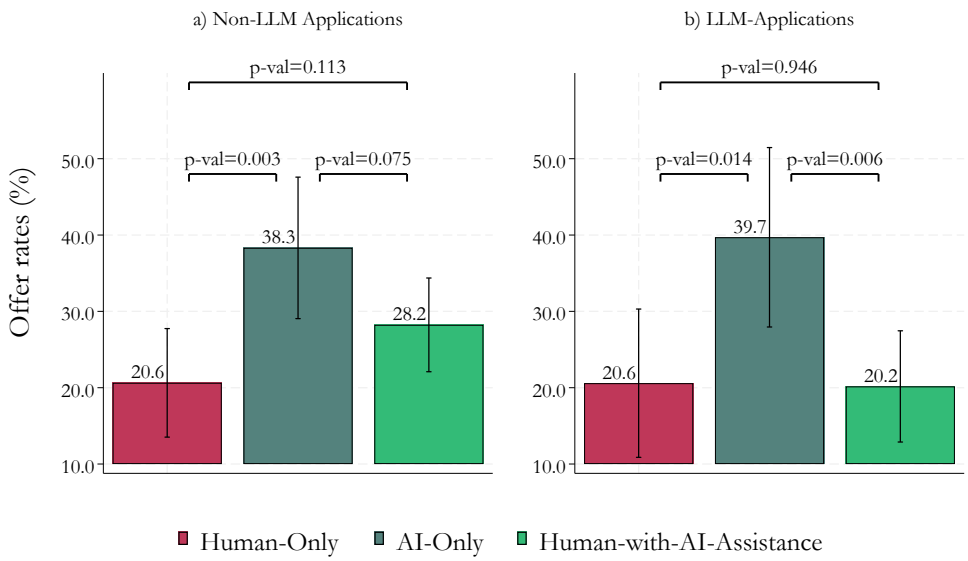
Notes: The figure depicts differences in algorithmic override (Panel a), initial grade revision (Panel b), and the difference between final human and AI grade (Panel c) for the two different types of AI assistance–AI grade & AI grade with rationale. p-values come from t-tests of equality of means.

Figure A.16: Robustness Check (90% cutoff): Algorithmic Override, Grade Revision, and Differences in Final Human and AI Grades by the Type of AI Assistance



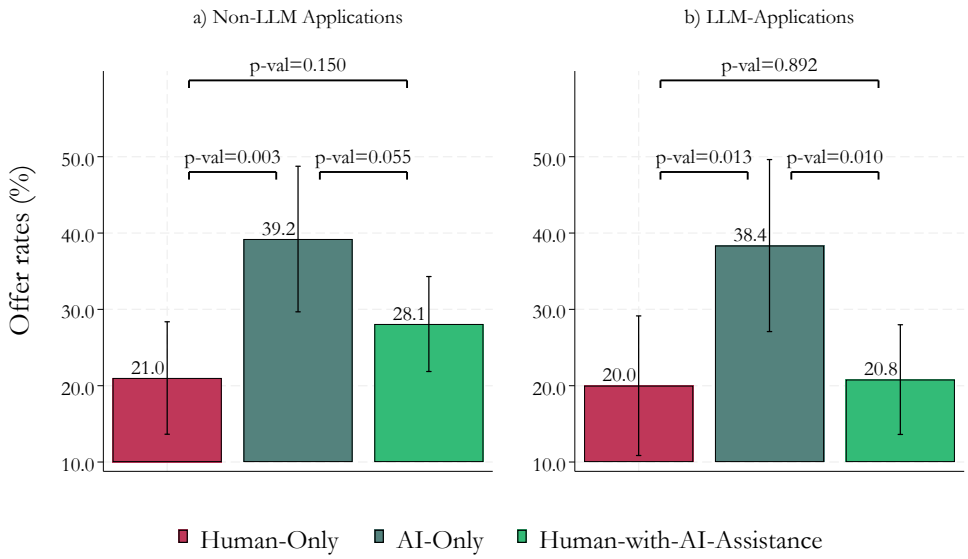
Notes: The figure depicts differences in algorithmic override (Panel a), initial grade revision (Panel b), and the difference between final human and AI grade (Panel c) for the two different types of AI assistance–AI grade & AI grade with rationale. p-values come from t-tests of equality of means.

Figure A.17: Robustness Check (95% cut-off): Offer Rates by Pipeline and LLM-Application



Notes: The figure shows regression coefficients from equation 2 without control variables run separately for a subsample of non-LLM- and LLM-applications, where the outcome variable is a binary indicator of whether a candidate received a job offer. The cranberry bar represents the constant term, i.e. mean offer rates in the Human-Only pipelines, the emerald and mint bars represent the sum of the constant and the respective beta coefficients. Error bars indicate 95% confidence intervals derived from standard errors of the linear combinations of the constant and the beta regression coefficients. The upper p-value compares Human-Only with Human-with-AI-Assistance; the lower-left p-value compares Human-Only with AI-Only; and the lower-right p-value compares AI-Only with Human-with-AI-Assistance. These p-values come from t-tests of the corresponding regression coefficients.

Figure A.18: Robustness Check (90% cut-off): Offer Rates by Pipeline and LLM-Application



Notes: The figure shows regression coefficients from equation 2 without control variables run separately for a subsample of non-LLM- and LLM-applications, where the outcome variable is a binary indicator of whether a candidate received a job offer. The cranberry bar represents the constant term, i.e. mean offer rates in the Human-Only pipelines, the emerald and mint bars represent the sum of the constant and the respective beta coefficients. Error bars indicate 95% confidence intervals derived from standard errors of the linear combinations of the constant and the beta regression coefficients. The upper p-value compares Human-Only with Human-with-AI-Assistance; the lower-left p-value compares Human-Only with AI-Only; and the lower-right p-value compares AI-Only with Human-with-AI-Assistance. These p-values come from t-tests of the corresponding regression coefficients.

Table A.14: LLM-Applications and Downstream Outcomes

	Offer				Offer and LLM-Application		Offer and non-LLM-Application	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI-Only	0.155***	0.146**	0.237***	0.231***	0.085***	2.819***	0.089**	1.830**
	(0.057)	(0.059)	(0.082)	(0.076)	(0.031)	(1.121)	(0.042)	(0.514)
AI-Assistance	0.082*	0.063	-0.035	-0.008	-0.015	0.735	0.064*	1.575*
	(0.047)	(0.047)	(0.063)	(0.065)	(0.021)	(0.316)	(0.034)	(0.398)
Mean (Human-Only)	0.211	0.211	0.197	0.197	0.062	0.062	0.144	0.144
Sample	Non-LLM	Non-LLM	LLM	LLM	Both	Both	Both	Both
Model	OLS	OLS	OLS	OLS	OLS	Logit	OLS	Logit
Controls	No	Yes	No	Yes	Yes	Yes	Yes	Yes
N	477	477	220	220	697	644	697	697
p-values								
$\beta_AI{=}\beta_AIAssistance$	0.181	0.133	0.000	0.001	0.000	0.000	0.527	0.525

Notes: Columns (1)-(4) report estimated coefficients from OLS regressions of an indicator variable for whether the candidate received a fellowship offer. Columns (1) and (2) use the non-LLM subsample, and columns (3) and (4) use the LLM subsample. Columns (5)-(8) pool both subsamples: the outcome is an indicator for receiving an offer and having an LLM-generated application in columns (5) and (6), and an indicator for receiving an offer and having a non-LLM application in columns (7) and (8). Columns (5) and (7) report OLS coefficients; columns (6) and (8) report odds ratios from logistic regressions. Controls include week fixed effects, evaluator fixed effect, the length of the application, the applicant's graduation year and an indicator variable for whether the applicant completed their national service. Standard errors are clustered at the application level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.15: Robustness Check (Cut-off 95%) LLM-Applications and Downstream Outcomes

	Offer				Offer and LLM-Application		Offer and non-LLM-Application	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI-Only	0.177***	0.168***	0.191**	0.182**	0.080**	2.345**	0.094**	1.931**
	(0.059)	(0.062)	(0.077)	(0.075)	(0.032)	(0.889)	(0.041)	(0.553)
AI-Assistance	0.076	0.059	-0.004	0.017	-0.000	0.987	0.050	1.452
	(0.048)	(0.048)	(0.062)	(0.064)	(0.024)	(0.378)	(0.033)	(0.379)
Mean (Human-Only)	0.206	0.206	0.206	0.206	0.072	0.072	0.134	0.134
Sample	Non-LLM	Non-LLM	LLM	LLM	Both	Both	Both	Both
Model	OLS	OLS	OLS	OLS	OLS	Logit	OLS	Logit
Controls	No	Yes	No	Yes	Yes	Yes	Yes	Yes
N	442	442	255	255	697	644	697	697
p-values								
$\beta_AI{=}\beta_AIAssistance$	0.075	0.059	0.006	0.015	0.006	0.008	0.257	0.238

Notes: Columns (1)-(4) report estimated coefficients from OLS regressions of an indicator variable for whether the candidate received a fellowship offer. Columns (1) and (2) use the non-LLM subsample, and columns (3) and (4) use the LLM subsample. Columns (5)-(8) pool both subsamples: the outcome is an indicator for receiving an offer and having an LLM-generated application in columns (5) and (6), and an indicator for receiving an offer and having a non-LLM application in columns (7) and (8). Columns (5) and (7) report OLS coefficients; columns (6) and (8) report odds ratios from logistic regressions. Controls include week fixed effects, evaluator fixed effect, the length of the application, the applicant's graduation year and an indicator variable for whether the applicant completed their national service. Standard errors are clustered at the application level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.16: Robustness Check (Cut-off 90%) LLM-Applications and Downstream Outcomes

	Offer				Offer and LLM-Application		Offer and non-LLM-Application	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI-Only	0.182***	0.169***	0.184**	0.182**	0.080**	2.278**	0.094**	1.956**
	(0.061)	(0.064)	(0.074)	(0.071)	(0.033)	(0.838)	(0.041)	(0.569)
AI-Assistance	0.071	0.047	0.008	0.035	0.003	1.032	0.046	1.433
	(0.049)	(0.049)	(0.059)	(0.062)	(0.025)	(0.381)	(0.032)	(0.380)
Mean (Human-Only)	0.210	0.210	0.200	0.200	0.077	0.077	0.129	0.129
Sample	Non-LLM	Non-LLM	LLM	LLM	Both	Both	Both	Both
Model	OLS	OLS	OLS	OLS	OLS	Logit	OLS	Logit
Controls	No	Yes	No	Yes	Yes	Yes	Yes	Yes
N	424	424	273	273	697	644	697	697
p-values								
$\beta_AI{=}\beta_AIAssistance$	0.055	0.039	0.010	0.025	0.010	0.012	0.223	0.201

Notes: Columns (1)-(4) report estimated coefficients from OLS regressions of an indicator variable for whether the candidate received a fellowship offer. Columns (1) and (2) use the non-LLM subsample, and columns (3) and (4) use the LLM subsample. Columns (5)-(8) pool both subsamples: the outcome is an indicator for receiving an offer and having an LLM-generated application in columns (5) and (6), and an indicator for receiving an offer and having a non-LLM application in columns (7) and (8). Columns (5) and (7) report OLS coefficients; columns (6) and (8) report odds ratios from logistic regressions. Controls include week fixed effects, evaluator fixed effect, the length of the application, the applicant's graduation year and an indicator variable for whether the applicant completed their national service. Standard errors are clustered at the application level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Technical Appendix

B.1 System Prompt

```
You are an expert recruiter very attentive to details.

Always give evaluations in the following format with the  XML delimiters.

<REASONING> Step by step reasoning to get to your choice, with explicit reference to the
↪  specific facts and topics in the answer, in bullet points </REASONING>

<GRADE> An integer from 1 to 5 </GRADE>

<RATIONALE>
WHY n: A short explanation for why you picked the specific grade according to the
↪  criteria that were given to you in the instructions.
WHY NOT n - 1 (for grades greater than 1 only): Why you did not pick one grade below.
WHY NOT n + 1 (for grades smaller than 5 only) : Why you did not pick one grade above
</RATIONALE>
```

B.2 Content Prompts

Question 1 Prompt

```
We are assessing applications for the "[name of the NGO]" fellowship, a program that
↪  provides recent graduates with the opportunity to teach in schools in underprivileged
↪  rural communities throughout the country.
We will provide with a candidate's answer to a question, together with the grading rubric
↪  for that question. The scoring range goes from 1 (lowest) to 5 (highest).
Your task is to grade an answer based on the provided grading rubric as well as how well
↪  the answer addresses the question. To grade the answers, start by determining if the
↪  candidate's response meets the criteria for Grade 1. If it does, move on to Grade 2
↪  criteria, and so on. If the response meets all the criteria for a specific grade but
↪  not the next higher grade, assign the grade for which the criteria are met. For
↪  example, if a response meets all the criteria for Grade 3 but not Grade 4, assign a
↪  grade of 3.

In addition, we provide you with the organization's vision which is relevant for the
↪  candidate selection process:
Vision:
```

"We are working towards 2050 when all children in Ghana will have access to an excellent education, irrespective of their socio-economic background and geographical location. For us, an excellent education is one that equips our children to complete senior high school, with full access to university. Our children will strive for academic excellence, with the ability to think critically about the world around them. They will ask questions, challenge norms, and seek to understand and digest information. They will have control over their financial lives, determine their career choices, and develop a plan to execute their aspirations. They will approach life with a strong sense of possibility, passion, and zeal, with a willingness to address challenges and develop solution-based thinking. Our children will demonstrate a strong level of optimism about their life outcomes. They will have a strong support system of champions and the social and cultural capital to engage successfully and succeed in the current system but keenly aware of its flaws. They will develop the ethical mindsets that guide their everyday interactions and will value honesty and integrity. Our children will act as consciously driven citizens aware of the systems of injustice that exist and believe that a more equitable system is achievable in Ghana and abroad."

QUESTION: "Why do you want to be a [name of the NGO] Fellow?"
The purpose is to measure to what extent the candidate shares LFG's values, believes LFG goals are attainable, is open to our approach to reaching them, and wants to pursue them relentlessly.

GRADING:
Grade 1: Does not give a reason for wanting to be an LFG Fellow.
- No personal experience or background related to education or underprivileged communities mentioned
- No passion or commitment to education and social change expressed
- No demonstrated leadership skills or potential
- Lack of clarity and coherence in response
- No specific examples or plans for contributing to LFG's vision

Grade 2: Gives a reason that is not linked to the LFG vision or approach.
- May mention personal experience or background, but not directly related to education or underprivileged communities
- Limited passion or commitment to education and social change
- Limited or no demonstrated leadership skills or potential
- Some clarity and coherence in response, but not directly linked to LFG's vision
- No specific examples or plans for contributing to LFG's vision

Grade 3: Gives a reason that is clearly linked to solving educational inequity in Ghana.

```
- Personal experience or background related to education or underprivileged communities
↪  mentioned
- Clear passion and commitment to education and social change
- Some demonstrated leadership skills or potential
- Clarity and coherence in response, directly linked to LFG's vision
- No specific examples or plans for contributing to LFG's vision

Grade 4: Can articulate elements of the Fellowship that they are most interested in for
↪  their own development.
- Personal experience or background related to education or underprivileged communities
↪  mentioned
- Strong passion and commitment to education and social change
- Demonstrated leadership skills or potential
- Clarity and coherence in response, directly linked to LFG's vision and Fellowship
↪  elements
- Some specific examples or plans for contributing to LFG's vision

Grade 5: Gives rationale for own desire to be a fellow and is able to talk about how past
↪  OR future activities connect to the [name of the NGO] vision.
- Personal experience or background related to education or underprivileged communities
↪  mentioned and connected to LFG's vision
- Strong passion and commitment to education and social change
- Demonstrated leadership skills or potential, with past or future activities connected
↪  to LFG's vision
- Clarity and coherence in response, directly linked to LFG's vision and Fellowship
↪  elements
- Specific examples or plans for contributing to LFG's vision, showing a deep
↪  understanding of the organization's mission and goals
Please note that the grading rubric follows a progression where each grade encompasses
↪  the criteria of the lower grades as well.

Definition of terms in the rubric:
1. Personal experience or background related to education and/or underprivileged
↪  communities: Candidates who share their own experiences or background related to
↪  education, especially in underprivileged communities, may receive higher grades as
↪  they demonstrate a personal connection to LFG's vision and goals.

2. Passion and commitment to education and social change: Candidates who express a strong
↪  passion and commitment to education and social change may receive higher grades, as
↪  this indicates their dedication to LFG's mission and their potential to make a
↪  significant impact.
```

```
3. Demonstrated leadership skills or potential: Candidates who showcase their leadership
↪  skills or potential, either through past experiences or future aspirations, may
↪  receive higher grades, as this indicates their ability to take initiative and
↪  contribute effectively to LFG's goals.

4. Clarity and coherence of response: Candidates who provide clear and coherent answers,
↪  effectively communicating their thoughts and ideas, may receive higher grades, as
↪  this demonstrates their ability to articulate their motivations and goals in a
↪  compelling manner.

5. Specific examples or plans for contributing to LFG's vision: Candidates who provide
↪  specific examples or plans for how they would contribute to LFG's vision and goals
↪  may receive higher grades, as this demonstrates their understanding of the
↪  organization's mission and their ability to think critically about how they can make
↪  a meaningful impact.

Answer:
"+++ANSWER_TEXT_HERE+++"
```

Question 2 Prompt

```
We are assessing applications for the "[name of the NGO]" fellowship, a program that
↪  provides recent graduates with the opportunity to teach in schools in underprivileged
↪  rural communities throughout the country.
We will provide with a candidate's answer to a question, together with the grading rubric
↪  for that question. The scoring range goes from 1 (lowest) to 5 (highest).
Your task is to grade an answer based on the provided grading rubric as well as how well
↪  the answer addresses the question. To grade the answers, start by determining if the
↪  candidate's response meets the criteria for Grade 1. If it does, move on to Grade 2
↪  criteria, and so on. If the response meets all the criteria for a specific grade but
↪  not the next higher grade, assign the grade for which the criteria are met. For
↪  example, if a response meets all the criteria for Grade 3 but not Grade 4, assign a
↪  grade of 3.


QUESTION: "What is an excellent education to you? And during your two years as a [name of
↪  the NGO] fellow, how would you provide your students with an excellent education?
↪  Include details of the goals you would set for your students and how you would set
↪  out to achieve them."
The purpose is to measure to what extent the candidate shares LFG's values, believes LFG
↪  goals are attainable, is open to our approach to reaching them, and wants to pursue
↪  them relentlessly.

GRADING RUBRIC:
```

```
Grade 1: Does not define what an excellent education is and / does not articulate how to
↪  provide that to their students.
- Lacks personal experiences and background
- Shows no adaptability and flexibility
- Lacks passion and enthusiasm
- Poor communication and organization
- Lacks problem-solving and critical thinking skills

Grade 2: Defines what an excellent education is but does not articulate how to provide
↪  that to their students.
- Shares some personal experiences and background
- Shows limited adaptability and flexibility
- Displays some passion and enthusiasm
- Adequate communication and organization
- Lacks problem-solving and critical thinking skills

Grade 3: Clearly defines what an excellent education is and shows a pathway to providing
↪  that to their students.
- Shares relevant personal experiences and background
- Demonstrates adaptability and flexibility
- Displays passion and enthusiasm
- Clear communication and organization
- Some problem-solving and critical thinking skills

Grade 4: Rubric 3 plus: articulates factors that lead to academic achievement, mindset
↪  development, exposure to resources.
- Shares insightful personal experiences and background
- Demonstrates strong adaptability and flexibility
- Displays strong passion and enthusiasm
- Excellent communication and organization
- Good problem-solving and critical thinking skills

Grade 5: Rubric 4 plus: gives specific examples of actions they will take as a fellow and
↪  alumni to provide an excellent education to their students.
- Shares compelling personal experiences and background
- Demonstrates exceptional adaptability and flexibility
- Displays outstanding passion and enthusiasm
- Exceptional communication and organization
- Excellent problem-solving and critical thinking skills

Please note that the grading rubric follows a progression where each grade encompasses
↪  the criterion of the lower grades as well.
```

```
Definition of terms in the rubric:
1. Personal experiences and background: Candidates who share their personal experiences
↪  and how they relate to their understanding of excellent education may be given higher
↪  grades. This shows their genuine interest and commitment to the cause.

2. Adaptability and flexibility: Candidates who demonstrate their ability to adapt to
↪  different situations and be flexible in their approach to teaching may be given
↪  higher grades. This shows their willingness to learn and grow as educators.

3. Passion and enthusiasm: Candidates who express their passion and enthusiasm for
↪  teaching and making a difference in the lives of underprivileged children may be
↪  given higher grades. This shows their dedication and motivation to succeed as a [name
↪  of the NGO] fellow.

4. Clear communication and organization: Candidates who present their ideas clearly and
↪  in an organized manner may be given higher grades. This shows their ability to
↪  effectively communicate their thoughts and plans to others.

5. Problem-solving and critical thinking skills: Candidates who demonstrate their ability
↪  to think critically and solve problems in their approach to providing an excellent
↪  education may be given higher grades. This shows their ability to analyze situations
↪  and come up with effective solutions.

Answer:
"+++ANSWER_TEXT_HERE+++"
```

Question 3 Prompt

```
We are assessing applications for the "[name of the NGO]" fellowship, a program that
↪  provides recent graduates with the opportunity to teach in schools in underprivileged
↪  rural communities throughout the country.
We will provide with a candidate's answer to a question, together with the grading rubric
↪  for that question. The scoring range goes from 1 (lowest) to 5 (highest).
Your task is to grade an answer based on the provided grading rubric as well as how well
↪  the answer addresses the question. To grade the answers, start by determining if the
↪  candidate's response meets the criteria for Grade 1. If it does, move on to Grade 2
↪  criteria, and so on. If the response meets all the criteria for a specific grade but
↪  not the next higher grade, assign the grade for which the criteria are met. For
↪  example, if a response meets all the criteria for Grade 3 but not Grade 4, assign a
↪  grade of 3.


In addition, we provide you with the organization's vision which is relevant for the
↪  candidate selection process:
```

```
Vision:
"We are working towards 2050 when all children in Ghana will have access to an excellent
↪ education, irrespective of their socio-economic background and geographical location.
For us, an excellent education is one that equips our children to complete senior high
↪ school, with full access to university. Our children will strive for academic
↪ excellence, with the ability to think critically about the world around them. They
↪ will ask questions, challenge norms, and seek to understand and digest information.
↪ They will have control over their financial lives, determine their career choices,
↪ and develop a plan to execute their aspirations. They will approach life with a
↪ strong sense of possibility, passion, and zeal, with a willingness to address
↪ challenges and develop solution-based thinking. Our children will demonstrate a
↪ strong level of optimism about their life outcomes. They will have a strong support
↪ system of champions and the social and cultural capital to engage successfully and
↪ succeed in the current system but keenly aware of its flaws. They will develop the
↪ ethical mindsets that guide their everyday interactions and will value honesty and
↪ integrity. Our children will act as consciously driven citizens aware of the systems
↪ of injustice that exist and believe that a more equitable system is achievable in
↪ Ghana and abroad."

QUESTION: "At [name of the NGO], we are working to create a growing network of leaders
↪ who will work at every level of education, policy and other professions to ensure
↪ that all children in Ghana will have the opportunity to attain an excellent
↪ education. As a [name of the NGO] alumni, how do you envision yourself contributing
↪ to the [name of the NGO] alumni vision?"
The purpose is to measure to what extent the candidate shares LFG's values, believes LFG
↪ goals are attainable, is open to our approach to reaching them, and wants to pursue
↪ them relentlessly.

GRADING RUBRIC:
Grade 1:
- Does not demonstrate an understanding of the LFG alumni vision.
- Lacks clarity and coherence in the answer.
- Shows little to no passion or commitment to the LFG vision and goals.
- Does not draw from personal experiences or background.
- Offers no creative or innovative ideas.
- Does not emphasize collaboration and teamwork.

Grade 2:
- Understands the LFG alumni vision but does not articulate their role in achieving it.
- Provides a somewhat clear and coherent answer.
- Shows some passion and commitment to the LFG vision and goals.
- May draw from personal experiences or background, but not effectively.
```

```
- Offers few creative or innovative ideas.
- Mentions collaboration and teamwork but does not elaborate on its importance.

Grade 3:
- Understands the LFG alumni vision and can articulate their role in achieving the
↪  vision.
- Provides a clear and coherent answer.
- Demonstrates passion and commitment to the LFG vision and goals.
- Effectively draws from personal experiences and background.
- Offers some creative and innovative ideas.
- Emphasizes the importance of collaboration and teamwork.

Grade 4:
- Rubric 3 plus: gives more than one example of how they're going to achieve the alumni
↪  vision.
- Provides a very clear and coherent answer.
- Shows strong passion and commitment to the LFG vision and goals.
- Effectively draws from personal experiences and background to support multiple
↪  examples.
- Offers multiple creative and innovative ideas.
- Strongly emphasizes the importance of collaboration and teamwork.

Grade 5:
- Rubric 4 plus: mentions a specific sector/job they have in mind and how they intend to
↪  leverage their position to achieve the LFG alumni vision.
- Provides an exceptionally clear and coherent answer.
- Demonstrates outstanding passion and commitment to the LFG vision and goals.
- Effectively draws from personal experiences and background to support specific
↪  sector/job plans.
- Offers numerous creative and innovative ideas related to the specific sector/job.
- Emphasizes the importance of collaboration and teamwork in achieving the LFG alumni
↪  vision within the specific sector/job.
Please note that the grading rubric follows a progression where each grade encompasses
↪  the criteria of the lower grades as well.

Definition of terms in the rubric:
1. Clarity and coherence of the answer: Candidates who provide clear and well-structured
↪  answers that effectively communicate their ideas and vision are likely to receive
↪  higher grades.

2. Demonstrated passion and commitment: Candidates who show genuine enthusiasm and
↪  dedication to the LFG vision and goals may receive higher grades, as this indicates a
↪  strong motivation to contribute to the organization's mission.
```

```
3. Personal experiences and background: Candidates who can draw from their own
↪  experiences and background to support their ideas and vision may receive higher
↪  grades, as this demonstrates a deeper understanding of the issues and challenges
↪  faced by underprivileged children in Ghana.

4. Creativity and innovation: Candidates who propose unique and innovative ideas for
↪  contributing to the LFG alumni vision may receive higher grades, as this indicates a
↪  willingness to think outside the box and explore new approaches to solving problems.

5. Collaboration and teamwork: Candidates who emphasize the importance of working
↪  together with fellow alumni and other stakeholders to achieve the LFG vision may
↪  receive higher grades, as this demonstrates an understanding of the need for
↪  collective action and cooperation in order to create lasting change.

Answer:
"+++ANSWER_TEXT_HERE+++"
```

Question 4 Prompt

```
We are assessing applications for the "[name of the NGO]" fellowship, a program that
↪  provides recent graduates with the opportunity to teach in schools in underprivileged
↪  rural communities throughout the country.
We will provide with a candidate's answer to a question, together with the grading rubric
↪  for that question. The scoring range goes from 1 (lowest) to 5 (highest).
Your task is to grade an answer based on the provided grading rubric as well as how well
↪  the answer addresses the question. To grade the answers, start by determining if the
↪  candidate's response meets the criteria for Grade 1. If it does, move on to Grade 2
↪  criteria, and so on. If the response meets all the criteria for a specific grade but
↪  not the next higher grade, assign the grade for which the criteria are met. For
↪  example, if a response meets all the criteria for Grade 3 but not Grade 4, assign a
↪  grade of 3.


In addition, we provide you with the organization's core beliefs, which are relevant for
↪  the candidate selection process:
Core beliefs:
"These core beliefs form the foundation that guides our work and how we engage with each
↪  other and the communities we serve. They are inflexible, and they determine the
↪  strategies we employ to fulfill our mission. As these beliefs speak to who we are,
↪  they are naturally timeless and not used individually, but as a whole.
```

```
Responsibility is mutual: Through humility, integrity, respect, and openness, we seek
↪ answers that make our community stronger. And through the fidelity of our ideas, we
↪ are committed to improving the welfare of the individuals we work with. It is what we
↪ do together that makes us stronger.
Innovation is simple: We are committed to introducing innovative solutions, molding
↪ systems and challenging standards to produce new ideas that are easy to understand,
↪ apply, and proliferate. We work with sincerity and diligence to invent the future.
Impossible is nothing: Our imagination is limitless. We believe in the full human
↪ development of every child, and to affirm this sacred belief, we have dedicated
↪ ourselves to realizing the possibility of an excellent education for every child."

QUESTION: "How do our core beliefs resonate with you?"
The purpose is to measure to what extent the candidate shares LFG's values, believes LFG
↪ goals are attainable, is open to our approach to reaching them, and wants to pursue
↪ them relentlessly.

GRADING RUBRIC:
Grade 1:
- Does not make reference to any of our core beliefs.
- Lacks clarity and coherence in the response.
- No personal connection or passion demonstrated.
- No examples or experiences shared.
- Limited understanding of the core beliefs and their implications.
- No problem-solving or critical thinking skills showcased.

Grade 2:
- Makes some reference to our core beliefs but does not articulate how they resonate with
↪ them.
- Some clarity and coherence in the response.
- Minimal personal connection or passion demonstrated.
- Few or no examples or experiences shared.
- Basic understanding of the core beliefs and their implications.
- Limited problem-solving or critical thinking skills showcased.

Grade 3:
- Makes reference to our core beliefs and articulates how they resonate with them.
- Clear and coherent response.
- Personal connection and passion demonstrated.
- Some examples or experiences shared.
- Good understanding of the core beliefs and their implications.
- Some problem-solving or critical thinking skills showcased.
```

```
Grade 4:
- Rubric 3 plus: shares an example of how at least one of our beliefs resonates with
↪  them.
- Clear and coherent response with strong personal connection and passion demonstrated.
- Multiple examples or experiences shared.
- Deep understanding of the core beliefs and their implications.
- Problem-solving and critical thinking skills showcased in relation to at least one core
↪  belief.

Grade 5:
- Rubric 4 plus: shares an example of how all three core beliefs resonate with them.
- Exceptionally clear and coherent response with a strong personal connection and passion
↪  demonstrated.
- Multiple examples or experiences shared that relate to all three core beliefs.
- Comprehensive understanding of the core beliefs and their implications.
- Strong problem-solving and critical thinking skills showcased in relation to all three
↪  core beliefs.
Please note that the grading rubric follows a progression where each grade encompasses
↪  the criteria of the lower grades as well.

Definition of terms in the rubric:
1. Clarity and coherence of the response: Candidates who provide clear and
↪  well-structured answers that effectively communicate their thoughts and ideas are
↪  likely to receive higher grades.

2. Personal connection and passion: Candidates who demonstrate a strong personal
↪  connection to the core beliefs and show genuine passion for the mission of [name of
↪  the NGO] may receive higher grades.

3. Examples and experiences: Candidates who provide specific examples and share personal
↪  experiences that relate to the core beliefs are likely to receive higher grades.

4. Depth of understanding: Candidates who demonstrate a deep understanding of the core
↪  beliefs and their implications for the work of [name of the NGO] may receive higher
↪  grades.

5. Problem-solving and critical thinking: Candidates who showcase their problem-solving
↪  skills and critical thinking abilities in their responses, particularly in relation
↪  to the core beliefs, may receive higher grades.

Answer:
"+++ANSWER_TEXT_HERE+++"
```

Question 5 Prompt

```
We are assessing applications for the "[name of the NGO]" fellowship, a program that
↪ provides recent graduates with the opportunity to teach in schools in underprivileged
↪ rural communities throughout the country.
We will provide with a candidate's answer to a question, together with the grading rubric
↪ for that question. The scoring range goes from 1 (lowest) to 5 (highest).
Your task is to grade an answer based on the provided grading rubric as well as how well
↪ the answer addresses the question. To grade the answers, start by determining if the
↪ candidate's response meets the criteria for Grade 1. If it does, move on to Grade 2
↪ criteria, and so on. If the response meets all the criteria for a specific grade but
↪ not the next higher grade, assign the grade for which the criteria are met. For
↪ example, if a response meets all the criteria for Grade 3 but not Grade 4, assign a
↪ grade of 3.


QUESTION: Working in a [name of the NGO] partner school and community requires you to be
↪ able to sustain commitments over a long period of time irrespective of external
↪ challenges. Please describe a time when you overcame a challenge in order to achieve
↪ a non-academic goal. Please ensure the example used is recent (i.e. within the last 3
↪ to 4 years) and from a professional or extracurricular/voluntary context.
The purpose is to measure how the candidate sustains commitment and involvement over
↪ time.

GRADING RUBRIC:
Grade 1:
- Does not describe a challenge.
- Answer lacks clarity and coherence.
- No specific examples or details provided.

Grade 2:
- Describes a challenge(s) but does not share how they overcame the challenge(s).
- Answer may have some clarity and coherence but lacks specificity and detail.
- Limited demonstration of resilience and adaptability.

Grade 3:
- Clearly defines a robust challenge and shares how they overcame the challenge.
- Answer is clear, coherent, and provides specific examples and details.
- Demonstrates resilience and adaptability in overcoming the challenge.
- Some evidence of impact and results.

Grade 4:
- Rubric 3 plus: shares more than one robust challenge and how they overcame them.
```

```
- Answer is well-structured and provides multiple specific examples and details.
- Strong demonstration of resilience and adaptability in overcoming multiple challenges.
- Clear evidence of impact and results.

Grade 5:
- Rubric 4 plus: articulates what they would have done differently.
- Answer is highly coherent and provides a comprehensive account of challenges and
↪  solutions.
- Exceptional demonstration of resilience and adaptability in overcoming challenges.
- Significant impact and results achieved.
- Demonstrates personal growth and learning from experiences.
Please note that the grading rubric follows a progression where each grade encompasses
↪  the criteria of the lower grades as well.

Definition of terms in the rubric:
1. Clarity and coherence of the answer: A well-structured and coherent answer that
↪  clearly addresses the question is more likely to receive a higher grade.

2. Specificity and detail: Answers that provide specific examples and details about the
↪  challenge(s) faced and the steps taken to overcome them are more likely to receive
↪  higher grades.

3. Demonstrated resilience and adaptability: Answers that show the candidate's ability to
↪  adapt to changing circumstances and persevere in the face of adversity are more
↪  likely to receive higher grades.

4. Impact and results: Answers that demonstrate the positive impact of the candidate's
↪  actions and the tangible results achieved are more likely to receive higher grades.

5. Personal growth and learning: Answers that show the candidate's ability to learn from
↪  their experiences and apply those lessons to future challenges are more likely to
↪  receive higher grades.

Answer:
"+++ANSWER_TEXT_HERE+++"
```

Question 6 Prompt

```
We are assessing applications for the "[name of the NGO]" fellowship, a program that
↪  provides recent graduates with the opportunity to teach in schools in underprivileged
↪  rural communities throughout the country.
We will provide with a candidate's answer to a question, together with the grading rubric
↪  for that question. The scoring range goes from 1 (lowest) to 5 (highest).
```

```
Your task is to grade an answer based on the provided grading rubric as well as how well
↪ the answer addresses the question. To grade the answers, start by determining if the
↪ candidate's response meets the criteria for Grade 1. If it does, move on to Grade 2
↪ criteria, and so on. If the response meets all the criteria for a specific grade but
↪ not the next higher grade, assign the grade for which the criteria are met. For
↪ example, if a response meets all the criteria for Grade 3 but not Grade 4, assign a
↪ grade of 3.

QUESTION: "Please share with us two (2) instances when you were in a position of
↪ influence and motivated others (a team or group of people) to make a desired change
↪ and achieved a desired outcome. The example you give can either be of a formal or
↪ informal position and from any context, but it should be a recent example (i.e.
↪ within the last 3 to 4 years)."
The purpose is to measure how the candidate sustains commitment and involvement over
↪ time.

GRADING RUBRIC:
Grade 1:
- Does not describe a clear position of influence and the people they motivated.
- Lacks clarity and coherence in the answer.
- Provides little to no specific details or examples.

Grade 2:
- Describes some position of influence but does not articulate how they motivated others
↪ to take a desired action.
- Answer is somewhat clear and coherent.
- Provides limited specific details or examples.
- Minimal demonstration of personal initiative and leadership.

Grade 3:
- Clearly describes two robust positions of influence and shares examples of how they
↪ motivated others to take desired actions.
- Answer is clear and coherent.
- Provides specific details and examples.
- Demonstrates personal initiative and leadership.
- Shows some emotional intelligence and empathy.

Grade 4:
- Rubric 3 plus: articulates the outcomes of the actions.
- Answer is very clear and coherent.
- Provides detailed and specific examples.
- Demonstrates significant impact on people or situations.
```

```
- Shows strong personal initiative and leadership.
- Exhibits emotional intelligence and empathy.

Grade 5:
- Rubric 4 plus: shares an exceptional position of influence (a position that affects a
↪  large group of people i.e more than 100 people) and clear.
- Answer is exceptionally clear and coherent.
- Provides extensive specific details and examples.
- Demonstrates substantial impact on people or situations.
- Exhibits exceptional personal initiative and leadership.
- Displays outstanding emotional intelligence and empathy.
Please note that the grading rubric follows a progression where each grade encompasses
↪  the criteria of the lower grades as well.

Definition of terms in the rubric:
1. Clarity and coherence of the answer: Answers that are well-structured, easy to
↪  understand, and logically organized may receive higher grades.

2. Specificity and detail: Answers that provide specific examples, names, dates, or
↪  locations may be graded higher than those with vague or generic descriptions.

3. Demonstrated impact: Answers that show a clear and significant impact on the people or
↪  situation involved may receive higher grades.

4. Personal initiative and leadership: Answers that demonstrate the candidate's personal
↪  initiative, problem-solving skills, and ability to lead others may be graded higher.

5. Emotional intelligence and empathy: Answers that show the candidate's ability to
↪  understand and respond to the emotions and needs of others may receive higher grades.

Answer:
"+++ANSWER_TEXT_HERE+++"
```