Finding Talent in the Age of AI *

Kobbina Awuah† Urša Krenk‡ David Yanagizawa-Drott§

April 10, 2025

Abstract

The recent mass adoption of generative artificial intelligence (AI), catalyzed by the rise of ChatGPT, has changed matching in labor markets. On the demand side, employers increasingly rely on AI to search for and screen potential candidates, while on the supply side, job seekers use AI to enhance their applications. We study this two-sided phenomenon by experimentally embedding a standard AI algorithm (GPT-4) into an organization's screening process for teacher recruitment. We find that fully automating this process increases hiring success by approximately 11 percentage points (a 73% improvement). However, when using the same algorithm as an assistant, evaluators frequently override its recommendations, and this assistance does not improve downstream hiring outcomes. We check whether this demand-side behavior can be explained by the widespread use of LLMs on the supply side, when writing application essays. We document that approximately 60% of applicants rely on LLM-generated text. Compared to the algorithm, human screeners tend to discount such applications, despite this not being part of the formal screening criteria. They also override the algorithmic decisions more often when the application essays are written using an LLM. Our results indicate that AI usage on the supply side affects decision-making on the demand side, including the effectiveness of algorithms as recruitment aids. Additionally, we suggest that automating the screening process can enhance labor market matching efficiency.

Keywords: Artificial intelligence, technology adoption, screening, labor markets.

*We are grateful to Maria Korobeynikova, Minh Trinh, Andrin Pluess, and Alessandro Vanzo for excellent research assistance. 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

The widespread adoption of generative artificial intelligence (AI) tools—exemplified by ChatGPT and other large language models (LLMs)—is reshaping labor markets. On the supply side, job seekers increasingly rely on AI to create polished application materials, including resumes and essays that highlight their personal experiences, values, and skill sets. On the demand side, employers experiment with integrating AI into their search and screening processes, with the aim of more efficiently identifying top talent. However, the impact of these shifts remains unclear. For instance, does making it easier to produce refined application materials weaken the ability of employers to discern genuine talent and fit? How do human evaluators adjust their assessments when AI-generated content is involved? And when AI tools are used to assist in screening, how do these dynamics interact?

On a conceptual level, there are several nontrivial factors that could affect the matching process. To fix ideas, consider an opening for a teacher position in a school. First, the use of generative AI can reduce the cost of producing polished resumes and cover letters demonstrating a good fit for the school. The technology could lead to a proliferation of seemingly high-quality applications, especially since soft skills, typically emphasized in cover letters, are important for the role. However, the true underlying quality of the applicant pool may not be higher. In fact, the real signal from these applications could vanish altogether, as the ability to craft well-written documents becomes less indicative of a candidate's true capabilities. On the other hand, on the demand side, organizations face a difficult challenge: they need to consider whether AI usage, or proficiency in AI skills itself, should be a screening criterion. For example, consider an applicant who uses the tool to explain their teaching philosophy, outlining which pedagogical approaches they believe best help children learn. The tool may help the applicant articulate their true views. However, it may also simply generate text that sounds impressive, based on the general consensus for the job, without genuinely reflecting applicant's beliefs. If the text is obviously AI-generated, this could also impact the perceived quality of the applicant, as evaluators may interpret reliance on AI tools as indicative of lower effort, motivation, or authenticity. Finally, organizations face another policy choice: whether to use generative AI as a screening tool to augment human decision-making or to automate the process entirely. The latter option is especially appealing to firms, as it allows them to screen applicants at scale, significantly reducing the cost and increasing speed compared to human labor. Although theories of signaling (Spence, 1973 and Stiglitz, 1975) and labor market search (Mortensen and Pissarides, 1994) generally point

out that costs of signaling and search matter, it is unclear how the rise of generative AI – when used on both sides – ultimately influences labor market outcomes. This paper presents evidence on this.

To study these phenomena, we partner with a nonprofit organization that hires recent university graduates to teach in deprived rural schools in Ghana. These teaching fellowships are prestigious and competitive, with approximately 15–20% of applicants ultimately receiving offers. Essays play a crucial role: candidates are asked to write about specific experiences and perspectives on certain topics, and their responses are ranked against a detailed rubric designed to measure candidates' fit, tenacity, and leadership potential.[1] Historically, these essays were written by applicants, and the organization relied on its staff to screen candidates for advancement to in-person interviews. Writing essays was perceived as a highly costly signal, intended to demonstrate not only the quality of applicants but also their motivation to join the organization, as reflected by the effort invested in crafting the essays. However, the rise of generative AI has made essay production significantly easier, raising questions about how evaluators might respond—particularly if they also use AI tools in their assessment processes. We embed an AI-based grading algorithm into this screening process and randomly assign applications to one of three evaluation pipelines: (1) *Human-Only*, where evaluators rely solely on their judgment as done historically; (2) *Human-with-AI-Assistance*, where evaluators record an initial grade and then see a GPT-4-generated recommendation before finalizing it; and (3) *AI-Only*, where GPT-4's evaluation dictates who advances. First, this design lets us compare how conventional human evaluation, partial AI integration, and full automation perform in identifying candidates who excel at later in-person assessments and ultimately receive job offers. Note that, although the final decision on which grade counts is randomized, we implement parallel grading, ensuring that every essay is evaluated

[1]For example, the rubric assigns numeric grades (1–5) based on how well essays address several open-ended questions related to the organization's vision, the applicant's understanding of what constitutes an excellent education, their familiarity with alumni goals, their alignment with core values, their resilience in overcoming challenges, and their leadership ability. As we will describe in detail below, the AI screening tool was designed to follow the criteria, just as the human screeners were. Two things are noteworthy. First, whether applicants were deemed to have used LLMs to generate answers was not a criterion. In practice, human screeners deviated from these instructions, whereas the algorithm did not. Second, we study generative AI as a screening tool. This is in contrast with supervised machine learning approaches, where prediction models are first trained based on historical, human-labeled, data. The latter approach may be superior from a talent prediction accuracy perspective, for example on downstream measures of quality, but is typically more costly to develop. It also requires the firm to have expertise in machine learning, which most firms do not have. These factors likely explain why generative AI has become so widely adopted, whereas supervised machine learning tools have not. As such, our paper studies a technology that potentially has widespread implications for matching in labor markets.

by both a human (with or without AI assistance) and by the AI. Second, it enables us to examine how human grading – with or without AI assistance – responds to AI-generated content.

First, we document that when we remove humans from the pipeline entirely and rely solely on the AI model, the downstream offer and hiring rates increase substantially—approximately an 84% and 73% improvement, respectively, over the Human-Only baseline. However, when we provide evaluators with the same algorithm as an assistant, we find that this does not improve downstream offer rates or hiring success significantly. We find that in a majority of cases, humans override the algorithm, which aligns well with research on algorithmic aversion (Dietvorst et al., 2015), and we find that AI assistance even slightly increases grading time without delivering strictly better downstream outcomes.

Second, we investigate whether the above result can be due to the usage of AI-written application materials, which in that recruitment year was a novel phenomenon. In our setting, there was widespread usage of AI by applicants when writing their application materials (we call those essays LLM-essays). Using a state-of-the-art LLM detection tool with low rates of false positives,[2] we find that approximately 60% of applicants use LLMs for their essays. LLM-essays are longer, more complex, and contain less specific information and applicants who predominately rely on LLMs, complete their applications faster. Humans assign significantly higher scores to those LLM-generated essays. Over time, however, as they review more applications, they appear to learn to identify AI-generated content more reliably. They become more skeptical and begin assigning lower grades to these essays, thereby halving the gap in scores between LLM- and non-LLM essays. Essentially, the evaluators begin penalizing LLM-generated content in relative terms. A similar pattern occurs for AI-assistance—evaluators are about 25% less likely to follow the AI recommendation when grading an LLM-essay. Intuitively, AI recommendations might offer stable reference points, aiding human decision-making in a noisier setting. Evaluators initially incorporate the model's suggestions, but as they see that the algorithm does not penalize AI-generated essays—and sometimes diverges from their own intuition—they lose trust in the tool. Rather than increasingly agreeing with AI over time, evaluators become less dependent on it.

Our findings contribute to multiple strands of literature. First, our work enhances the understanding of AI's impacts on labor markets, particularly in the context of recruiting and hiring workers. The ability of AI to enhance recruitment processes has led to widespread

[2]We describe in detail in Section 3 what tool we use, how we check for false positives, and how classification rates change at different thresholds and levels of aggregation.

adoption of these technologies in organizations (Vrontis et al., 2022) . Much of the existing research in recruitment has focused on the effects of AI on diversity of the hires and biases against certain populations within the applicant pool (Avery et al., 2023; Li et al., 2024; Agan et al., 2023). Our study adds to a small body of work suggesting that AI-driven candidate selection can lead to selection of higher quality candidates. For example, Cowgill (2020) finds that candidates selected by AI systems are generally better performers and more productive and Chalfin et al. (2016) demonstrates that AI can assist in the selection of less violent police officers and more effective teachers. Additionally, Wiles et al. (2023) show that using AI to construct applications leads to greater clarity of those applications and therefore enables the recruiters to extract the quality signal more easily.

Second, we also contribute to recent work documenting that generative AI can boost productivity in tasks like coding, writing, and general consulting tasks (Brynjolfsson et al., 2023; Bubeck et al., 2023; Dell'Acqua et al., 2023; Noy and Zhang, 2023; Peng et al., 2023; Kumar et al., 2023). In our setting, even a capable model does not guarantee improvements unless humans trust and effectively integrate its output, which is related to the literature showing that people do not systematically agree with AI recommendations due to algorithmic aversion (Dietvorst et al., 2015), bias against AI-generated content (Parshakov et al., 2025), priors that are far from algorithmic recommendations (Glaeser et al., 2021), or cognitive constraints (Agarwal et al., 2023).

To our knowledge, this paper is one of the few, alongside Otis et al. (2023), that analyses the capabilities, productivity, and performance effects of AI assistance powered by novel LLMs in the context of developing countries. Most other work in this context focuses on designing AI-tutors aimed at improving learning outcomes (Chen et al., 2024; De Simone et al., 2025).

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".[3] Every year, a cohort of between 50 and 150 fellows is assigned to schools in rural areas[4] and earn a stipend comparable to the average entry-level salary in Ghana.[5] The position is considered prestigious, and the candidate selection process is competitive, with only 15-20% of applicants being offered positions.[6] 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 remains 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 supply and demand sides of the recruitment process for the partner non-profit organization, as well as the design of our policy experiment.[7] 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.[8] Recruitment is cyclical and typically occurs between March and July. If a candidate accepts the offer, they begin their fellowship between October of the 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

[3]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.

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

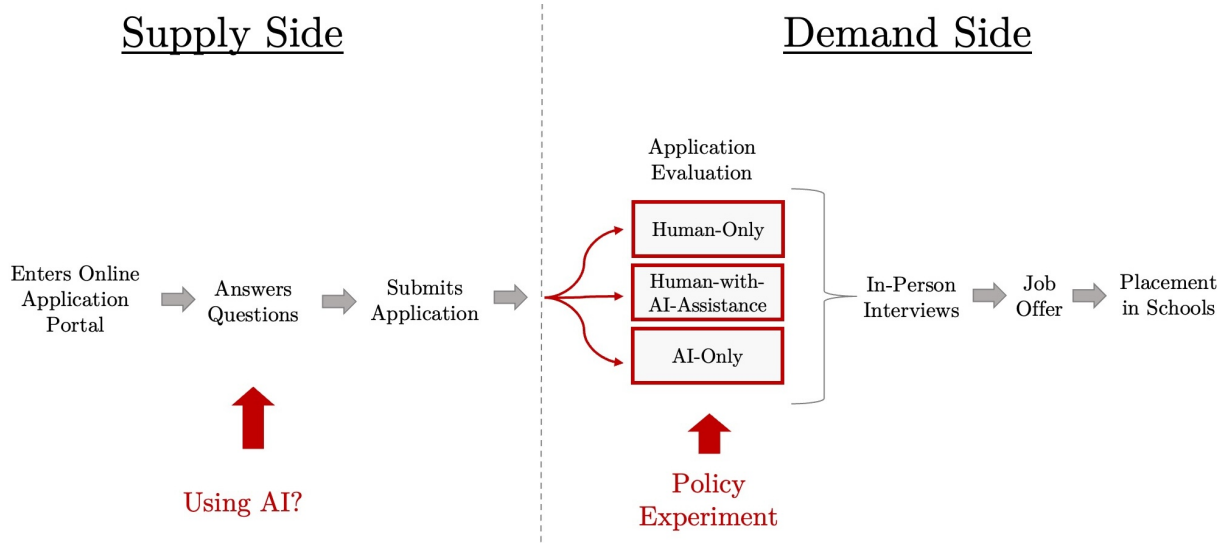
[5]The stipend received during the fellowship exceeds what the person would normally receive during National Service in the public sector

[6]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.

[7]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.

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

Figure 1: Supply and Demand Side in the Recruitment Process



Notes: The figure shows the supply and demand sides of the recruitment process for the partner non-profit organization. On the supply side, potential applicants were required to enter the online application portal, register, and answer six essay questions before submitting the application. Since ChatGPT had become widely available a few months prior, many applicants used it to assist in answering the essay questions. After 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, 189 received fellowship offers, and 129 accepted the offers.

are graded on a scale from 1 (lowest) to 5 (highest), based on clear grading criteria, unknown to the applicants. Applicants who achieve a total score of 18 or higher 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[9] do not get graded.

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 ap-

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

plicant'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 assess 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.* While applicants do not have access to these grading criteria, it is easy to see why using LLM input to answer such questions would be advantageous. LLMs can produce well-structured answers that align with common expectations for strong responses, potentially giving applicants an edge in writing more compelling essays, and saving them a substantial amount of time.

In-Person Interviews 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). 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 or higher are offered a fellowship position.

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 randomize applications to one of three policy pipelines; Human-Only, Human with AI-Assistance, and AI-Only- thereby affecting the final grade which determines if 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 Human with AI-Assistance pipeline, the grade is provided by human evaluators who receive help from the AI. For half of the applications in this group, the evaluators receive only the AI-generated grade as input, while for the other half, the AI grade is accompanied by a rationale (also generated by the AI), explaining why that particular grade was assigned to the response. This design 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 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.

Figure 2 illustrates the process the evaluators followed for grading.[10] 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. This was followed by a screen presenting a question and its answer, along with 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 a grade for a question answer, the process differed depending on the random assignment of AI assistance. Applications assigned to receive no AI assistance proceeded to the next question. However, for applications assigned to receive AI-assistance, evaluators were shown another screen after submitting their grade. On this screen, the evaluators were shown the answer and the grading criteria again, as well as the grade that the algorithm suggested. As mentioned above, to identify potential mechanisms, in half the cases, evaluators were also provided with a justification for the algorithm's recommendation.

[10]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.

At the end of that screen, evaluators were required to re-enter the grade of 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.

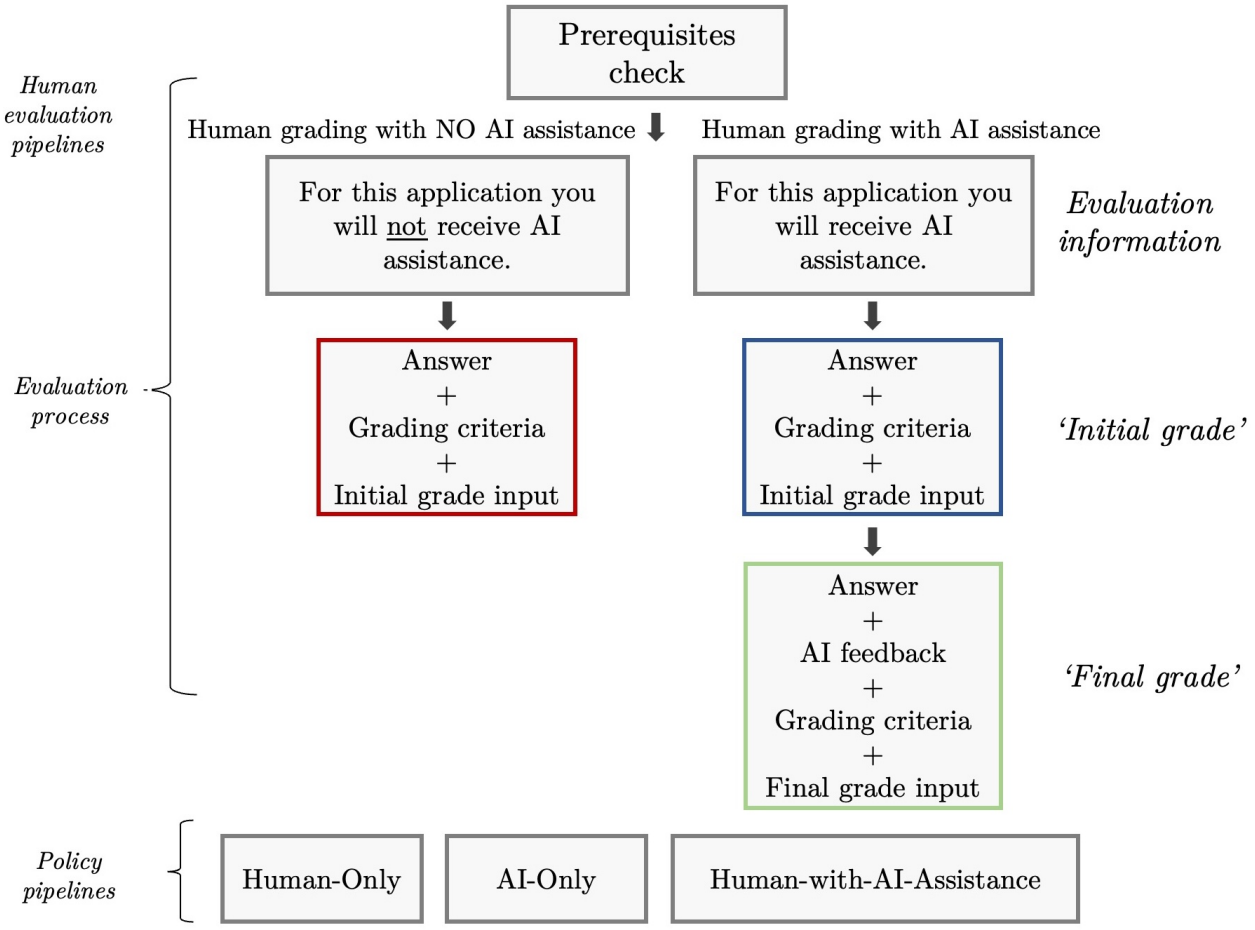
The goal of this design was threefold. First, the parallel grading—where each essay is graded by both humans (with or without AI assistance) and the AI—allows us to compare differences between human initial and AI grades, as well as between human final and AI grades, for the same set of essays. This significantly increases the precision of our analysis. Second, the double grading helps us assess how "noisy" the grading process is and provides insights into why the AI might outperform human evaluators in certain contexts. Third, by randomizing applications into different policy pipelines, we can evaluate the causal impact of each grading approach on downstream outcomes such as job offer rates and hiring.

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), utilizing a "zero-shot" approach. This means that 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 (Appendix Section C.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., 2023). 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.[11] The core of the content prompts (Appendix Section C.2) consisted of instructions from the evaluator manual, including the grading criteria for each grade (1 to 5)

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

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").

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"[12] 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 Data, Outcomes, and Empirical Strategy

3.1 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 a question answer was 2238 words (373 words for each answer), 44.9% of essay answers were generated by an LLM[13], 60.0% of the applicants have at least one LLM-generated essay, and 31.6% of the applications can be classified as being entirely LLM-generated. Due to a change in the non-profit organization's data privacy policy during the course of the

[12]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.

[13]We explain in detail how we define whether an answer, or the entire application, is LLM-generated in Section 3.3 below.

experiment, we were able to obtain detailed background information for approximately 75% of applicants. Among these applicants, 35.7% identify as female, 57.2% have completed their national service, 4.7% hold a Master's degree or higher, and 12.8% had previously applied to the program. 86.4% 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). 38.9% of applicants originally come from one of Ghana's Northern regions, 14.2% from Volta region and the remainder from Ashanti (12.3%), Greater Accra (7.6 %), and other regions in Southern and Central Ghana (27.0%).

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 effect. 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).

3.2 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. Finally, we explain how we identify and define LLM-generated essays and LLM-generated applications, and we describe their prevalence in our setting.

3.2.1 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. 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.

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; b) *algorithmic override*: a dummy variable that equals one if the human final grade is not equal to the AI grade, 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), for the subset of applications for which AI assistance was given; d) *difference between initial grade and AI grade*: the initial grade minus the AI grade; e) *difference between final grade and AI grade*: the final grade minus the AI grade, for the subset of applications for which AI assistance was given.

Grading Time We record the time that the evaluators spent grading application answers through our Qualtrics Survey platform. Grading time can approximately be interpreted as productivity - the longer the time needed to grade an answer, the lower the productivity. We use two time frames: *time up to initial grade*, which captures the time taken to assign the initial grade, and *time up to final grade*, which includes the time taken up to the final grade assignment. For question answers assessed without AI-assistance, the *time up to final grade* is equivalent to the *time up to initial grade*. However, for question answers where AI assistance was provided, the *time up to final grade* is the sum of the *time up to initial grade* and the time spent on the AI feedback page. *Time up to final grade* reflects the overall impact of AI assistance on grading time. Analyzing *time up to initial grade* allows us to investigate potential anticipation effects from receiving AI assistance, as evaluators were informed beforehand about whether they will receive such assistance.

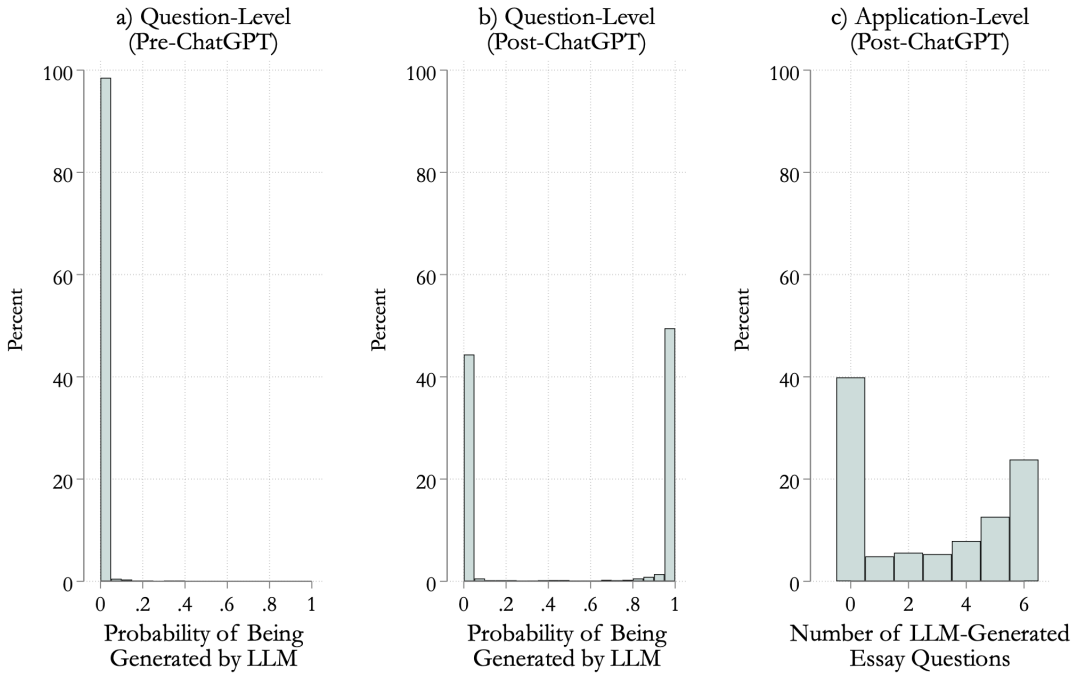
3.2.2 Downstream Outcomes (Application-Level Outcome Variables)

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 question grades with equal weight given to each. 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. For applicants that were advanced to the in-person assessment stage of the application process in each of our pipelines (i.e. were awarded a total application grade of at least 18 points), we observe additional downstream outcomes, and create the following variables; a) *attended assessment center*: a dummy variable 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 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.

3.3 Identifying AI-Generated Essays

Figure 3: 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.

To detect AI-generated content, we use a transformer-based neural network called Pangram Text, developed by Pangram Labs (Emi and Spero, 2024). This tool has very low overall error rates and low false positives rates. When analyzing a document, the software estimates the probability that the text was AI-generated and identifies the likely model used

(e.g., GPT-4, GPT-3.5, Gemini, etc.). If the probability that an essay is written by an LLM is 0.99 or higher, we classify the essay as LLM-generated. We also test the robustness of our results by adjusting the probability cutoff (e.g., using 0.9 instead of 0.99), and our results remain qualitatively unchanged.

To classify whether an entire application is LLM-generated, we calculate the average probability of all answers being AI-generated at the question level for each application. If this average probability is 0.99 or higher, we classify the entire application as LLM-generated. We validated the model by testing Pangram Text on application essays submitted before ChatGPT's commercial release (for the previous application cycle in Spring 2022), where we expect the ground truth for LLM-generation to be 0%. Approximately 96% of these pre-ChatGPT essays had an estimated probability of being LLM-generated below 0.01, and none had a probability above 0.44. Using the 0.99 likelihood cutoff for classification, this results in a false positive rate of 0%.

Figure 3 shows the distribution of estimated probabilities for essay answers to be LLM-generated, for essays in the pre-ChatGPT period (Panel a)) and in post-ChatGPT (Panel b)) period. Panel c)) represents the number of questions classified as LLM-generated according to our metric in each application. LLM-generated essays are very common in our experimental setting. Using the method described above, we classify approximately 45% of essay questions as LLM-generated (Figure 3, panel b))[14]. Additionally, 60% of applications have at least one LLM-generated essay, and about 31.6% of applications are classified as fully LLM-generated according to our method.

3.4 Characteristics of LLM-Generated Answers

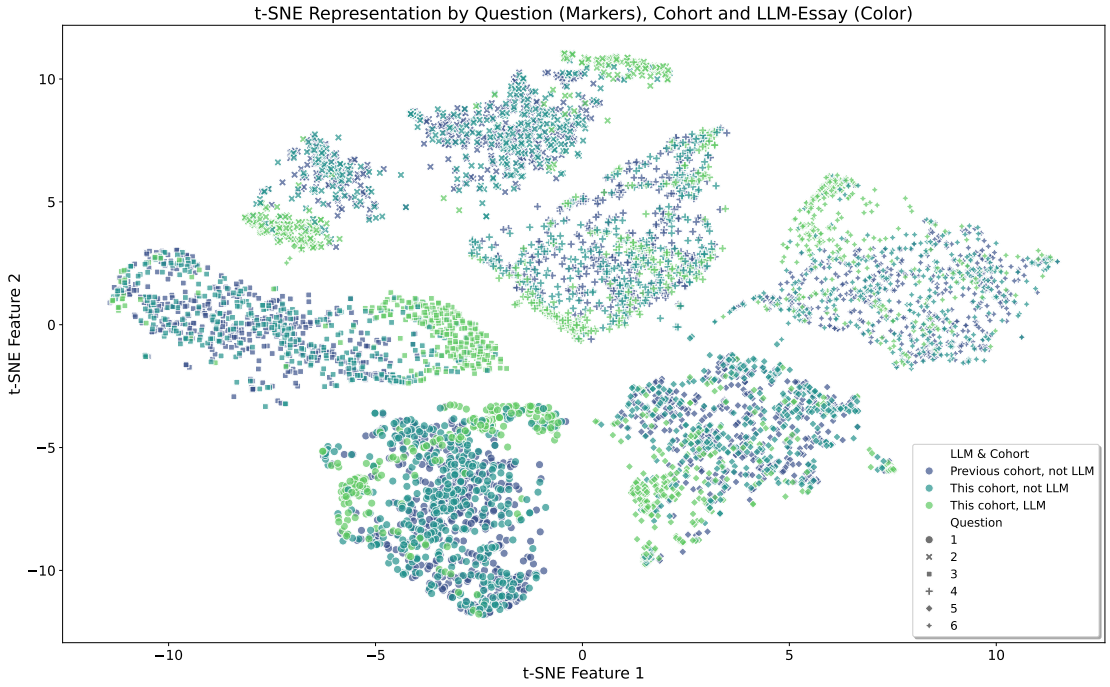
LLM-generated essays are 55% less likely to include specific information, such as details about the applicant's university or gender, they are 40 words (11%) longer, and have lower readability scores (Flesch, 1948)[15]. (see Appendix Figure A.8). Additionally, LLM-generated essays

[14]This percentage is based on a 0.99 cutoff; for a 0.9 cutoff, the corresponding percentage is 50%

[15]Readability reflects the ease with which a reader comprehends written text; higher readability scores indicate less effort required for the reader. We use 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's Word text 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.

are semantically distinct from non-LLM essays. Figure 4, which visualizes high-dimensional essay embeddings in two dimensions, shows that LLM-generated essays (light green) occupy different regions of the semantic space compared to non-LLM essays (dark green) and form smaller, more compact clusters. This semantic distinction is further supported by our analysis of the principal components of the embedding vectors, which reveals that the distributions of LLM and non-LLM essays are significantly different (see Appendix Figure A.9.

Figure 4: 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-lite-02-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.

What Predicts LLM Usage? In our study, the use of LLMs to produce application materials was not randomly assigned; it is a choice made by the candidates themselves. Appendix Figure A.10 shows that the strongest predictors of LLM usage are whether the person applied to the fellowship before, whether they had a personal referral, whether they completed the national service (all negative predictors), as well as whether the person submitted

their application late (in July), and had a low GPA (between 1.0-2.0 out of 4) (both positive predictors). However, we should interpret the low GPA predictor with caution. Only 9 people (1.75% of the sample with available demographics) fall into this category, and for all but one of them the application was classified as LLM-generated. It is reasonable to assume that candidates who applied to fellowship before used LLMs less frequently, as they likely reused essays from previous applications. Similarly, applicants who learned about fellowship through personal connections (e.g., campus events or word-of-mouth) may differ from those who discovered the program through social or traditional media in their likelihood of knowing that LLMs can help them write their applications.

3.5 Empirical Strategy

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

$$y_i = \alpha + \beta_1 AIOnly_i + \beta_2 AIAssistance_i + X_i'\lambda + \gamma_i + \epsilon_i \quad (1)$$

where $AIOnly_i$ and $AIAssistance_i$ are indicator variables equal to one if the application was assigned to AI-Only and Human-with-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.[16]

Additional Analysis We perform a number of additional analyses on our question-level data, the details of which are mentioned in Section 4 below.

4 Main Results

This section presents the overall effects of incorporating an AI algorithm into the organization's recruitment process. We begin with a question-level analysis and describe how AI grades differ from human grades. In Section 4.1, we show that there is substantial disagreement between human initial grades (that is, grades assigned prior to obtaining algorithmic feedback) and AI grades. The two types of grades match in only about a third of the cases, and AI grades are consistently higher on average. In Section 4.2, we proceed with

[16] 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.

the analysis of our policy pipelines– we compare the downstream outcomes of applicants in Human-Only and AI-Only pipelines. We find that the applicants in the AI-Only pipeline have substantially better downstream outcomes; for example, they are 84% more likely to receive an offer and 73% more likely to be hired.

After establishing that our algorithm does an excellent job in selecting applicants who eventually receive an offer, we investigate what happens when the same AI algorithm is provided to evaluators as an assistant in Section 4.3. We document that algorithmic overriding is common— evaluators override the algorithm in over 80% of the cases where their initial grade is different from the grade provided by the algorithm. Lastly, we find that the Human-with-AI-Assistance pipeline did not result in higher job-matching rates than the Human-Only baseline.

4.1 Initial Human and AI Grades

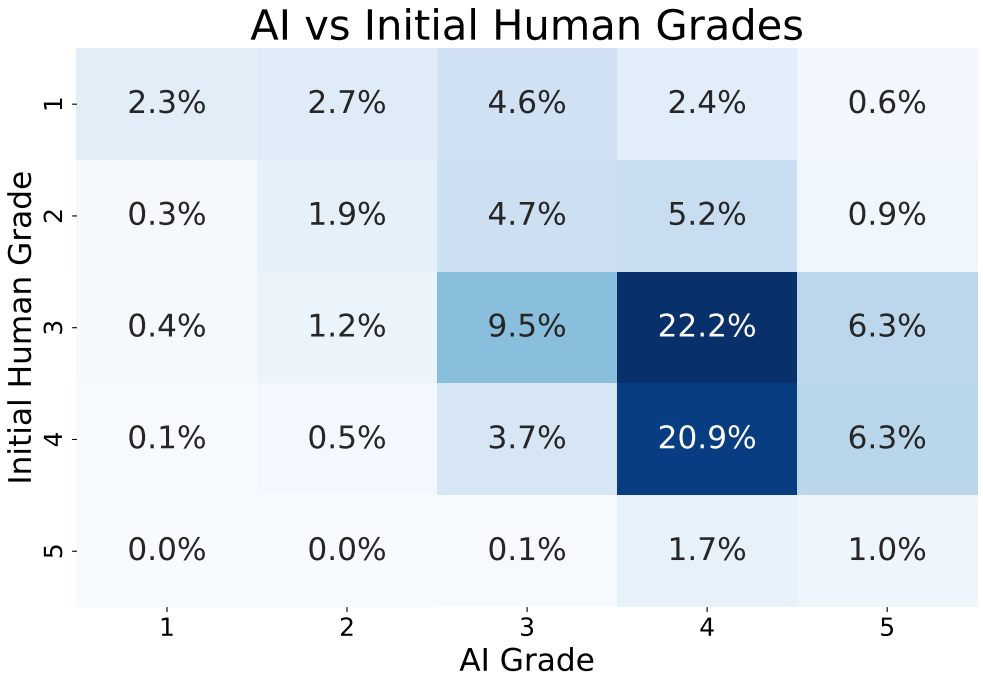
In this section we document the agreement rates between Human initial and AI grades, as well as the agreement rates between initial human grades for essays that were independently graded twice by different evaluators.

AI-Only vs. Human Initial Grades Figure 5 shows a 5x5 matrix that depicts the distribution of grades, each cell representing agreement frequencies between *initial human* and *AI* grades for each individual essay answer (note that there are 6 essays per application). The diagonal (top-left to bottom-right) indicates agreement between grades. Areas below (above) the diagonal represent cases where the initial human grade is higher (lower) than the AI grade. We observe that there is about 36 % agreement in the AI and human initial grades. In the majority of all cases (56%), the AI grades are higher than the human grades, while only in about 8% of cases are human grades larger than the AI grades (conditioning on disagreement, these numbers are 87% and 13 %, respectively). The average difference between human and AI grades is -0.71 (-1.1 conditional on there being a disagreement)[17], which is substantial, given that the average initial human grade is around 2.9. Overall, this suggests that the AI seems to be a lot more generous when grading the essays although there is substantial heterogeneity across questions (Appendix Figure A.1), with agreement rates ranging from 26% (question 4, core beliefs) to 47% (question 6, influencing and motivating others).

Agreement in Human Grades To be able to 'judge' whether these disagreement rates

[17]The absolute differences are 0.9 and 1.4, respectively.

Figure 5: Initial human grades vs. AI grades



	Number
Agreement rate (initial human & AI grades)	36 %
Grade difference (overall)	-0.71
Grade difference (if diff. >0)	-1.1

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).

are large or small, and to investigate whether a task is straightforward or not, we had a subset of applications graded by two independent evaluators. Appendix Figure A.2 shows a 5x5 matrix that depicts the distribution of grades, each cell representing agreement percentages between initial human grades and AI grades for answers from applications that were graded twice. Interestingly, the grades across the two grading rounds are the same only in 44% of the cases, indicating that the task is not straightforward and it is difficult for the two distinct graders to agree on the grade. There is also some heterogeneity in agreement across questions (Appendix figure A.3), with agreement rates ranging from 36% (question 1-why

do you want to become a fellow) to 53% (question 6-influencing and motivating others). We further benchmark these agreement rates by comparing them to agreement rates both within the same LLM model and across different LLM models, including the one used in our experiment (GPT-4 (gpt-4-0314))—see Appendix Table A.4. While disagreement rates on grades across models are relatively high (the highest being around 58% between GPT4o and Gemini), disagreement rates for repeated grading by the same model are much lower than for humans. For the model we used, disagreement rates are around 20%, compared to 56% for humans.

4.2 AI-Only Policy Pipeline

Table 1 summarizes the applicants' progression through the selection process under different recruitment pipelines: "Human-Only", "AI-Only" and "Human-with-AI-Assistance". It reports estimated coefficients from OLS regressions for application grading outcomes (Panel A) and downstream outcomes (Panel B). Odd columns contain only stratum (week) fixed effects, and the even columns add demographic control variables.[18] The specification with the control variables (columns (2) and (4)) shows that compared to applicants assigned to "Human-Only" pipeline, applicants assigned to "AI-Only" pipeline receive 4.2 (24%) points more on average for their application and are 29.5 p.p. (50%) more likely to achieve the cut-off grade of 18 and be invited to the in-person interview phase. This is consistent with the fact that the AI tends to award more generous grades. In Panel B, the specification with the control variables (columns (2), (4) and (6)) shows that applicants in the AI-Only pipeline are also 18.4 p.p. (65%) more likely to attend the assessment center, 17.4 p.p. (84%) more likely to receive an offer and 10.9 p.p. (73%) more likely to accept the offer than the applicants in the Human-Only baseline.

Why do candidates in the AI-Only pipeline end up performing so much better than candidates in the Human-Only baseline? One possible explanation is that, because the AI advances a much larger number of candidates, it minimizes the likelihood of screening out candidates of high quality who are capable of receiving an offer (analogous to minimizing Type II error). To shed light on this, Table 2 shows the likelihood of an applicant receiving an offer based on their ranking in each pipeline. Specifically, it reports the probabilities for applicants ranked in the top–50 (column 1), top–30 (column 2), and top–10 (column 3), using their application grades as the basis for ranking. We can see that, even when the

[18] As mentioned in Section 3 above, we include only a subset of control variables in our regressions, as we do not have demographic controls for everyone.

Table 1: Application-level and Downstream Outcomes for Policy Pipelines

Panel A: Grading outcomes

	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)
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				
AI=AI-Assistance	0.000	0.000	0.000	0.000

Panel B: 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)
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						
AI=AI-Assistance	0.001	0.003	0.002	0.004	0.013	0.024

Notes: Panel A: Columns 1-4 report estimated coefficients from OLS regressions respectively 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)). Panel B: Columns 1-6 report estimated coefficients from OLS regressions respectively of an indicator variable 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, that is accepted the job offer (Columns (5) and (6). Note that the variables in columns 3-8 are unconditional, meaning that they take a value of zero if the person has not reached that stage. In both Panels, all columns include stratum (week) fixed effects, in Panels A and B the even columns additionally include controls for 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

number of candidates advanced to the interview stage is held constant, candidates in the AI-Only are still significantly more likely (55%, 81%, and 113% for top–50, top–30, and top–10, respectively) to receive an offer.[19] This is a strong indication that there is more to the AI screening than simply advancing more candidates to the interview stage. In fact, when we examine the raw correlations between application grades and interview day grades, and determine which grades (human initial or AI grades) predict interview grades better for candidates who attended the interview, AI grades exhibit stronger correlations and have 46% to 114% larger coefficients than human initial grades (see Appendix Figure A.4 and Table A.6). This implies that the AI grades are more informative of candidate quality than initial human grades. We investigate this further by delving into a concept of "semantic signal". What we call "semantic signal in grade" is a simple information criterion based on how semantically similar the essay answers are within and across grades that were assigned to them. The reasoning is simple: if grades are informative (i.e., if there is signal contained within a certain grade), one would expect question answers within the same grade to be more semantically similar than question answers across different grades. We indeed find that according to this method, AI grades contain substantially more signal than human initial grades. Our Appendix Section B provides a detailed explanation of this concept.

4.3 AI-Assistance

Having established that our AI grader performs remarkably well in this setting, we investigate what happens when evaluators receive AI assistance—-that is, when they are shown the grade recommended by the AI for the essay answer.

Usage of the AI-Assistant and Algorithmic Override We define AI-Assistance usage as any grade revision that occurs after receiving the algorithmic recommendation, even if the revision is only partial (i.e the grade is not revised fully up to or down to the AI grade), among the subset of cases where initial human and AI grades disagree. Similarly, we define as algorithmic override any case where the final human grade is not equal to the algorithmic recommendation. Figure 6 depicts the proportion of times the evaluators override the algorithm (Panel a), revise their initial grade (Panel b), and the amounts they revise for (Panel c), categorized by initial grade disagreement. Algorithmic override is common–when the initial human grade differ from the AI grade, evaluators override the recommendation 80.6 % of the time. They override the recommendation more often when the AI grade is above

[19]Note that within each pipeline, many candidates received the same grade, i.e. many people share the ranks, so there are significantly more people in the sample than just n in each pipeline.

Table 2: Offers Given to Top Candidates

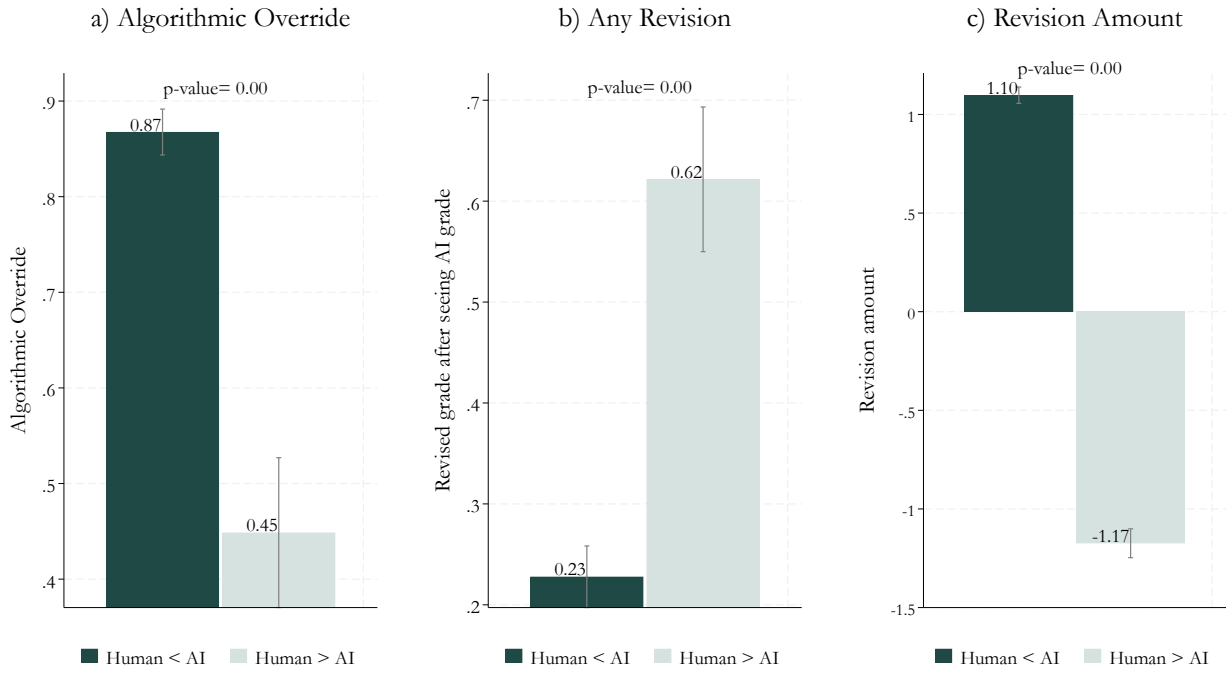
	Offer Received		
	(1)	(2)	(3)
AI-Only	0.174**	0.254**	0.375**
	(0.083)	(0.114)	(0.164)
AI-Assistance	0.082	0.114	0.281
	(0.079)	(0.110)	(0.170)
Mean (Human-Only)	0.319	0.312	0.333
Sample	Top-50	Top-30	Top-10
Stratum FE	Yes	Yes	Yes
Controls	No	No	No
N	221	125	63
p-values			
(AI=AI-Assistance)	0.258	0.186	0.504

Notes: Table reports estimated coefficients respectively from OLS regressions of the indicator variable for wheter the applicant received an offer for top-n candidates based on application scores from each pipeline: top-50 (Column 1), top-30 (column 2) or top-10 (column 3). Nothe that the offer rate here is unconditional–meaning in this case, that if the candidate did not attend the in-person interview, the variable will take a value of zero. Standard errors are clustered at the application level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

their initial grade, than when it is below (in 86.8 % and 44.9 % of the cases, respectively). Evaluators revise their grade in approximately 28.5 % of the cases if the grade provided by the AI assistant does not match their own grade, but they often do not adjust the grade all the way to the AI grade. Evaluators are significantly more likely to adjust their way down, than up. Specifically, evaluators revise in approximately 23% percent of the cases when their initial grade is below the AI grade, and in approximately 62% of the cases if their initial grade is above the AI grade.[20] AI assistance therefore raises the agreement rate from 36% to about 47.7%, an increase of about 33% (depicted in Appendix Figures A.5, A.6, and A.7). When revising, most evaluators adjust for approximately one point, which explains why the revision rate of 28.5% translates into only 11.7 p.p increase in the agreement rate.

[20]There is a negligible number of revisions where human and AI grades are in agreement–a total of 10 cases representing 0.67% of the sample where AI assistance was given.

Figure 6: 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 amounts they revise for (Panel c), categorized by initial human and AI grade disagreement; 95% confidence intervals; *p-values* calculated from a t-test for equality of means.

4.4 Human-with-AI-Assistance Policy Pipeline

Table 1 Panel A, shows that applicants assigned to Human-with-AI-Assistance pipeline, receive on average a 0.736 (54% column 2) higher total grade. However, this increase in total grade is not large enough to statistically significantly affect the advancement rate to the next stage. When it comes to downstream outcomes, Table 1 Panel B shows that applicants in Human-with-AI-Assistance pipeline do no have a statistically significantly higher likelihood of receiving the offer or being hired compared to applicants in the Human-Only baseline. Moreover, when we look at how application grades correlate with grades in the in-person assessment, we can see that the correlation is somewhere in between the AI-Only and Human-Only correlations (Appendix Figure A.4). Moreover, our results from Section B on "signal" in grade also seem to indicate that the amount of signal in human grades in the Human-with-AI-Assistance pipeline falls between the Human-Only and AI-Only pipelines.

4.5 Grading Time and Productivity

We next turn to analyzing the effects of AI assistance on grading time, which we use as a proxy for productivity. We look both at the time taken up to the initial grade (that is time needed to read the essay question and enter the initial grade), as well as time taken up to the final grade (the total time taken on a question answer that includes initial grading, time spent on AI feedback page, and entering the final grade). Appendix table A.7 presents the regression results. The first result is that essay answers for which evaluators receive AI assistance take 13-17% longer to be graded (Columns (4) and (5)), indicating, if anything, lower productivity. This is driven by cases where there is a disagreement between human and AI grades, which increases grading time by 26% (Column (6)). This effect is consistent with evaluators partially re-reading the essay and re-evaluating their own grade when it does not match the AI grade, suggesting that AI assistance actually reduces productivity (as we know it does not bring any apparent benefits to downstream outcomes), which is in contrast to what most of the recent work on AI-assistance suggests (e.g. Noy and Zhang, 2023).

Looking at time up to initial grade, Columns (1) - (3) of Appendix Table A.7 show that there seems to be an anticipation effect of AI assistance. Evaluators get told whether they will receive assistance as soon as they open the application file and if the application they are reviewing is randomly assigned to be receiving AI assistance, they spend about 10% less time initially reading the question answers. However, this initial gain in time is not enough to compensate for the extra time the evaluators spend on the AI page, since the total effect on time spent grading is positive.

5 The Role of LLM-Generated Essays in Explaining Our Results

The results of our policy experiment, presented in Section 4, reveal that candidates in the AI-Only pipeline are significantly more likely to receive a job offer and be hired than those in the Human-Only pipeline — and, perhaps surprisingly, also those in the Human-with-AI-Assistance pipeline. The worse performance of the pipeline where AI assistance is used for grading occurs because people frequently override algorithmic recommendations when the AI is used as an assistant. In this section, we provide evidence of the central role that LLM-generated[21] essays play in explaining these findings.

[21] We will use LLM-generated and AI-generated interchangeably.

5.1 How are LLM-Essays Graded?

In this section, we use data on essay grading, taking advantage of the fact that all essays were graded in parallel, with both human grades (without assistance) and AI grades available for each essay.

Table 3: 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 initial grades and AI grades (Columns (1) and (2)) and an indicator variable for whether the human initial 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 3.)

How do Human Evaluators, Absent Algorithmic Assistance, Respond to LLM-Generated Essays? In this section we show that while both human graders and the AI award a grade premium to LLM-essays, relative to the AI grade (10% and 13% respectively),[22] humans award a smaller premium LLM-essays. Table 3 shows that humans give a smaller premium compared to AI, that is humans tend to discount LLM-essays relative to the algorithm. Specifically, the gap between human and AI grades is about 25% higher for LLM-essays (Column (2)), and humans are about 4.5 percentage points (12%, Column (4)) less likely to agree with the AI grade for LLM essays than for non-LLM essays. Overall, these results suggest that applicants benefit from using LLM-generated application materials, as such materials receive higher grades—whether graded by AI or humans—and, consequently, increase the likelihood of being invited to an interview.

[22]See Appendix Table A.8 for regression results, and Appendix Figure A.11 for the full disagreement matrix in initial and AI grades for LLM- vs. non-LLM essays.

Why does this difference between human and AI grades for LLM-generated versus non-LLM essays occur? If we assume that the algorithm is "unbiased" toward LLM essays because it strictly adheres to the grading criteria, then the higher grades for LLM-essays likely reflect their superior quality according to those predefined standards. This does not necessarily imply that the candidate is of higher quality; rather, it may simply indicate that LLM-generated essays are clearer or better structured (Wiles et al., 2023). In contrast, if human graders assign relatively lower grades to LLM-generated essays, it suggests they may hold negative perceptions about candidates who use LLMs. For instance, they might view such candidates as lazy, low-effort, disinterested in the position, or even dishonest about their skills. Interestingly, this human bias against LLM essays does not remain constant but evolves over time. Initially, humans award a similar premium to LLM essays as the algorithm does, but as grading progresses, they gradually reduce this premium. By the final third of the graded applications, humans assign overall grades that are 35% lower for LLM essays compared to those assigned by AI.

Table 4: 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***	0.080***	0.091***	0.071***	-0.084***	-0.057**	-0.184***	-0.171***
	(0.023)	(0.023)	(0.022)	(0.022)	(0.026)	(0.025)	(0.040)	(0.041)
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 1-6 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 difffer (Columns (6) and (5)) 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 (what we call human fnitial grade in Section 3.)

How do Human Evaluators, with Algorithmic Assistance, Respond to LLM-Generated Essays? Table 4 shows that when grading LLM-essays with algorithmic assistance, humans tend to override the algorithm 16% more often (Column (2)), they are 18% less likely to make any revisions (Column (6)), and the differences between final human and AI grades is about 31% higher (Column (8)). Similar to when grading without

AI assistance, evaluators initially override the algorithm equally for both LLM- and non-LLM essays. However, over time, they start overriding the algorithm more, especially for LLM-generated essays (see Appendix Table A.9). Our experimental design allows us to examine how the differences in algorithm-overriding rates between LLM- and non-LLM essays vary depending on whether evaluators received a justification for the grade suggested by the AI assistant. The results, presented in Appendix Figure A.12, show that the significant differences in algorithmic overriding and revision rates are primarily driven by applications assigned to the AI-Grade-with-Rationale treatment group. When evaluators are provided with a justification for the AI grade, they tend to follow algorithmic recommendations more frequently—but only for non-LLM essays. 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.

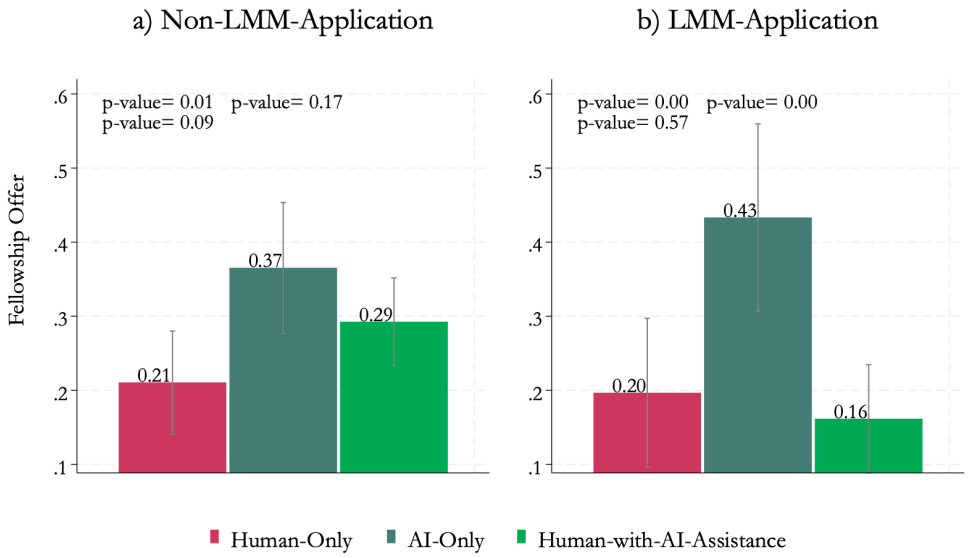
5.2 LLM-Applications and Downstream Outcomes

In Section 5.1, we demonstrated that human graders, when evaluating without algorithmic assistance, assign relatively lower grades to LLM-generated essays compared to non-LLM essays. Additionally, when using AI as an assistant, they tend to override the algorithm more frequently when grading LLM-generated essays. In this section, we investigate how our outcomes for different policy pipelines vary by whether the application was LLM-generated or not.

Figure 7 presents the main findings and Table 5 presents these results in a regression format. In the Human-Only pipeline, the likelihood of receiving a fellowship offer is nearly identical for LLM-generated and non-LLM-generated applications (20 vs. 21% respectively). However, there is a striking difference in the Human with AI-Assistance pipeline: participants with non-LLM-generated applications receive offers at significantly higher rates (13 percentage points or 80% higher) compared to those with LLM-generated applications. In fact, for non-LLM applications, we cannot reject the null hypothesis that the coefficients on AI-Only and Human with AI-Assistance pipelines are the same, while for the LLM-generated applications we can reject this hypothesis. Columns (1) and (3) replicate the findings shown in Figure 7, while Columns (2) and (4) include additional control variables. Additional columns in Table 5 further confirm this finding. The outcome variables in Columns (5)-(8) are the interaction between receiving an offer and being an LLM-application (Columns (5)-(6)) or being a non-LLM-application (Columns (7)-(8)).

These results align with the evidence presented earlier, which is that evaluators override the AI algorithm significantly more often for LLM compared to non-LLM applications. Since full automation seems to be, at least in our setting, the best option for achieving favorable downstream outcomes, not following the AI-recommendation when LLM-applications are involved causes worse downstream outcomes.

Figure 7: Offer Rates by Pipeline and LLM-Application



Notes: The figure presents the raw offer rates for our three policy pipelines–Human-Only, AI-Only, and Human-with-AI-Assistance–for Non-LLM Applications (Panel a) and LLM-Applications (Panel b). 95% confidence intervals; *p-values* calculated from a t-test for equality of means. The "firs" *p-value* refers to the difference between the red and dark green bars, the"second" *p-value* (below the first one) compares the red and light green bars, and the "third" *p-value* (next to the first one) compares the dark green and light green bars.

Table 5: 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								
AI=AI-Assistance	0.181	0.133	0.000	0.001	0.000	0.000	0.527	0.525

Notes: Panel A: Columns (1)-(5) and (7) report, respectively, estimated coefficients from OLS regressions of an indicator variable for whether the candidate received a fellowship offer (Columns (1)-(4)), and of an interaction between the indicator variable for whether the candidate received a fellowship offer and the indicator variable for whether the application was LLM-generated (Columns (5) and (7)). Columns (1) and (2) estimate the coefficients for a subsample of applications which were LLM-generated, columns (3) and (4) for the subsample which was not-LLM-generated, and columns (5) and (7) for the entire sample. Columns (6) and (8) report odds ratios from a logistic regression of an interaction between the indicator variable for whether the candidate received a fellowship offer, and the indicator variable for whether the application was LLM-generated. 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6 Conclusion

The fast development of new generative AI models that can serve as AI-assistants has shown promise in improving screening processes for hiring workers as well as helping job applicants produce polished application materials. In this paper we provide evidence that using AI tools on the demand and supply sides of a recruitment process can have interaction effects. When humans are involved in the screening process, whether they follow algorithmic recommendations depends critically on whether LLMs are used by the applicants on the supply side. We show that when human evaluators assess LLM-generated essays using AI assistance, they are more likely to override the assistant's recommendations. This occurs because evaluators likely perceive that the underlying quality of candidates who use LLMs is lower than the quality of their essays suggests. These dynamics then have significant negative consequences for job offer and hiring rates, compared to full automation.

There are, however, several limitations to our study. First, we use off-the-shelf GPT-4, so we cannot comment on what the optimal grading algorithm would be. In our setting, full automation using algorithmic grading appears to achieve outcomes closer to the optimum than human grading with or without algorithmic assistance. However, in settings where

in-person interview costs (i.e., the cost of advancing a candidate of poor quality) are higher, the results might be different. Our results align with those of Wiles et al. (2023), who find that algorithm usage on the supply side can improve labor market outcomes, and with various other studies documenting algorithmic aversion. We add to this literature by showing that algorithmic aversion is greater when LLMs are used by job applicants on the supply side. Second, while we capture some early demand-supply dynamics over time, the long-term effects remain unclear. In particular, applicants might learn that using LLMs in their applications is penalized by the demand side and adjust their behavior accordingly.

References

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." *NBER Working Papers*, https://ideas.repec.org//p/nbr/nberwo/30981.html, Number: 30981 Publisher: National Bureau of Economic Research, Inc.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." July, https://papers.ssrn.com/abstract=4505053.

Avery, Mallory, Andreas Leibbrandt, and Joseph Vecci. 2023. "Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech." February. 10.2139/ssrn.4370805.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. 2023. "Generative AI at Work." April. 10.3386/w31161.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan et al. 2023. "Sparks of Artificial General Intelligence: Early experiments with GPT-4." April. 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.

Chen, Yiling, Tao Lin, Ariel D. Procaccia, Aaditya Ramdas, and Itai Shapira. 2024. "Bias Detection Via Signaling." 10.48550/ARXIV.2405.17694.

Cowgill, Bo. 2020. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Re´sume´ Screening."

De Simone, Martin, Wuraola Mosure, Federico TIberti, Federico Manolio, Maria Barron, and Eliott Dikoru. 2025. "From chalkboards to chatbots: Transforming learning in Nigeria, one prompt at a time." January, https://blogs.worldbank.org/en/education/From-chalkboards-to-chatbots-Transforming-learning-in-Nigeria.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan R. Mollick et al. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." September. 10.2139/ssrn.4573321.

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, Place: US Publisher: American Psychological Association.

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

Flesch, Rudolph. 1948. "A new readability yardstick.." *Journal of applied psychology* 32 (3): 221, Publisher: American Psychological Association.

Glaeser, E., Andrew N. Hillis, Hyunjin Kim, and S. Kominers. 2021. "How Does Compliance Affect the Returns to Algorithms? Evidence from Boston's Restaurant Inspectors." https://www.semanticscholar.org/paper/How-Does-Compliance-Affect-the-Returns-to-Evidence-Glaeser-Hillis/52e23534071223906b9cb10d8347e88fee69e3af.

Kumar, Harsh, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. "Math Education with Large Language Models: Peril or Promise?." November. 10.2139/ssrn.4641653.

Li, Danielle, Lindsey Raymond, Peter Bergman, and UT Austin. 2024. "Hiring as Exploration."

Mortensen, Dale T., and Christopher A. Pissarides. 1994. "Job Creation and Job Destruction in the Theory of Unemployment." *The Review of Economic Studies* 61 (3): 397–415. 10.2307/2297896, Publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

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, Publisher: American Association for the Advancement of Science.

Otis, Nicholas G, Berkeley Haas, Rowan Clarke, and Rembrand Koning. 2023. "The Uneven Impact of Generative AI on Entrepreneurial Performance."

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." February. 10.48550/arXiv.2302.06590, arXiv:2302.06590 [cs].

Spence, Michael. 1973. "Job Market Signaling." *The Quarterly Journal of Economics* 87 (3): 355–374. 10.2307/1882010, Publisher: Oxford University Press.

Stiglitz, Joseph E. 1975. "The Theory of "Screening," Education, and the Distribution of Income." *The American Economic Review* 65 (3): 283–300, https://www.jstor.org/stable/1804834, Publisher: American Economic Association.

Vrontis, Demetris, Michael Christofi, Vijay Pereira, Shlomo Tarba, Anna Makrides, and Eleni Trichina. 2022. "Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review." *The International Journal of Human Resource Management* 33 (6): 1237–1266. 10.1080/09585192.2020.1871398, Publisher: Routledge _eprint: https://doi.org/10.1080/09585192.2020.1871398.

Wei, Jason, Xuezhi Wang, Dale Schuurmans et al. 2023. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." January. 10.48550/arXiv.2201.11903, arXiv:2201.11903 [cs].

Wiles, Emma, Zanele T. Munyikwa, and John J. Horton. 2023. "Algorithmic Writing Assistance on Jobseekers' Resumes Increases Hires." January. 10.3386/w30886.

"Finding Talent in the Age of AI"

Online Appendix

Kobbina Awuah
Ursa Krenk
David Yanagizawa-Drott

Table of Contents

A Figures and Tables

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 disagreement	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 inital-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 3.2.1.

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: Disagreement Rates: Model Comparisons

Percentage Disagreement in Grades

	GPT4	GPT4o	CLAUDE	GEMINI
GPT4	21.500	37.040	41.224	58.919
GPT4o	37.040	15.333	37.374	50.813
CLAUDE	41.224	37.374	5.667	49.067
GEMINI	58.919	50.813	49.067	27.000
Average Grade	3.701	3.511	3.486	3.071

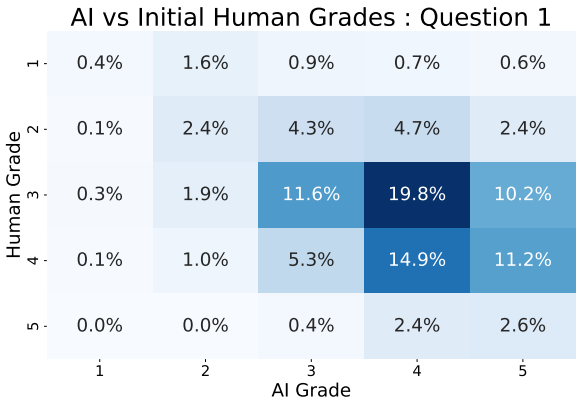
Notes: The table displays disagreement rates in grades awarded both across and within different LLMs, including GPT4 (gpt-4-0314), GPT4o (gpt-4o-2024-05-13), Claude (claude-3-5-sonnet-20240620), and Gemeni (gemini-1.5-pro-001). Disagreement across models (off-diagonal values) is represented as the share of instances where distinct grades are given. Disagreement within models (diagonal values) reflects variation in grades when rerunning the same model with different random seeds. The final row presents the average grade assigned by each model across all N=4182 essays.

Table A.5: Downstream Outcomes (conditional)

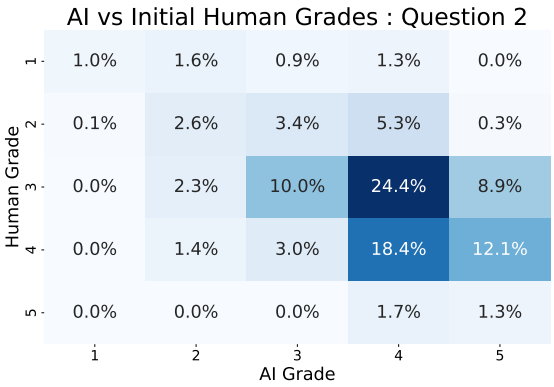
	Attended AC	Offer	Hired
	(1)	(2)	(3)
AI-only	0.0377	0.0836	-0.0411
	(0.061)	(0.075)	(0.091)
AI-Assistance	0.00881	0.0442	-0.0690
	(0.058)	(0.073)	(0.089)
Mean (Human-only)	0.478	0.727	0.725
Stratum FE	Yes	Yes	Yes
Controls	No	No	No
N	494	247	191

Notes: Columns 1-4 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 conditional on attending the assessment center (column 2) and was hired, that is accepted the job offer (column 4) conditional on receiving the offer. All columns include stratum (week) fixed effects; Note that the variables in columns 2-3 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

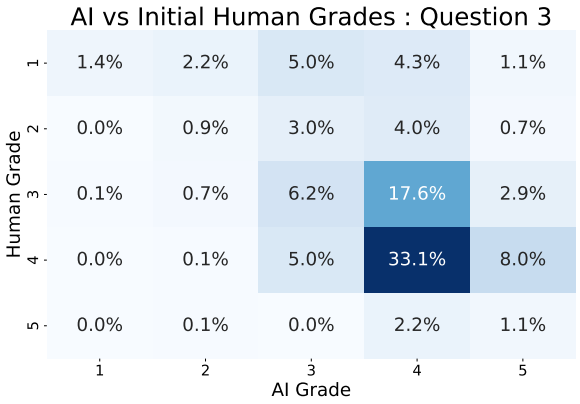
Figure A.1: Initial Human Grades vs. AI Grades by Question



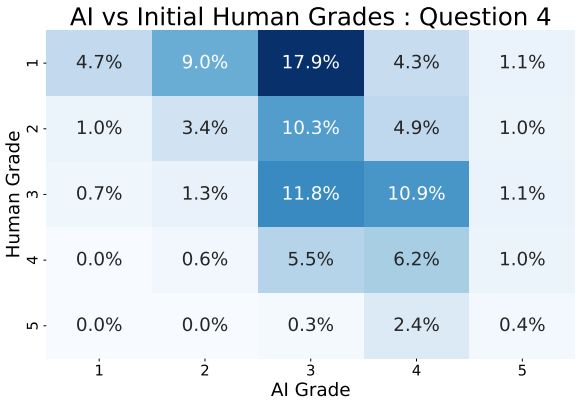
a) Question 1



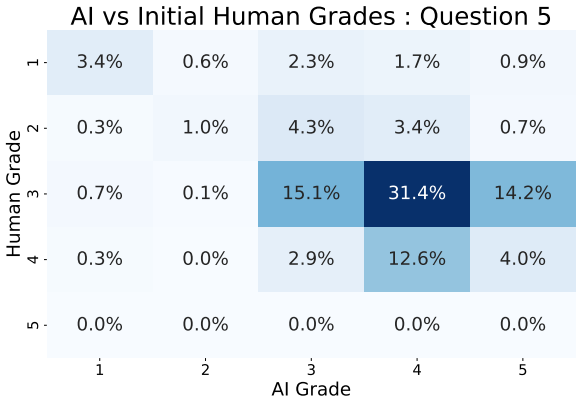
b) Question 2



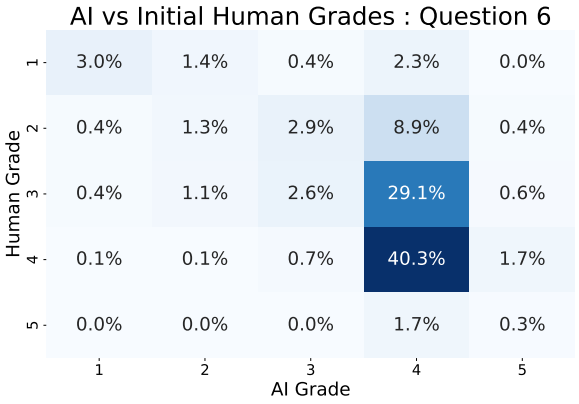
c) Question 3



d) Question 4



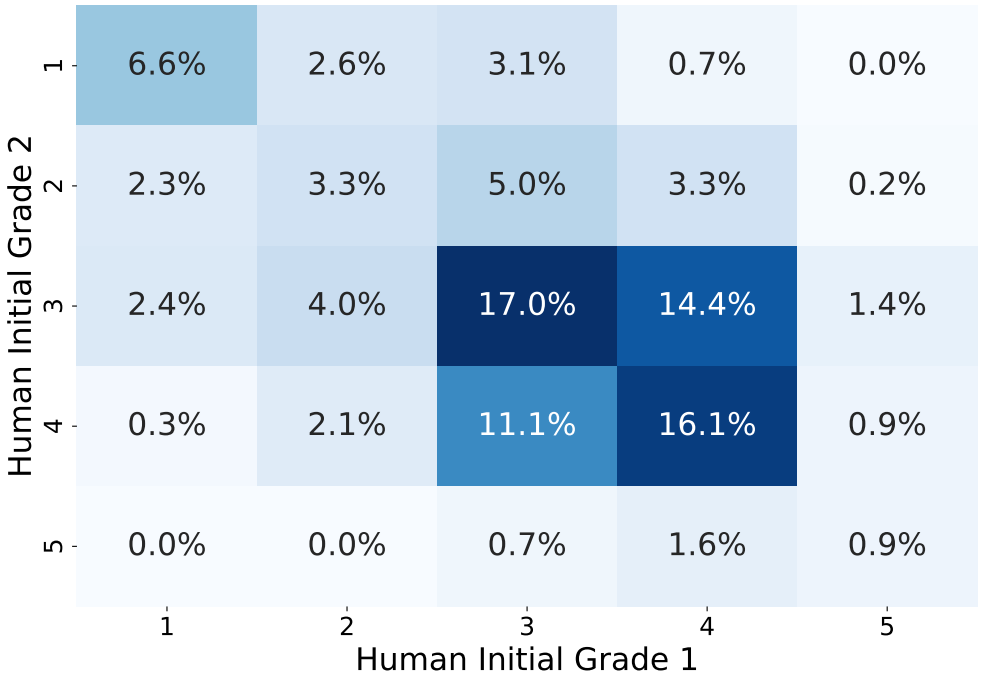
e) Question 5



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.

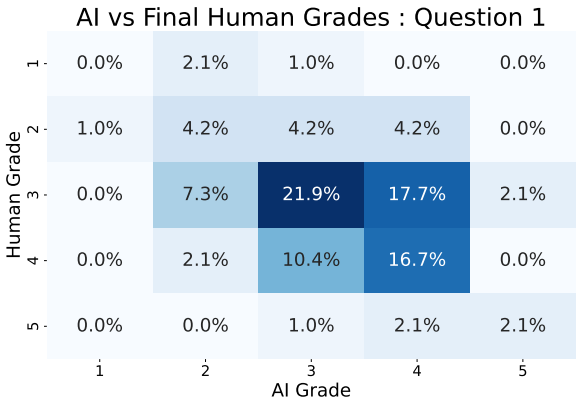
Figure A.2: Initial Human Grades Consistency



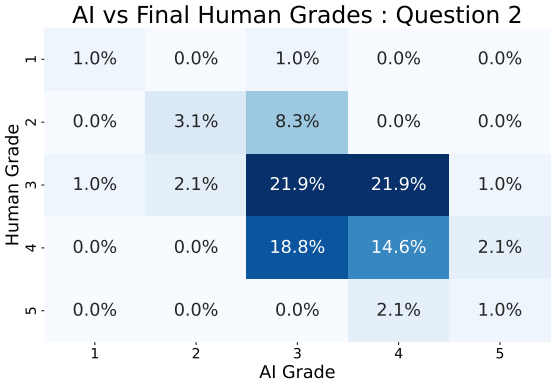
	Number	Percentage
Human grade 1!=Human grade 2	323	56
Human grade 1=Human grade 2	253	44

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.

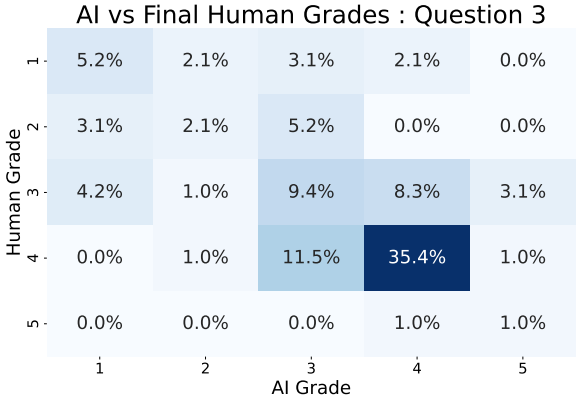
Figure A.3: Initial Human Grade Consistency by Question



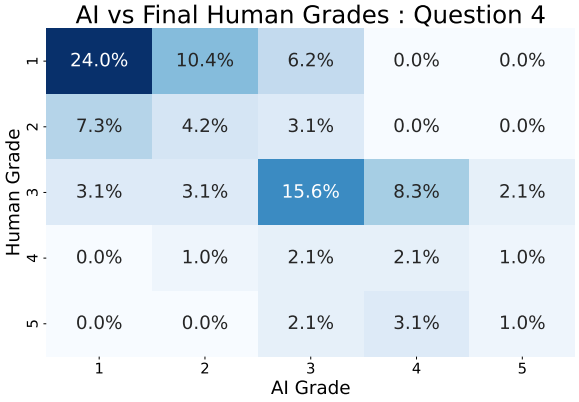
a) Question 1



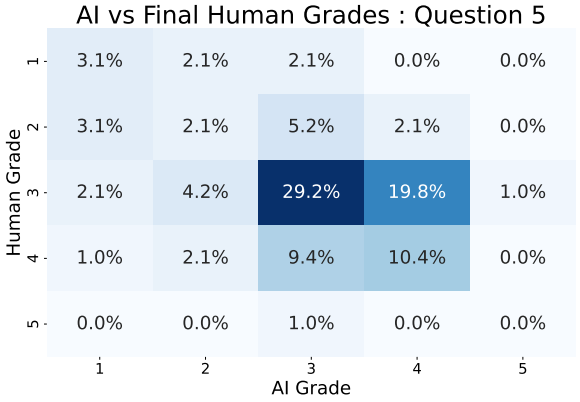
b) Question 2



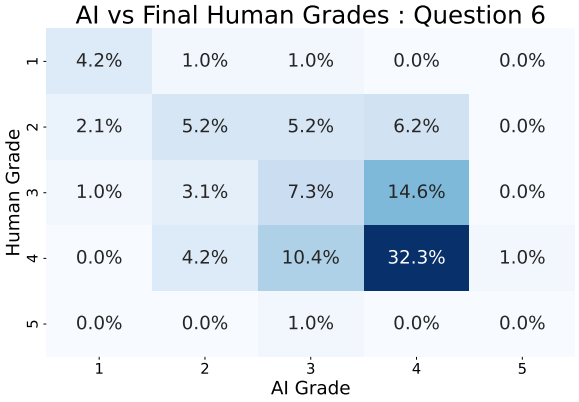
c) Question 3



d) Question 4



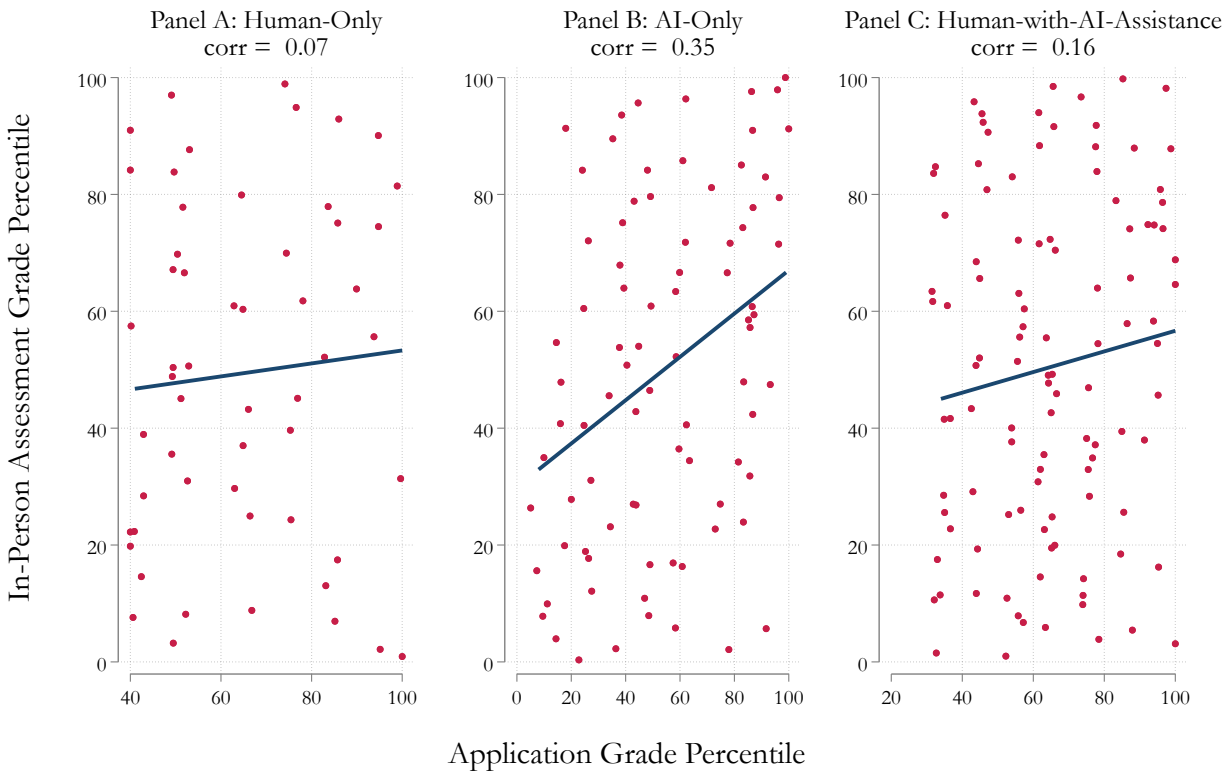
e) Question 5



f) Question 6

Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement frequencies between initial human grades (ranging from 1 to 5) for applications graded twice, separately for each question. 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.

Figure A.4: The Correlations Between Application Grades and In-Person-Assessment Grades



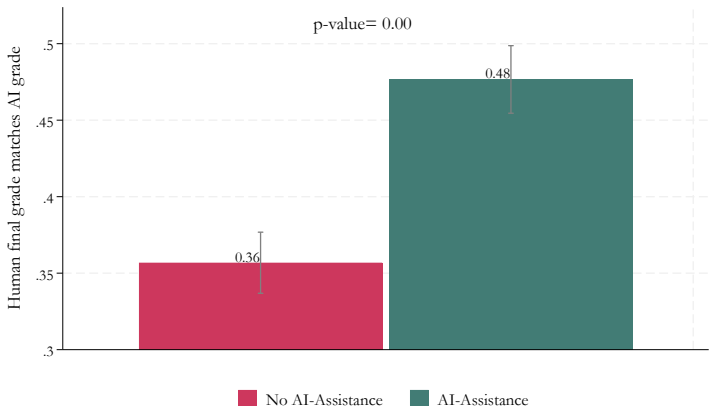
Notes: The figure shows scatter plots of pipeline-specific in-person assessment grades (percentiles) versus application grades for the three different pipelines: Human-Only, AI-Only, and Human-with-AI-Assistance. Each subplot includes a blue fitted line to indicate the correlation between in-person assessment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

Table A.6: How Do Application Grades Predict In-Person-Assessment grades?

	Total in-person assessment grade	
	(1)	(2)
Total human grade	0.363	0.647**
	(0.243)	(0.307)
Total AI grade	0.779***	0.947***
	(0.280)	(0.342)
Mean (Human-only)	55	55
Type of human grade	Initial Grade	Initial Grade
Stratum FE	Yes	Yes
Controls	No	Yes
N	247	247
p-value (Human=AI)	0.364	0.600

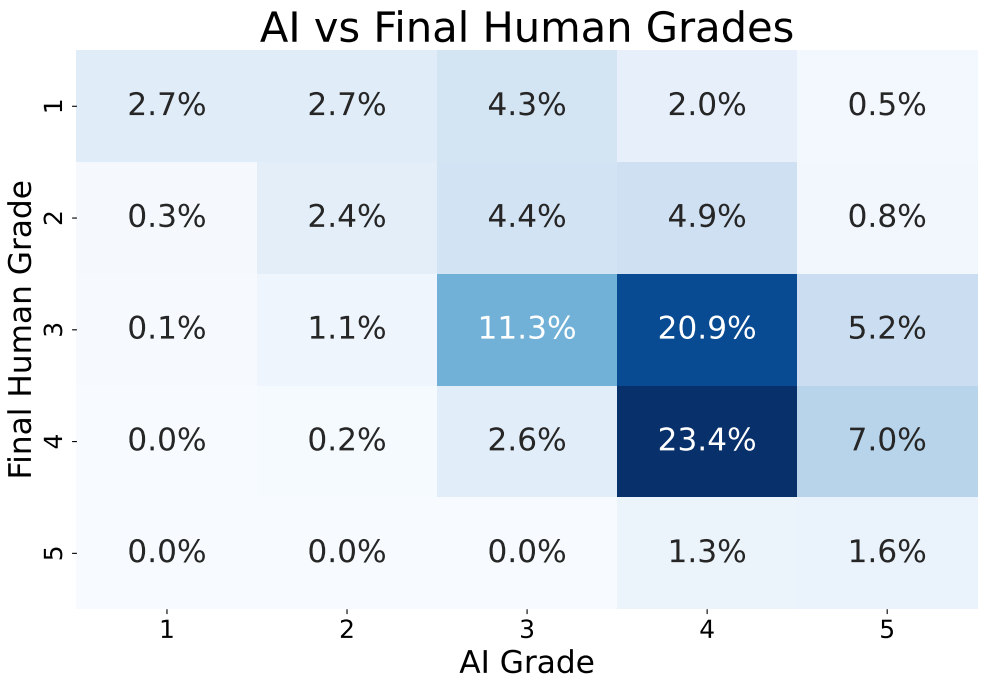
Notes: Columns 1-2 report estimated coefficients from OLS regressions of total in-person assessment grades on initial human total applications grades and the total AI grades, for people who were advanced to, and attended the in-person assessment. All columns include stratum (week) fixed effects; columns additionally includes 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. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.5: Average Agreement in Final Grades



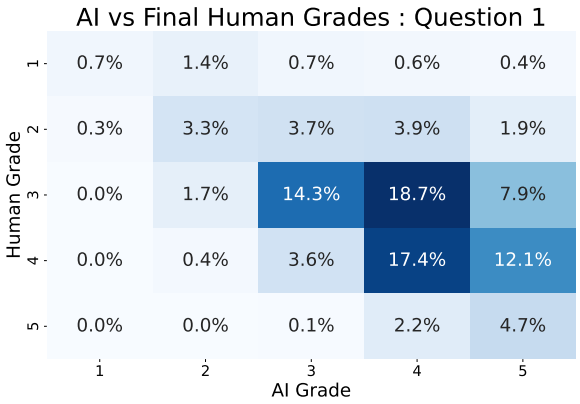
Notes: The figure shows the proportion of questions where the final human grade (after receiving AI Assistance) matched the AI grade for applications randomized into the Human-with-AI-Assistance treatment group. p-values are calculated from a t-test for equality of means.

Figure A.6: Final Human Grades vs. AI Grades

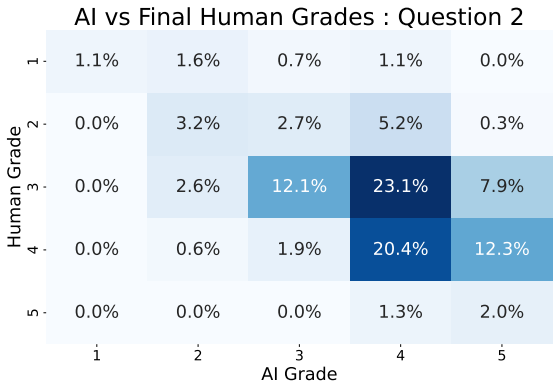


Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement frequencies between final human and AI grades (both ranging from 1 to 5). The diagonal (top-left to bottom-right) indicates complete agreement. Areas below (above) the diagonal represent cases where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

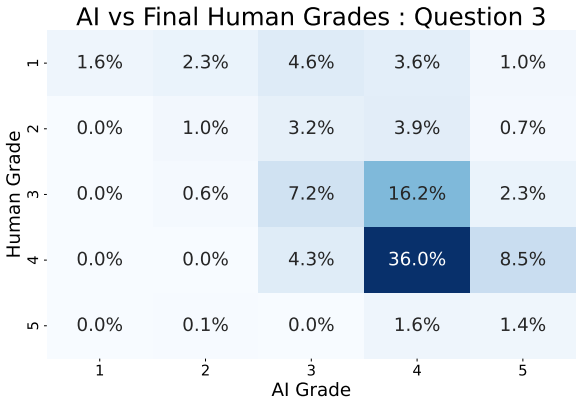
Figure A.7: Final Human Grades vs. AI Grades by Question



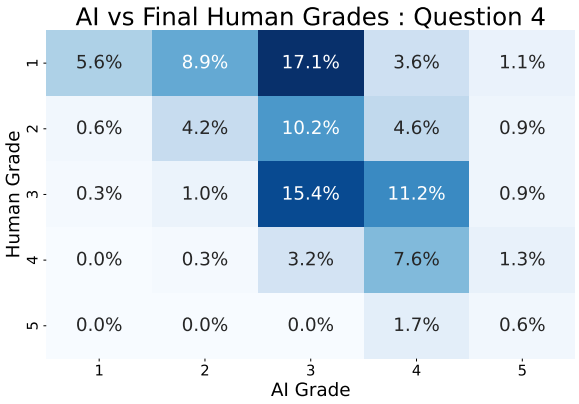
a) Question 1



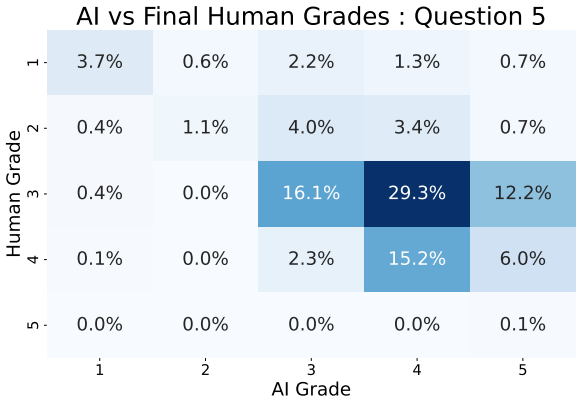
b) Question 2



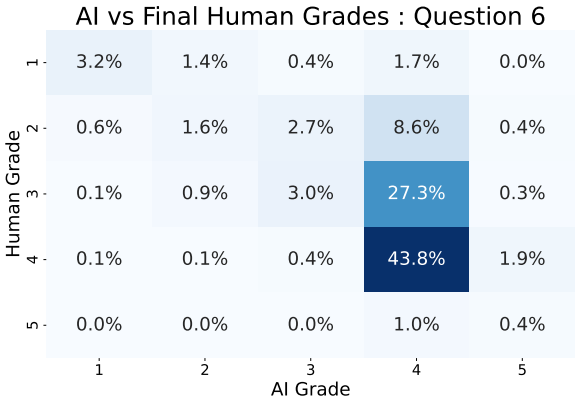
c) Question 3



d) Question 4



e) Question 5



f) Question 6

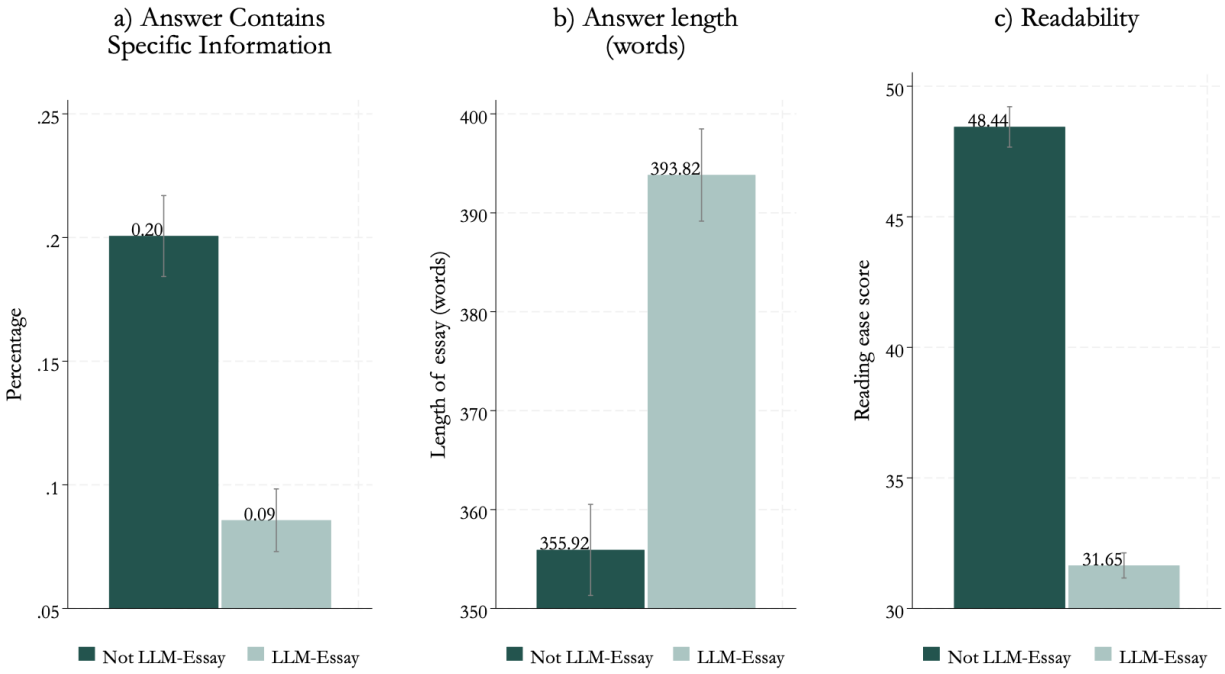
Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement frequencies between final human and AI grades (both ranging from 1 to 5), separately for each question. The diagonal (top-left to bottom-right) indicates complete agreement. Areas below (above) the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

Table A.7: Time Spent on Application

	Time to initial grade (log)			Time to final grade (log)		
	(1)	(2)	(3)	(4)	(5)	(6)
AI assistance	-0.102*	-0.146***	-0.203***	0.167***	0.127***	0.017
	(0.060)	(0.044)	(0.062)	(0.055)	(0.041)	(0.058)
Disagreement in grade			0.083*			0.087**
			(0.042)			(0.042)
Disagreement x AI-Assistance			0.090			0.173***
			(0.063)			(0.059)
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) and (4) 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 95 th percentile on question-level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$..

Figure A.8: Characteristics of AI-generated answers

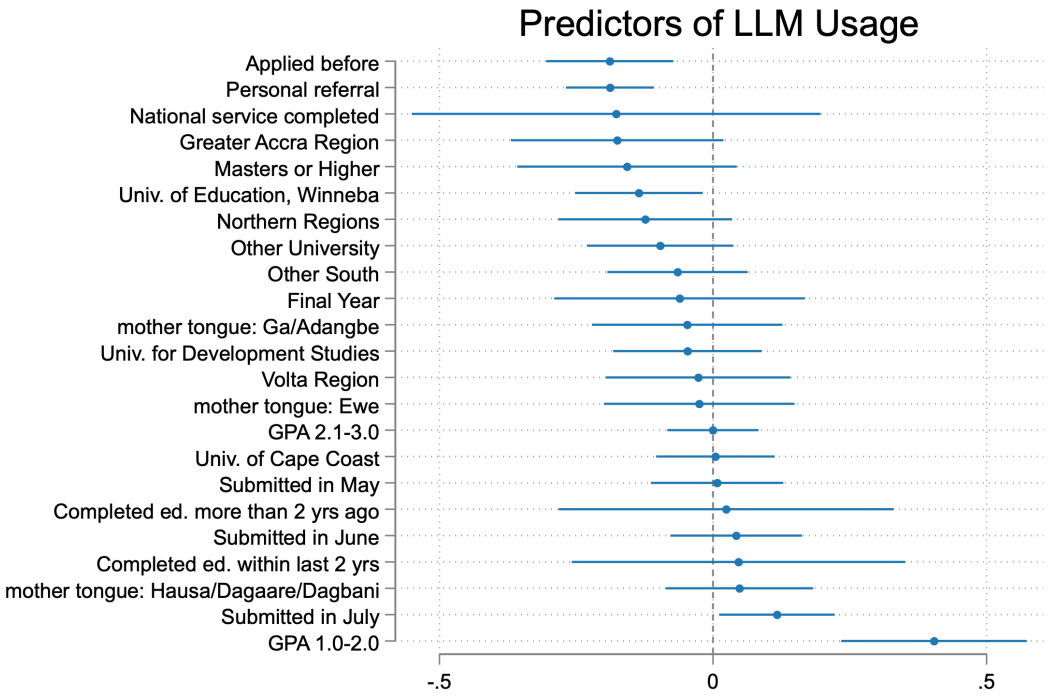


Notes: The figure depicts the characteristics of LLM- and non-LLM-essays. Panel a: Proportion of answers that contain specific information (for example on applicant's gender or university). Panel b: Answer length in words. Panel c: The complexity as measured by the Flesch reading ease (Flesch, 1948), a widely used metric that depends on sentence length and the number of syllables in words used in sentences.The exact formula is:The exact formula is: Reading Ease $= 206.835 - 1.015\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) - 84.6(\frac{\text{Total Syllables}}{\text{Total Words}})$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.9: Is Semantic Content Different Across LLM and Non-LMM answers?

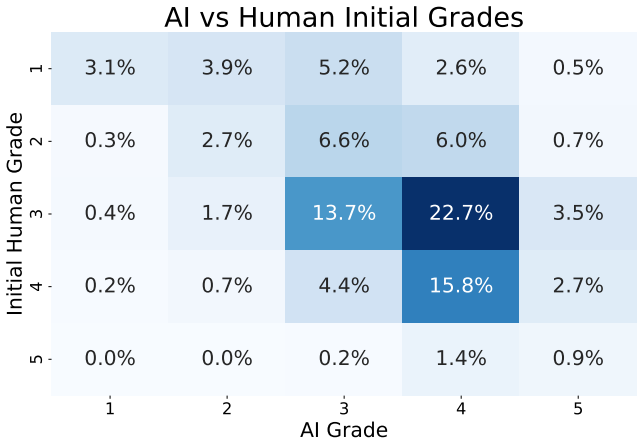
Notes: The figure depicts the distribution of first four principal components (out of 10 that were generated) of the vector embeddings that were generated using "voyage-lite-02-instruct" model from Voyage AI for LLM- and non-LLM-essays, and the Test statistic and the p-value of the Komolgorov-Smirnov test for equality of distributions.

Figure A.10: What Predicts LLM Usage?



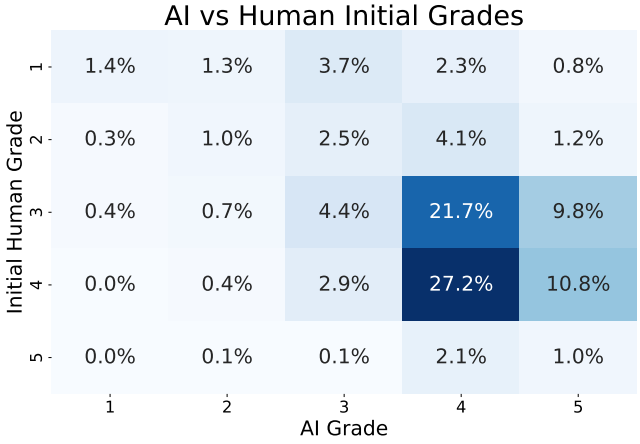
Notes: The figure displays coefficients from an OLS regression at the application level of mean of the question-level likelihood of being LLM-generated on different demographic controls, for a subset of people for whom we have all these controls available.

Figure A.11: Agreement between Human Initial Grades and AI Grades by LLM-essay



a) Not LLM-generated

	Number
Agreement rate (initial human & AI grades)	36 %
Absolute grade difference (overall)	0.88
Absolute grade difference (if diff. >0)	1.38



b) LLM-generated

	Number
Agreement rate (initial human & AI grades)	35 %
Absolute grade difference (overall)	0.93
Absolute grade difference (if diff. >0)	1.43

Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement frequencies between human initial grades and AI, by LLM-generated essays. grades (ranging from 1 to 5). The diagonal (top-left to bottom-right) indicates complete agreement. Areas above (below) the diagonal represent cases where the initial human grade grade was higher (lower) than the AI grade. The tables summarizes question counts off (row 1) and on (row 2) the diagonal.

Table A.8: Initial Human, AI, and Final Human Grades are Higher for LLM-essays

	Human initial grade		AI grade		Human final grade	
	(1)	(2)	(3)	(4)	(5)	(6)
LLM-essay	0.300***	0.284***	0.484***	0.452***	0.339***	0.315***
	(0.040)	(0.041)	(0.037)	(0.039)	(0.055)	(0.056)
Mean (non-LLM)	2.818	2.818	3.482	3.482	3.482	3.482
Controls	No	Yes	No	Yes	No	Yes
N	4,182	4,182	4,182	4,182	1,968	1,968

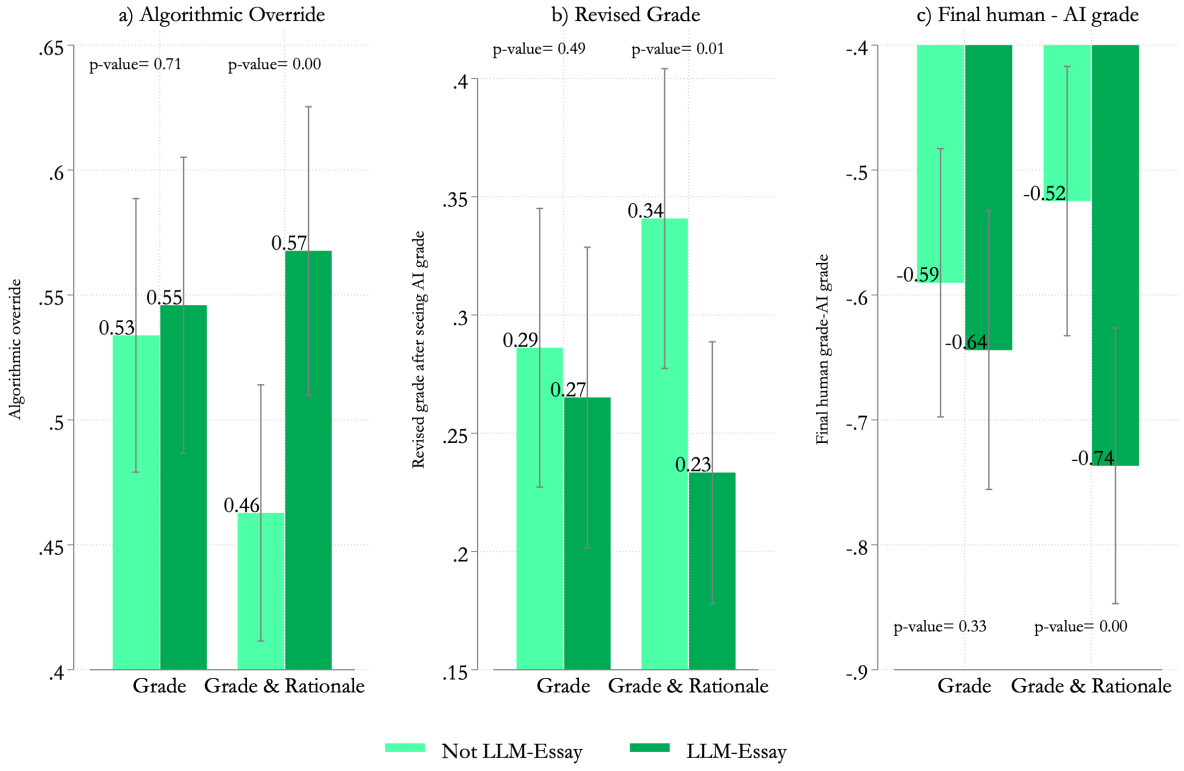
Notes: Columns 1-6 report estimated coefficients from OLS regressions respectively of Human initial grades (Columns (1) and (2)), AI grades (Columns (3) and (4)), and human final grades (Columns (5) and (6)). All columns include controls for evaluator fixed effect, the even columns additionally include controls for 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.

Table A.9: 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 1-6 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)) and the difference between final human and AI grades (Columns (6) and (5)). All columns include controls for 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.12: 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.

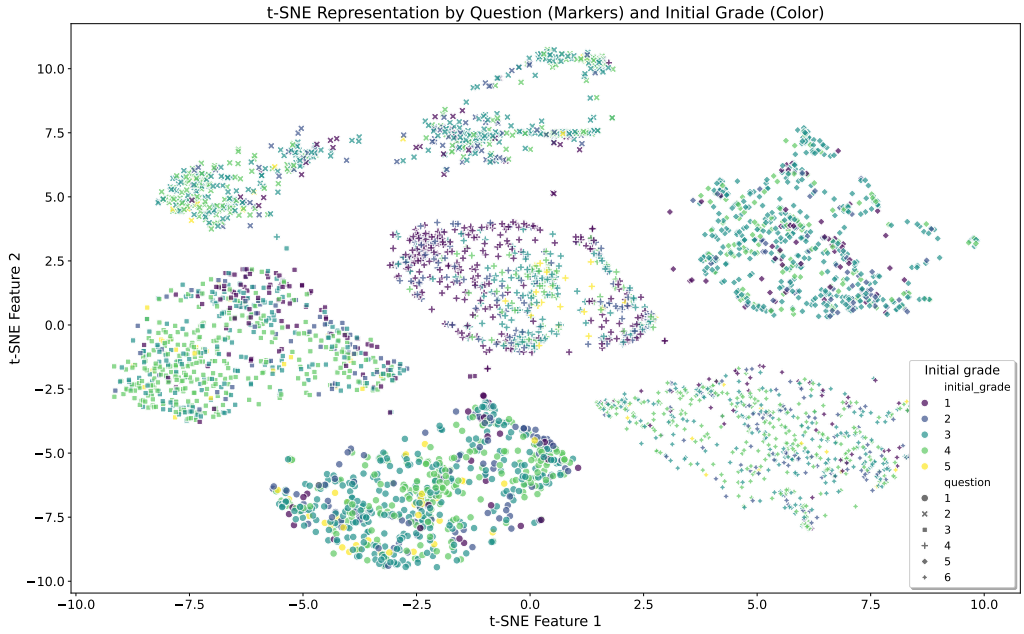
B Signal in Grades

This section presents the details of how we constructed the "semantic signal" variable mentioned in Section 4.

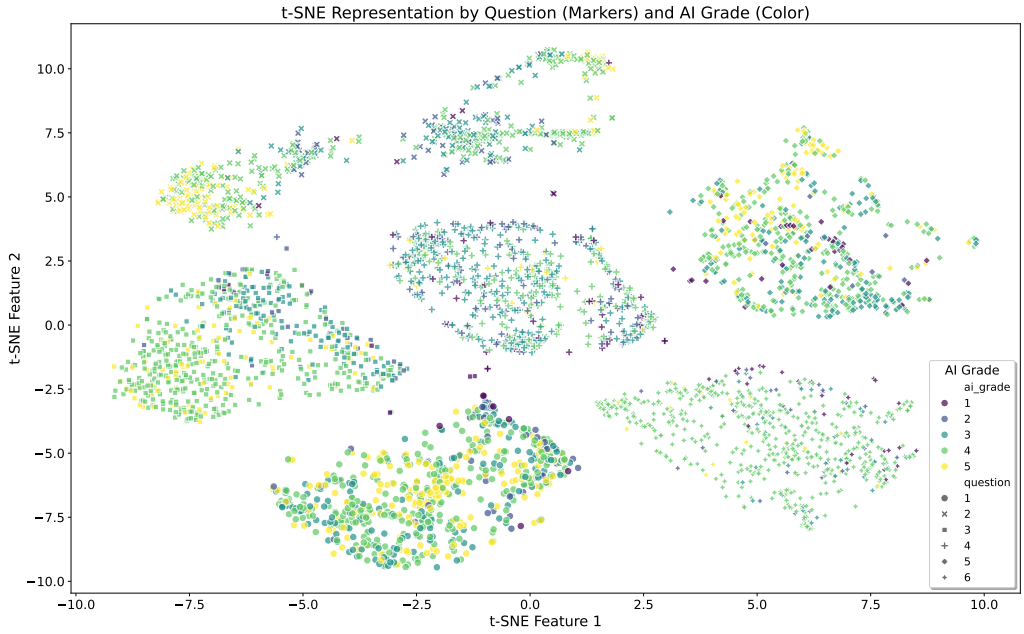
Vector Embeddings of the Essay Answers We first converted each essay answer into vector embeddings using the "voyage-lite-02-instruct" model from Voyage AI. The original 1024-component embedding vectors were first condensed to 50-dimensions using a PCA (Principal Component Analysis) reduction. Next, we used a t-SNE (t-distributed Stochastic Neighbour Embedding), a non-linear dimensionality reduction technique to project the data onto a two-dimensional plane, resulting in Figure B.13. The distance between points in the figure reflects the relative similarity of their respective high-dimensional vectors; the closer

the two points, the more similar the answers they represent. Figure B.13 reveals distinct clusters for each essay question, which suggests the embeddings effectively capture semantic features specific to the content addressed in each question. Questions 1 and 3 through 6 exhibit particularly tight clusters, indicating a high degree of thematic similarity in the responses. Interestingly, question 2 ("What is an excellent education to you, and how do you intend to provide that to your students?") stands out. Here, we observe two distinct clusters: one aligns more closely with the "alumni vision" (question 3) and the other with "core beliefs" (question 4). Moreover, consistent with the findings displayed in Figure 5, the AI awards higher grades more frequently across all questions. We can observe some minor clustering for very low grades, with the most noticeable pattern appearing for human initial grades for questions 2 (located at the 8-9 o'clock position) and 4 (centered).

Figure B.13: t-SNE Clustering of Answer Embeddings by Essay Question and Grade



a) Human initial grades



b) AI grades

Notes: The figures 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-lite-02-instruct" model from Voyage AI, codensed to 50 principal components using PCA and ultimateley to two components using t-SNE, a non-linear dimensionality reduction technique. Each point represents an individual answer's embedding.

Semantic Signal We next turn to comparing the semantic signal contained within each grade. We use the cosine similarity between each answer within a question as a proxy for signal contained in a grade, the idea being that the more signal the grades contain, the more similar to each other should the question answers be within a particulate grade than across grades.

To study semantic signal contained within grades across the three treatment pipelines, we estimate equation 2:

$$\theta_{ijq} = \alpha + \beta_1 SameScore_{ijq} + \beta_2 SameScoreXAIAssistance_{ijq} + \beta_3 SameScoreXAIOnly_{ijq} + \gamma_1 AIAssistance_{ijq} + \gamma_2 AIOnly_{ijq} + X'_{ijq}\lambda + Q_{ijq} + \epsilon_{ijq} \quad (2)$$

where θ_{ijq} are pairwise similarity scores, Q_{ijq} contains question fixed effects, and X_{ijq} is a vector of control variables that includes grades of texts i and j. Our main coefficients of interest are β_2 and β_3 which tell us how much more similar are texts *within* the same grades for grades generated by AI and Humans-with-AI-Assistance groups than for grades generated by Human-Only group, and γ_1 and γ_2, which tells us how much more (dis)similar are texts *across* the same grades for grades generated by AI and Humans with AI-Assistance groups than for grades generated by Human-Only group. Our measure of total signal for each treatment group will be the difference between β and γ coefficients; Human-Only, AI-assistance, and AI-Only signal in grade will be β_1; $\beta_1 + \beta_2$ - γ_1; and $\beta_1 + \beta_3$ - γ_2, respectively. To ease the interpretation of the coefficients, we standardise the cosine similarity scores, θ_{ijq}, with the mean and standard error of across grade Human-Only similarity scores. The coefficients can therefore be interpreted as standard deviation differences.

Figure B.14 visualizes the measure of total signal constructed from the coefficients presented of estimating equation 2. Figure B.14 panel A presents the total signal in grades by pipeline, and panel B additionally presents the signal by question. On average, AI-Only grades contain the most semantic signal, followed by Human-with-AI-Assistance grades. Specifically, Human-Only, Human-with-AI-assistance and AI-Only grades contain 0.045, 0.196 and 0.451 signal, respectively. In practice, this means that texts within grades are 0.045 SD more similar than texts across grades for the Human-Only group. For grades in Human-with-Assistance and AI-Only groups, this difference is 0.196 SD and 0.451 SD, respectively. There is substantial heterogeneity across questions (Panel b), with lowest difference between AI-Only and Human-Only for question 4 and the biggest difference for question

3.

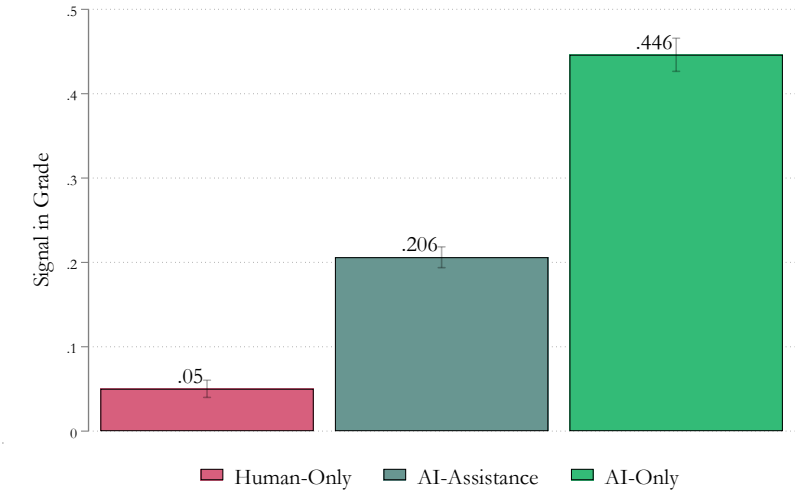
For an alternative measure of the amount of "signal" contained in grades, we train a random forest classifier on an 80% sample of the question answers to predict the grades on the rest of the sample. We then look at the model's performance metrics overall and for each grade. The idea is that when there is more text-based signal contained in each grade, the model will perform better. Table B.10 displays the performance metrics for a random forest classifier for initial human (Panel a) and AI (Panel b) grades. We look at precision (proportion of correctly classified grades among those the model predicted for a specific grades), recall (proportion of actual instances within a specific grade category that the model correctly identified), F1-score (a harmonic mean of precision and recall), accuracy (overall proportion of correctly classified grades), unweighted average and weighted average (average weighted by the number of observations). While the performance metrics vary significantly across grades, our results show that the when the random forest classifier is trained on AI grades, it has higher overall accuracy (for 6 p.p., 11%), and higher weighted average in precision (13.4 p.p., 30%), recall (6.6 p.p., 13%), F1-score (4.8 p.p., 10%).

Table B.10: Random Forrest Classifier

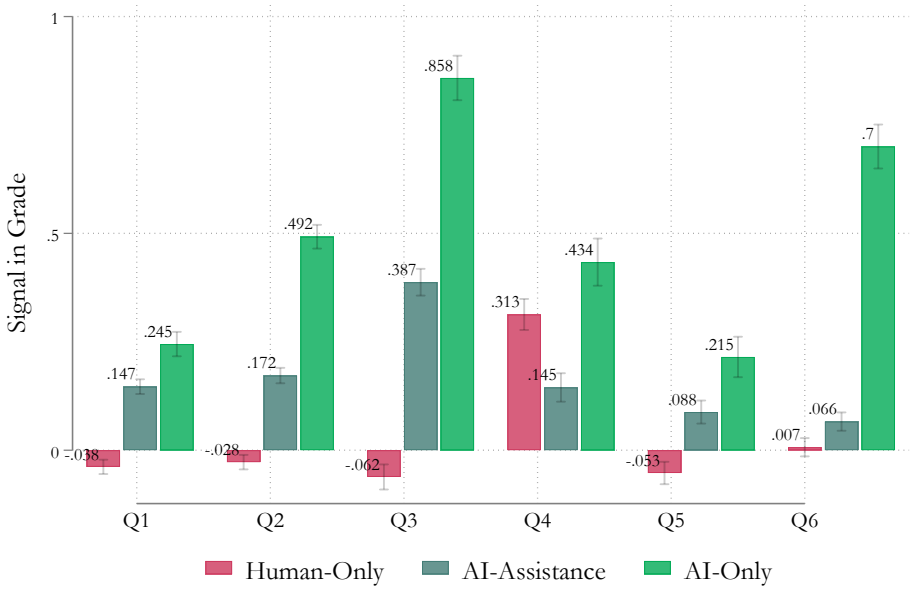
Panel a: Human				
Grade	Precision	Recall	F1 Score	Observations
1	0.620	0.648	0.633	88
2	0.143	0.009	0.017	109
3	0.503	0.669	0.574	338
4	0.502	0.537	0.519	270
5	0.000	0.000	0.000	32
Weighted Average	0.449	0.513	0.468	837
Accuracy	0.513			
Panel b: AI				
Grade	Precision	Recall	f1 Score	Observations
1	1.000	0.130	0.231	23
2	0.333	0.016	0.031	61
3	0.481	0.346	0.402	188
4	0.589	0.911	0.715	425
5	0.744	0.207	0.324	140
Weighted Average	0.583	0.579	0.516	837
Accuracy	0.579			

Notes: The table presents performance metrics for a random forest classifier, evaluated on the original text embeddings of the essay questions (both panels a and b). Panel a) represents the metrics for grades assigned by humans, and panel b) for grades assigned by the AI. For all panels, an 80-20 train-test split was used to assess the model's performance on unseen data. Precision: measures the proportion of correctly classified grades among those the model predicted for a specific category (e.g., Grade 3). Recall: measures the proportion of actual instances within a specific grade category (e.g., Grade 3) that the model correctly identified. F1-Score: a harmonic mean that combines precision and recall, providing a balanced view of the model's performance. Accuracy: Overall proportion of correctly classified grades across all categories. Weighted Average: The weighted average value for each metric (precision, recall, F1-score) calculated across all grade categories.

Figure B.14: Semantic Signal in a Grade



a) Total signal in grade



b) Total signal in grade by question

Notes: The figure shows the total signal contained in a grade, for applications that were assigned to Human-Only, Human-with-AI-Assistance and AI-Only decision pipelines. The signal was computed using regression coefficients from equation 2 using pairwise cosine similarity scores of the answer vector embeddings that were generated using "voyage-lite-02-instruct" model from Voyage AI. The coefficients were standardized so the size of the bars can be interpreted as SD deviation differences from "Human-Only" across-grade-similarity. Panel A: Total signal in grade in each policy pipeline. Panel B: Total signal in grade in each policy pipeline by question.

C Technical Appendix

C.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>
```

C.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+++"
```