Automation with Generative AI?

Evidence from a Teacher Hiring Pipeline *

Kobbina Awuah† Urša Krenk‡ David Yanagizawa-Drott§

July 11, 2025

Most recent version here

Abstract

Can generative AI improve hiring? We experimentally embed generative AI (GPT-4) into a teacher-recruitment screening process, comparing three pipelines: (i) human-only, (ii) human with AI assistance, and (iii) fully automated screening. Automation increases downstream hiring success by 11 percentage points (a 73% improvement) over the human-only baseline. In contrast, AI assistance neither improves outcomes nor productivity, as users systematically disregard its recommendations, perceiving it as incapable of distinguishing between signals of teacher quality. Our results provide evidence in favor of fully automated deployment of generative AI for an economically important task, highlighting potential limitations of hybrid approaches involving humans-in-the-loop.

Keywords: Artificial intelligence, automation, screening, labor markets.

*We are grateful to Maria Korobeynikova, Minh Trinh, Andrin Pluess, and Alessandro Vanzo for excellent research assistance. We also thank seminar participants at Collegio Carlo Alberto, UC Berkeley Development Lunch, AI in Social Science Conference at the University of Chicago, ESA Helsinki, Workshop in AI + Economics at the University of Zurich, Stone Inequality Seminar at UBC, and the Center for Behavioral Institutional Design Seminar at NYU Abu Dhabi.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

In recent years, advanced generative AI (GenAI) technologies have rapidly evolved, enabling machines to perform a wide range of complex tasks with exceptional accuracy. This progress and subsequent mass adoption have revived discussions about how these technologies should be applied in the labor market: whether they are best used to replace human labor entirely through automation or to assist and augment human workers in performing their tasks (Acemoglu, Autor, and Johnson, 2023; Acemoglu and Restrepo, 2019). Understanding whether "automation" or "augmentation" is more effective has significant economic implications, as GenAI is expected to impact a broad spectrum of occupations across industries (Eloundou et al., 2023). In this paper, we contribute empirical evidence to this ongoing debate by examining the effects of GPT-4, one of the earliest GenAI models to achieve widespread adoption after it was commercially released in April 2023. In the weeks after its launch, we conduct a field experiment focusing on screening for talent, a context characterized by widespread AI adoption by both recruiters and job applicants (Insight, 2024).

We partner with a nonprofit organization that recruits recent university graduates for prestigious teaching fellowships in rural schools in Ghana and embed GPT-4 into its hiring process. At the initial screening phase, which consists of grading short essays written by applicants, based on a fixed set of criteria defined by the organization, we randomly assign applications to one of three pipelines relevant for policy-making, each differing in how applications are graded: (1) *Human-Only*: the current status quo, where evaluators rely solely on their own judgment, and only their grade determines whether a candidate advances to the next stage of the hiring process; (2) *Human-with-AI-Assistance*: augmentation of human workers with GPT-4, where evaluators first record an initial grade, then review a GPT-4-generated grade recommendation before finalizing their evaluation, with the final grade deciding advancement; and (3) *AI-Only*: where GPT-4 fully automates the evaluation, and its assessment alone determines which candidates advance.[1] The question of whether artificial intelligence should augment human labor or fully automate it is central to contemporary economic and policy debates. While prior work has compared these augmentation and automation pipelines for traditional predictive models in real-world settings (e.g., Agarwal et al. (2024)), or has studied the powerful augmentation effects of generative AI in the field (Noy and Zhang (2023), Brynjolfsson, Li, and Raymond (2025)), an experimental comparison of all three policy pipelines – human-only, AI-assisted, and full automation – using generative AI has been absent from the literature. We conduct, to our knowledge, the first randomized controlled trial to test these three pipelines using a generative AI in a real-world labor market, where screening

[1]We prompted GPT-4 to not only provide the grading score but also the rationale for the score. In the assistance pipeline, we randomized whether the rationale was visible to the human user. We describe this in detail below, but unsurprisingly since in the vast majority of the cases the assistant recommendations were ignored, there is no average differential effect from providing the rationale.

decisions carry direct economic stakes in the form of influencing the likelihood of job offers and employment for applicants.[2]

Our design allows us to directly compare the effectiveness of conventional human evaluation, human augmentation via AI assistance, and full automation in identifying candidates who later perform well in in-person assessments and ultimately receive job offers. Importantly, while the policy pipeline used for advancement is randomly assigned, we employ parallel grading: every application is evaluated independently by both a human (with or without AI assistance) and by the AI. This approach enables us to conduct precise counterfactual analysis of grading outcomes for the same applicants.

We document that when we remove humans from the pipeline entirely and rely solely on the AI grading, the downstream offer and hiring rates increase substantially – approximately an 84% and 73% improvement, respectively, compared to the baseline where human screeners receive no assistance. This is direct evidence that the *AI-Only* policy outperforms the *Human-Only* policy. However, when we provide evaluators with the same algorithm as an assistant, we find no significant improvements in downstream offer rates or hiring success. In a majority of cases (over 80%), humans override the algorithm's recommendation. Our design also allows us to carefully measure how much time human graders spend on each applicant. Here, we find that AI assistance increases grading time by approximately 13-17%. Together with the lack of evidence of improvements in downstream outcomes, our experiment finds little to no support for the hypothesis that *Human-with-AI-Assistance* improves productivity on the task. If anything, it worsens productivity.

We then examine the potential behavioral mechanisms behind the relative underperformance of the pipeline with a human-in-the-loop. *Ex ante*, it is not clear why assistance would not be on par with, or even superior, to an AI-only pipeline. In theory, human decision makers may have complementary skills (manifested in various ways, such as via better contextual information, tacit knowledge, etc.) which can be combined with the narrow capabilities underlying a generative AI system like GPT-4. Presumably, such complementarities will be context specific. In our setting, experienced teachers were evaluating potential teachers. One would therefore expect complementary skills relevant for the context. Moreover, even if no such complementarity exists, a simple effort minimization approach could result in performance on par with an AI-only pipeline. That is, if an AI system is useful (as is the case here), a user can essentially 'copy and paste' the recommendation, achieving performance on par with AI-only – and superior to human-only – at a fraction of the

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

human-only effort.[3] Clearly, with an algorithmic override rate of 80% and more time (not less) spent on the task, this was not the case. To examine mechanisms behind the algorithmic override, we identify patterns of systematic disregarding of AI advice. In short, our evidence points to users misperceiving the AI assistant as incapable of distinguishing between high- and low-quality signals of teacher quality, based on the submitted material.

To understand the behavior behind the systematic algorithmic override, it is important to understand the context of the experiment. In particular, the organization experienced a new phenomenon as ChatGPT had been launched only a few months earlier: widespread submissions of "LLM-essays". While this was unprecedented at the time of the experiment, it is now well documented across labor markets that job applicants use LLM tools like ChatGPT to craft their application materials. For example, survey data from the U.S. by Insight (2024), collected in November 2024, suggest that more than 40% of job applicants in the labor market use LLMs to help with the tailoring of cover letters, resumes or other submission materials.[4] Almost 90% of hiring managers claim they can tell if an applicant is using LLMs. Whether or not hiring managers should be taking into account such tool usage, as a negative or positive signal of underlying candidate quality, is still an open question. Approximately half of managers say they do take it into account, and the other half say they do not. The same issue was debated among the evaluators during our experiment, and the data suggest it greatly influenced decision making.[5]

Using a state-of-the-art LLM detection tool with low rates of false positives, we find that approximately 60% of applicants use LLMs for their essays.[6] LLM-essays are longer, more complex, and contain less applicant-specific information. Applicants who rely on LLMs complete their applications faster. Evaluators initially 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 with AI-assistance. Evaluators are about 25% less likely to follow the AI recommendation when grading an LLM-essay. Evaluators initially incorporate the model's suggestions, but as they see

[3] An example such potential behavior under comes from the ChatGPT experiment by Noy and Zhang (2023). While strictly speaking there were only two policy pipelines, *Human-only* and *Human-with-AI-Assistance*, in practice the latter pipeline resulted in only very minor human edits of the AI output. Those edits were deemed largely ineffectual. From a simple treatment effects perspective, productivity increased as time spent on the task was reduced. The authors conclude that "*It is not obvious whether these dynamics should be interpreted as evidence that ChatGPT will displace human workers or evidence that it will augment them. Although ChatGPT directly substituted for participants' effort with little need for human input, it also enabled participants to complete tasks much faster.*"

[4] The practice appears widespread. A recent UK survey from March 2025 also found that over 50% of job applicants use generative AI, with 41% specifically using it to draft cover letters (Thomas, 2025).

[5] Throughout this paper we use the terms "AI-generated", "LLM-generated" and "LLM-essays" interchangeably to refer to essays written by job applicants with the assistance of generative AI tools.

[6] In Section 4, we describe the detection tool, explain how we assess false positives, and show how classification rates vary with different thresholds and levels of aggregation.

that the algorithm does not penalize LLM-generated essays, thereby diverging from their own evaluations, they lose trust in the tool. Rather than increasingly agreeing with AI over time, evaluators become less dependent on it. This pattern is also confirmed through qualitative interviews. The fact that the AI-assistant is unable to distinguish between human-written and LLM-written essays influences evaluators' views on its ability to distinguish high-quality from low-quality applicants. In this sense, our results echo the framework by Vafa, Rambachan, and Mullainathan (2024), in that humans may generalize about the ability of an AI in many ways, and if it performs poorly on one specific task humans may overreact in forming beliefs that it will also perform poorly on a related, but nevertheless distinct, task. We find evidence consistent with this mechanism, which helps explain the lack of hiring success in the *Human-with-AI-Assistance* pipeline.[7] By contrast, we show that the automation pipeline performs best in part because it scores a higher share of candidates being of high quality based, advancing them to the next interview stage based on the grading rubric set by the organization. We find additional supporting evidence that AI grades are less noisy, as captured by correlations with evaluation scores in the separate interview stage as well as how the semantic content of the essays consistently maps into grades. Put simply, the evidence is consistent with the GPT-4 algorithm applying the screening instructions (as defined by the organization's grading rubric) in a consistent manner, whereas in the other two pipelines human-induced deviations from the screening criteria dominate, which resulted in noisier grades, penalization of applicants that used LLMs and lower hiring success.

Our results should be interpreted with appropriate caution. While we find experimental evidence in favor of automation, this evidence should be assessed within the specific labor market context, time period, and AI technology in which our experiment was conducted. Moreover, from an external validity perspective, we find evidence that decision-making under *Human-with-AI-Assistance* is influenced by the widespread use of generative AI in the economy – applicants had started using LLMs. Evaluators appeared to lose trust in the AI tool because they perceived it as unaware of, or unable to account for, this widespread trend. Moreover, ChatGPT had been launched only a few months before the experiment, at a time when global mass adoption had not yet been reached. Beliefs about the technology have, presumably, evolved since then. More importantly, generative AI is an evolving technology which can be shaped in various ways. Our experiment deployed an early version, GPT-4, but generative AI can be tailored through training and fine-tuning strategies, prompting techniques, and systems orchestration approaches. In principle, contrary to GPT-4, AI can be designed and trained to maximize human-complementarity in specific tasks (McLaughlin and Spiess, 2024). Therefore, while our paper presents evidence which speaks in favor of automation

[7] A complementary way to think about the result is provided by McLaughlin and Spiess (2024). They consider AI-assistance through the lens of a principal-agent framework, where humans have some comparative or absolute advantage. Their framework considers the optimal design, including the benefits of withholding recommendations in some cases, which our paper does not address. Instead, we study a simple approach where the same information set and task is given to a generative AI algorithm (GPT-4), which arguably is commonplace in everyday use cases.

using GenAI, this evidence should be interpreted with appropriate nuance.

Our findings contribute to multiple strands of literature. First, we contribute to the literature studying the effects of integrating AI, and in particular GenAI, technologies, into the workplace either as aids to human workers or as tools for automation. Previous research has focused either on comparing traditional machine learning methods using observational data (Angelova, Dobbie, and Yang, 2023), or on using an experimental approach with analogous treatment groups (Agarwal et al., 2024), or on evaluating generative AI exclusively as a tool for "augmenting" humans, without including a treatment arm involving full task automation (e.g., Brynjolfsson, Li, and Raymond (2025); Noy and Zhang (2023). Our study, however, is to our knowledge, the first to empirically test GenAI's performance across all policy-relevant pipelines within a real-world economic setting with high stakes.

Given the growing adoption of AI in recruitment due to its potential to enhance these processes (Vrontis et al., 2022), our setting addresses an area of increasing relevance for organizations. Much of the existing research on recruitment has focused on how AI affects the diversity of hires and its potential to reduce biases against certain groups within the applicant pool (Avery, Leibbrandt, and Vecci, 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 the selection of higher-quality candidates (Cowgill, 2020; Chalfin et al., 2016) and that the use of AI by job applicants to craft applications may lead to a higher number of successful matches (Wiles, Munyikwa, and Horton, 2025). Note that, however, none of the above papers examine the effects of GenAI.

We also contribute more generally to recent work documenting that generative AI can boost productivity in tasks like coding, writing, and general consulting tasks (Brynjolfsson, Li, and Raymond, 2025; Bubeck et al., 2023; Dell'Acqua et al., 2023; Noy and Zhang, 2023; Peng et al., 2023; Kumar et al., 2023). However, even a highly capable model does not ensure improved outcomes unless users trust its output and are able to incorporate it effectively, which did not happen in our case. This relates to literature exploring why people might systematically disagree with AI recommendations, due to factors such as algorithmic aversion (Dietvorst, Simmons, and Massey, 2015), bias against AI-generated content (Parshakov et al., 2025), priors that are far from algorithmic recommendations (Kim et al., 2024), cognitive constraints (Agarwal et al., 2024), or the tendency to overgeneralize performance of LLMs across different domains (Vafa, Rambachan, and Mullainathan, 2024).

Finally, our research is among the first, alongside Otis et al. (2024), to systematically analyze the capabilities and productivity implications of generative AI assistance powered by novel large language models specifically in developing-country contexts. Most related research in these regions has focused primarily on AI-powered educational tools designed to improve learning outcomes (Chen et al., 2024; De Simone et al., 2025).

The rest of the paper is structured as follows. In Section 2, we explain the setting and the

experimental design. In Section 3, we cover a simple conceptual framework to structure the thinking around the signal extraction problem in the three policy pipelines. Section 4 covers the data and estimation framework. Section 5 presents the main results. Section 6 explores mechanisms behind the relative under-performance of the pipeline with humans-in-the-loop. Section 7 concludes.

2 Background and Experimental Design

2.1 Background and the Organization's Recruitment Process

We collaborate with a Ghanaian educational non-profit organization. The organization recruits recent university graduates and places them in disadvantaged rural schools nationwide for a two-year teaching fellowship program. Prior teaching experience is not required, but candidates must hold at least a bachelor's degree before starting the program. The organization provides extensive pre-placement training and on-the-job support throughout the two-year fellowship. Candidates can apply for this position as either a regular job or as part of their compulsory "National Service".[8] Every year, a cohort of between 50 and 150 fellows is assigned to schools in rural areas who earn a stipend comparable to the average entry-level salary in Ghana.[9][10] The position is considered prestigious, and the candidate selection process is competitive, with only 15–20% of applicants being offered positions.[11] Importantly, the organization usually sets a number of slots to fill, and if they do not find enough high-quality candidates to fill those slots, the positions remain unfilled. After the program, the majority of fellows (about 60%) stay in the education sector (either working in educational non-profit organizations or as teachers in schools). Among those who leave the education sector, many work for other non-profit organizations or in the public sector.

Figure 1 illustrates the supply and demand sides of the recruitment process for the partner non-profit organization, as well as the design of our policy experiment.[12] 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

[8] 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.

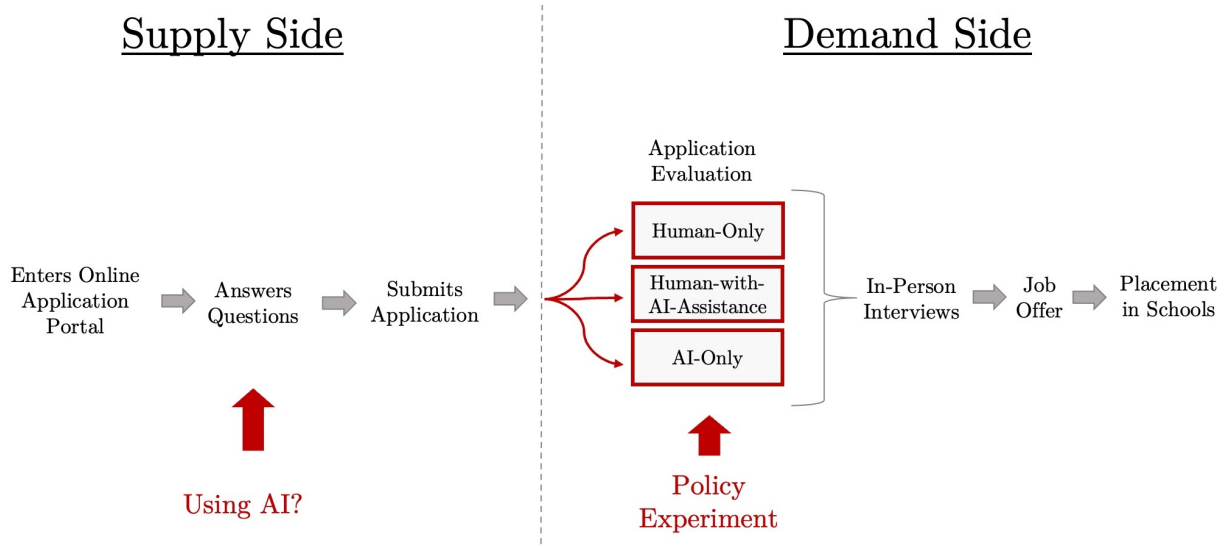
[9] Ghana has 16 regions in total, and the partner non-profit is present in 10 of them.

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

[11] 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.

[12] 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.

Figure 1: Supply and Demand Sides 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.

score are invited to in-person interviews, after which fellowship offers are given.[13] Recruitment is cyclical and typically occurs between March and July. If a candidate accepts the offer, they begin their fellowship between October of that year and January of the following year.

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

[13] 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.

status, country of residence, and graduation year. Applications of ineligible candidates are not graded.[14]

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

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

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

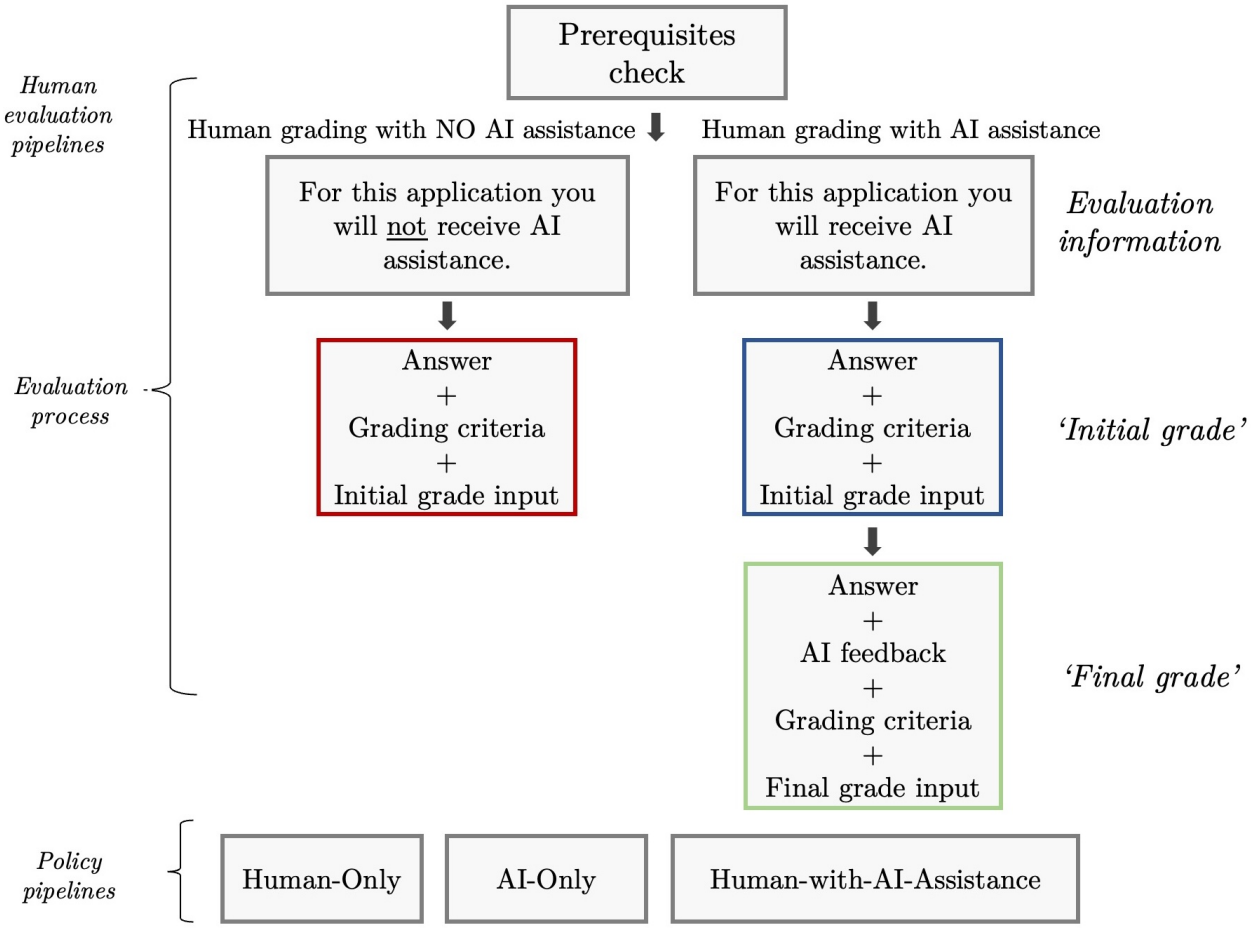
Figure 2 illustrates the process the evaluators followed for grading.[15] Evaluators were first shown information that determines applicants' eligibility for the program (i.e., "Prerequisites check"). If a participant failed to meet the eligibility criteria (most commonly, having a "Higher Education 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's 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. 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

[15] 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.

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.

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").

The goal of this design was twofold. First, by randomizing applications into three distinct policy-relevant pipelines, we can evaluate the causal impact of each grading approach on downstream

outcomes such as job offer rates and hiring, shedding light on the relative effectiveness of using GPT-4 as an assistant and tool for automation relative to conventional grading. Second, the parallel grading, where each essay is graded by both humans (with or without AI assistance) and the AI, allows us to compare differences between human initial and AI grades, as well as between human final and AI grades, for the same set of essays.

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 (for details see 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.[16] The core of the content prompts (for details see Appendix Section C.2) consisted of instructions from the evaluator manual, including the grading criteria for each grade (1 to 5) and definitions for relevant terms (e.g., a specific definition of "resilience and adaptability"). The prompts had the following structure:

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

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

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

4. The grading criteria. The criteria from the training manual were "augmented"[17] with grade-specific factors. For instance, for question 2, grade 3, the augmented criterion read (the

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

[17] 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

augmented part in italics): "Clearly defines an excellent education and outlines a path to offering it to students. *This includes a) sharing relevant personal experiences and background, b) demonstrating adaptability and flexibility, c) displaying passion and enthusiasm, d) demonstrating clear communication and organization, and e) exhibiting some problem-solving and critical thinking skills.*"

3 Conceptual Framework

For conceptual clarity, we now describe the screening task and our experimental policy pipelines in a structured manner. We start by describing how the organization conducted screening before the introduction of AI. We then discuss how introducing AI can be conceptualized either as an "automation" or as an "augmentation" technology, and how it relates to the fundamental signal extraction problem.

The Status Quo (Before AI). The goal of the organization is to screen for applicants who are likely to be a good fit, given its mission. We can think of teacher quality as an unobservable vector θ consisting of several dimensions $j = 1, \ldots, K$ each influencing student learning through an unknown production function $y = f(\theta)$. In our context, the organization views a sufficiently "good teacher" as someone above a (very demanding) quality threshold $\bar{\theta}$, and is willing to give offers to anyone deemed above that. Since rural areas are under-served, there is not an issue of hiring too many teachers. The organization's mission is to prioritize excellence by ensuring that only highly qualified teachers are placed in schools.

As described in Section 2.1 above, the organization screens for quality in two stages. First, it evaluates written submissions. The organization's leadership deems that the following dimensions are to be evaluated in the first stage: Motivation; Educational Philosophy; Alumni Vision; Value Alignment; Grit and Leadership. In practice, this is done by asking the applicant to answer a specific question. For example, the educational philosophy question: 'What is an excellent education to you, and how do you intend to provide that to your students?'. In the application process, each job applicant i submits their answer X_{ij}. The leadership of the organization specifies a set of criteria, C_j, along each dimension.

The grading task is to provide a grade S from 1 to 5, strictly following the criteria. Formally, the task assigned to the human evaluators conducting the screening can be described by the following function:

$$G_{ij} = f(X_{ij}, C_j).$$

In this formulation, the task given to human graders by the leadership effectively resembles a

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.

step-function that takes textual input, similar to classification tasks commonly found in Natural Language Processing. The grade determines whether applicants proceed to the next stage, which consists of a full day of in-person interviews and evaluations. An applicant proceeds only if they are deemed "above the bar", based on their average grade. For each applicant i, define their average grade across the $K = 6$ dimensions as $\bar{G}_i$. The organization's policy for advancing a candidate is:

$$\text{AboveBar}_i \;=\; \mathbf{1}\{\bar{G}_i \geq 3\}.$$

That is, the leadership of the organization assigns equal weight on all six dimensions and requires an average grade of 3 out of 5. Human evaluators are explicitly instructed to focus independently on each dimension (that is, on each answer), treating each separately and grading strictly according to the provided criteria. In practice, the screening process is sequential: evaluators assess the first dimension, then the second, and so forth. In the second, in-person evaluation stage, candidates are assessed on several additional dimensions, which are broader and distinct from those in the first stage. Applicants whose cumulative interview grade $\tilde{\theta}$ is above a set numerical threshold (50 out of 100 in our case), meant to capture $\bar{\theta}$, receive an offer. This means there is no crowding-out: all applicants who surpass the quality bar are offered positions. We are *not* claiming that this is the optimal policy to screen for talent. Rather, it is a detailed description of the organization's existing screening approach and the underlying mission-based philosophy guiding it.

Introducing AI for Screening. The description above refers to the organization's status quo, human-only, screening process up until 2022, before ChatGPT and other generative AI models became widely available. A key advantage of generative AI is that it is pre-trained and general-purpose: users can simply provide a task, and the model quickly generates text output at minimal cost, without requiring specific expertise or data. However, since generative AI is a black-box tool not trained specifically on the user's context, the accuracy of its output remains uncertain.[18]

In collaboration with us, the organization decided to experiment with AI in their hiring process. They wanted to answer two key policy questions:

1. Can generative AI improve the screening process compared to the status quo?
2. Is it better to automate the screening task entirely or to augment human graders with an AI-assistant?

Our experiment studies three alternative (randomized) screening policy pipelines. In theory, the

[18]One should distinguish between closed-source and open-source models. The latter can be downloaded and fine-tuned for specific tasks with a user's own data. Doing this requires machine learning expertise, however. At the time of writing this paper, closed-source models greatly dominated in terms of market shares (primarily models developed by OpenAI, Anthropic, and Google, typically accessed through web-based applications as well as their corresponding API services).

final grade for any given applicant i on some dimension j could be determined in one of three ways

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

Here G_H is the Human-Only grade (status quo), G_{AI} is the grade in the AI-Only condition (automation) and G_{HAI} is the final human graded grade after seeing the AI recommended grade (Human-with-AI-Assistance). We randomize the decision-making process at the applicant level, i. Comparing these three pipelines allows us to answer the key policy questions. Our design allows us to not only compare human decision-making with and without AI (G_H, G_{HAI}), compared to task automation (G_{AI}), but we can estimate treatment effects on hiring success; whether a candidate ultimately ended up above the bar ($\tilde{\theta}$) for a job offer.

Ex-ante, it is unclear how language models will evaluate essays relative to human graders, or how human effort and decision-making are influenced by AI assistance. The grading task is a complex and cognitively demanding signal extraction problem. It involves paying attention to often several paragraphs of text, in a high-dimensional semantic space, and assess which criteria are fulfilled. As such, inattention and errors due to complexity may be expected (Gabaix and Graeber, 2024; Gabaix, 2019). Moreover, as there are no explicit monetary rewards attached to grading in a certain way, and there is no verifiable way to unambiguously conclude a grade is incorrect, the incentives to provide effort may be weak. Human graders (experienced teachers in this case) may also disagree with the criteria set out by the organization, perhaps based on their own experiences and perspectives. They could also be biased towards certain groups. While no explicit information about gender and ethnicity was provided during grading, such information could potentially be inferred from the essay answers. In general, these issues can be characterized as a principal-agent problem between the organization and the human graders.

LLMs, unlike human graders, do not suffer from such issues. They can pay full attention to all text input as long as it fits into their "context window", which in our case was always satisfied. In this regard, its signal extraction capabilities may be very good. However, LLMs may not have tacit knowledge about the context that experienced teachers might have. Their training data is unlikely to include detailed information about the educational context in rural Ghana. Those additional, or nuanced, aspects of the signal extraction problem could be key for assessing the underlying teacher quality. LLMs could also be biased in a number of ways.

Finally, adding AI assistance can introduce complexity into decision-making since humans have to consider additional information, the AI-generated grade G_{AI}, and form beliefs about how it was generated. Since the AI grading function f_{AI} is a black-box, evaluators observe only its output, creating ambiguity about how G_{AI} should be interpreted. Human decision makers must therefore decide whether, given the input X_{ij}, to trust G_{AI}. Human intuitions about LLM performance across different tasks are often inaccurate, even within familiar domains such as mathematics or

moral reasoning, as documented by Vafa, Rambachan, and Mullainathan (2024). Specifically, LLMs often perform poorly when humans expect them to perform well, and vice versa.[19]

Thus, there are multiple reasons why grading outcomes might differ between human evaluators and language models. Determining which policy pipeline best identifies suitable candidates is ultimately an empirical question. We pay particular attention to cases where there is initial disagreement between human and AI grades (i.e $G_H \neq G_{AI}$) to observe if and when humans incorporate AI recommendations into their final grades (G_{HAI}). An advantage of our design is that all applications were graded by both an AI and a human, which allows us to examine differences in grading for the same applicant (see Figure 2). Additionally, we also examine how these total grades (G_H, G_{AI}, G_{HAI}) correlate with downstream outcomes $\tilde{\theta}$ from the in-person interviews, which serve as the organization's proxy for "good teachers". We would expect a stronger correlation and higher hiring rates in a good screening pipeline.

LLM Usage by Applicants. As Figure 1 illustrates, an important aspect is that generative AI may not only be used by employers during screening, but also by applicants themselves. Consequently, application materials may result from an iterative interaction between the human applicant and a generative AI platform such as ChatGPT, significantly complicating the screening process. For example, imagine the employer's screening criteria C_j are somewhat known (though imperfectly) from job descriptions or the employer's website, and suppose that applicants have some noisy beliefs about what the employer values. Then, when the employer is asking for things like "What is an excellent education to you, and how do you intend to provide that to your students?", an applicant could then feed those as input to ChatGPT and ask it to generate some text X_{ij} that fits the criteria.[20]

This raises a fundamental challenge for screening in labor markets: What factors should be taken into account? That is, what input predicts a "good teacher"? Depending on one's view, it will have implications for the three policy pipelines and the role of generative AI, as there could be misalignment. To see this, we discuss a set of misalignment possibilities next.

Misalignment Possibilities. The leadership has one view, represented by the criteria C_j, on what

[19] Beyond generative AI, there is a broader issue of how humans treat probabilities when provided with algorithmic assistance. In a recent study with radiologists by (Agarwal et al., 2024), radiologists perform equally well under Human-Only conditions as an AI-Only pipeline, but AI-assisted decision-making leads to more errors due to cognitive limitations. Although this paper did not specifically examine generative AI, it highlights general challenges and potential pitfalls associated with algorithmic assistance.

[20] Two important aspects arise here. First, using ChatGPT directly affects the generated text. Second, even if two essays were identical, one could ask if the method used to create the essay should influence grading. The underlying signal about candidate quality might differ depending on the generation process. This again highlights the question of what constitutes an optimal screening algorithm. However, this paper does not seek to determine the optimal algorithm but rather explores human decision-making when AI assistance might be misaligned with human preferences or beliefs.

predicts a "good teacher". Human graders, who in our case are experienced teachers themselves, having served as previous fellows, may agree with these criteria and follow them to the best of their ability. However, they might also hold alternative views on the fundamental signal extraction problem – what makes for a "good teacher" – and sometimes deviate from the official grading policy. When this happens, there is misalignment *among humans* inside the organization. In short, a classical principal-agent problem.

The key difference in the AI era is that leadership can now assign decision-making to two types of agents: human graders or AI. Consider how this relates to job applications generated by large language models (LLMs). If human graders take into account whether an application was AI-generated and discount it, then when they ask themselves, "Given the input X_{ij}, do I trust the output from G_{AI}?", the algorithm would be *misaligned with the human graders*. This occurs because generative AI models are trained using reinforcement learning to follow instructions (Ouyang et al., 2022), which in this case are set by the leadership, but human graders may adjust their trust based on their own judgment. We would therefore expect $G_{HAI} \neq G_{AI}$ and $G_H < G_{AI}$, i.e. there would be *algorithmic override* due to the misalignment in a specific direction. If the criteria C_j truly predicts a "good teacher" and those are identified later in the hiring funnel, the AI-Only pipeline would yield better hiring outcomes. By contrast, we would not necessarily expect that pattern when the input X_{ij} is not (perceived) to be generated by an AI, because there is less misalignment and algorithmic override.[21] In sum, there could be different types of misalignment: between leadership and its workers, between leadership and the algorithm, and between the workers and the algorithm.

A final note on important context: the use of ChatGPT by job applicants in our experiment was unexpected. This issue first arose within the organization during the experiment, as ChatGPT had not existed the previous year. It was intensely discussed inside the organization. Human graders claimed they could tell whether an essay was partially or fully generated by an LLM and felt that this should be considered a negative signal. Ultimately, leadership decided not to change their grading policy. Graders also knew that the AI recommendations strictly followed the established criteria. This awareness likely played a significant role in explaining the experimental results and the apparent lack of trust in the AI-generated grades.

[21] This intuition is similar to the principal-agent framework outlined by McLaughlin and Spiess (2024). This mechanism differs from other explanations for algorithmic override, which often refer to either information asymmetry (where the human and the AI have different information sets) or capability asymmetry (where the human and the AI differ in their ability to process information).

4 Data, Outcomes, and Empirical Strategy

4.1 Data

Our experiment involved the evaluation of 697 eligible applications, corresponding to 4182 question answers. Within this set, 101 applications were independently graded by two distinct evaluators. Table A.2 presents baseline summary statistics of our sample (Panel A displays question-level summary statistics, and Panel B displays application-level summary statistics), and Table A.3 presents application-level baseline balance checks. The average length of a question answer was 2238 words (373 words for each answer), 45% of essay answers were classified as generated by an LLM[22], 60% of the applicants have at least one LLM-generated essay, and 32% 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 experiment, we were able to obtain detailed background information for approximately 75% of applicants. Among these applicants, 36% identify as female, 57% have completed their national service, 5% hold a Master's degree or higher, and 13% had previously applied to the program. 86% of the applicants come from five universities in Ghana (KNUST, University of Development Studies, University of Cape Coast, University of Education (Winneba), and University of Ghana). 39% of applicants originally come from one of Ghana's Northern regions, 14% from Volta region and the remainder from Ashanti (12%), Greater Accra (7%), and other regions in Southern and Central Ghana (27%).

Assignment of applications to policy treatment groups is largely balanced across observable characteristics. Columns 13 and 14 of Appendix Table A.3 report the joint F-statistic and the related p-value of a regression for each of the row variables on the set of three treatment indicators and strata fixed effects. We also fail to reject the null hypothesis of zero effect in a joint test of orthogonality of all variables in the table on assignment to any treatment status (p-val=0.52).

4.2 Outcome Variables

In this section, we describe our outcome variables in detail. First, we outline the question-level outcome variables used in our grading analysis. Next, we describe the application-level ("downstream") outcome variables used to analyze the effects of the three different policy pipelines. Finally, we explain how we identify and define LLM-generated essays and LLM-generated applications, and we describe their prevalence in our setting.

4.2.1 Question-Level Outcome Variables

Grades We have three types of grades in our data; "initial grades", "final grades" and "AI grades". "Initial grades" and "final grades" are grades recorded by human evaluators, while "AI grades" are

[22]We explain in detail how we define whether an answer, or the entire application, is LLM-generated in Section 4.2.3 below.

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 ($S_H \neq S_{AI}$); b) *algorithmic override*: a dummy variable that equals one if the human final grade is not equal to the AI grade ($S_{HAI} \neq S_{AI}$) for the subset of applications given the AI assistance; c) *any revision*: a dummy variable that equals one if the evaluator revised their initial grade conditional on the initial and AI grades being different ($S_{HAI} \neq S_H \mid S_H \neq S_{AI}$) 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 ($S_H - S_{AI}$); e) *difference between final grade and AI grade*: the final grade minus the AI grade ($S_{HAI} - S_{AI}$) 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 is informative of productivity in performing the task, in the sense that conditional on a given grade or hiring success rate, 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.

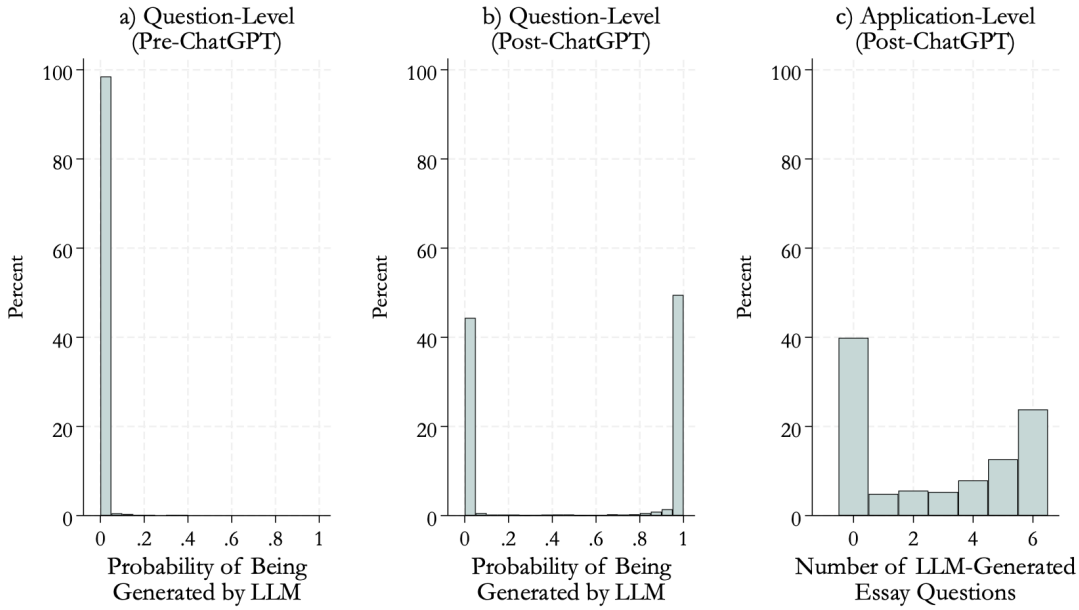
4.2.2 Downstream Outcomes (Application-Level Outcome Variables)

The total application grade, which is used to determine which candidates advance to the next phase of the selection process, is calculated by summing the individual 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.

4.2.3 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. This tool has low error rates, with an overall accuracy of 99.85%, 0.19% false positive rate, and 0.11% false negative rate (Emi and Spero (2024)). 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%. This lends credibility to our method.

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)[23]. This percentage varies considerably by question, ranging from 39% to 48%. The lowest share of LLM-generated answers is for the first question ("Why do you want to be a fellow?"), while the highest is for the third question ("Alumni vision"). See Appendix Figure A.1, panel c, for details.[24] Additionally, 60% of applications have at least one LLM-generated essay, and about 32% of applications are classified as fully LLM-generated according to our method.

4.2.4 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)[25]. See Appendix Figure A.2 for details. Additionally, LLM-generated essays 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

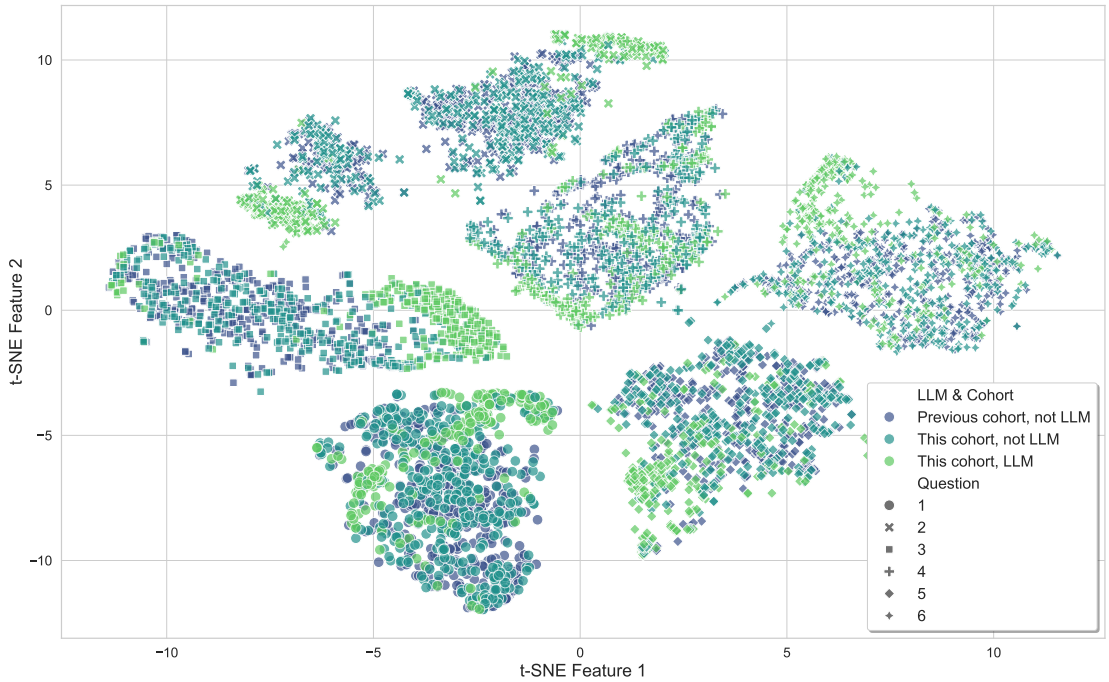
[23]This percentage is based on a 0.99 cutoff; for a 0.9 cutoff, the corresponding percentage is 50%

[24]The corresponding ranges for the 0.9 cut-off are 45% and 55%, respectively. See Appendix Figure A.1.

[25]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 = $206.835 - 1.015\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) - 84.6(\frac{\text{Total Syllables}}{\text{Total Words}})$. The Flesch reading ease score is a widely used metric for readability, and it is conveniently available in tools like Microsoft'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.

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.3).

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.4 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.

4.3 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:

$$y_i = \alpha + \beta_1 AIOnly_i + \beta_2 Human\_with\_AI\_Assistance_i + X_i'\lambda + \gamma_i + \epsilon_i \quad (2)$$

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.[26]

Additional Analysis We perform a number of additional analyses on our question-level data, the details of which are mentioned in Section 5 below.

5 Main Results

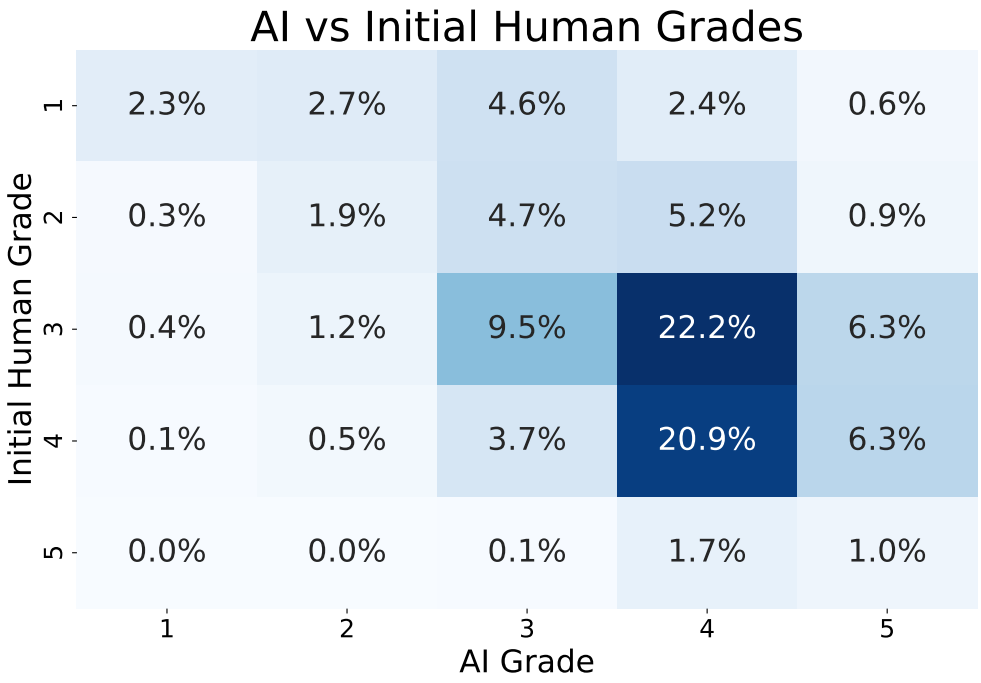
This section presents the overall effects of incorporating an AI algorithm into the organization's recruitment process. We begin with a question-level analysis and describe how AI grades differ from human grades. In Section 5.1, we show that there is substantial disagreement between human initial grades (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 5.2, we proceed with the analysis of our policy pipelines and 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 5.3. 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.

[26] 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.

5.1 Initial Human and AI Grades

Figure 5: Initial human grades vs. AI grades



	Number
Agreement rate (initial human & AI grades)	36 %
Grade difference (overall)	-0.71
Grade difference (if difference >0)	-1.11

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

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.

Human-AI Disagreement 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), which is substantial, given that the average initial human grade is around 2.9.[27] Overall, using equation (1), it is clear that on average $G_H < G_{AI}$.

Human-Human Disagreement Why would there be Human-AI disagreement? As we discuss in Section 3, there could be various forms of principal-agent problems. Human graders may provide little effort, especially as the task is complex, or they may disagree on what predicts a "good teacher" and deviate from the official grading policy. This is hard to test directly, since there is no unambiguous ground truth to benchmark against. That said, it is informative to study how two human graders view the same application. Therefore, our experimental design included double grading for a random subset of applications. Appendix Figure A.6 shows the results. There is vast disagreement. The grades across the two grading rounds are the same only in 44% of the cases.[28] This provides direct evidence that two humans do *not* follow the criteria homogeneously. Put differently, we document great heterogeneity in the human scoring function f_H.

AI-AI Disagreement To further benchmark these agreement rates by comparing them to agreement rates both within the same LLM model and across different leading LLM models, including the one used in our experiment (GPT-4 (gpt-4-0314))—see Appendix Table A.4.[29] 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. We conclude that human disagreement rates on grades for this task are relatively high.

5.2 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 even columns add demographic control variables.[30]

[27] The absolute differences are 0.9 and 1.4, respectively. There is some heterogeneity across questions (Appendix Figure A.5), with agreement rates ranging from 26% (question 4, Value Alignment) to 47% (question 6, Leadership).

[28] There is also some heterogeneity in agreement across questions (Appendix Figure A.7), with agreement rates ranging from 36% (question 1) to 53% (question 6).

[29] What is considered a state-of-the-art LLM has changed many times since our experiment was implemented.

[30] As mentioned in Section 4 above, we include only a subset of control variables in our regressions, as we do not have demographic controls for everyone.

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 higher grades on average, $G_H < G_{AI}$.

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 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.[31]

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 Figure 6 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 examining the "semantic signal" of the grade. Specifically, we employ 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. The submitted essay should predict a given score well, whereas inconsistent grading would imply giving different grades to otherwise similar candidates. 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 how we implement this. In short, the AI-only pipeline displays the highest consistency in grading, substantially greater than both the

[31]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 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)
Human-with-AI-Assistance	0.873**	0.736**	0.070	0.060
	(0.366)	(0.308)	(0.044)	(0.039)
Mean (Human-Only)	17.691	17.691	0.593	0.593
Stratum FE	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
N	697	697	697	697
p-values				
$\beta\_AI = \beta\_AIAssistance$	0.000	0.000	0.000	0.000

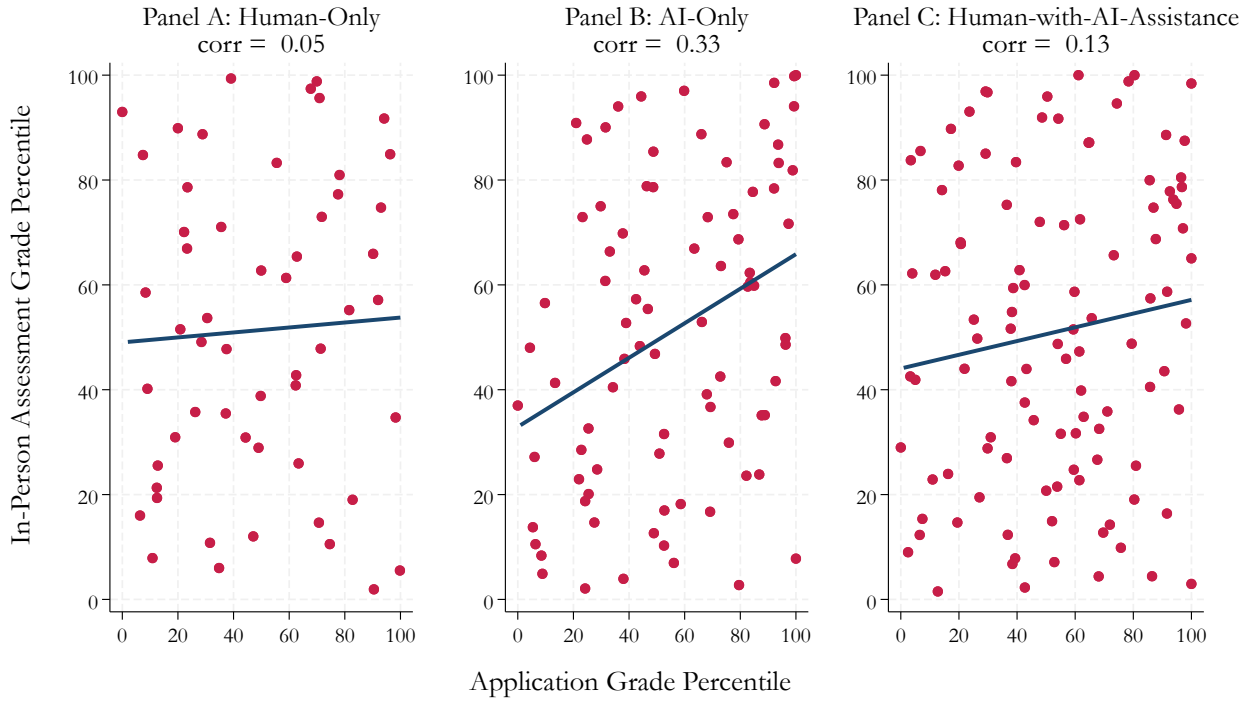
Panel B: Downstream Outcomes

	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)
Human-with-AI-Assistance	0.044	0.050	0.046	0.049	0.015	0.019
	(0.042)	(0.040)	(0.038)	(0.038)	(0.033)	(0.033)
Mean (Human-Only)	0.284	0.284	0.206	0.206	0.149	0.149
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes
N	697	697	697	697	697	697
p-values						
$\beta\_AI = \beta\_AIAssistance$	0.001	0.003	0.002	0.004	0.013	0.024

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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

human-only and the assistance pipeline do. This evidence therefore implies that GPT-4 appears to have consistently applied the grading rubric as defined by the organization, whereas in both conditions with humans-in-the-loop there are deviations from the grading rubric. As we will show

Figure 6: 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.

in the next section, these deviations were not purely random but instead systematic, based on whether applicants had used LLMs to generate the essays. Human evaluators consistently gave such applicants lower scores, despite the fact this was not a part of the grading rubric as defined by the organization, and what the human graders had been tasked to implement.

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)
Human-with-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			
$\beta\_AI = \beta\_AIAssistance$	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). Note that because many candidates have the same application score, the number of candidates in each bin is greater than n. Also note 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

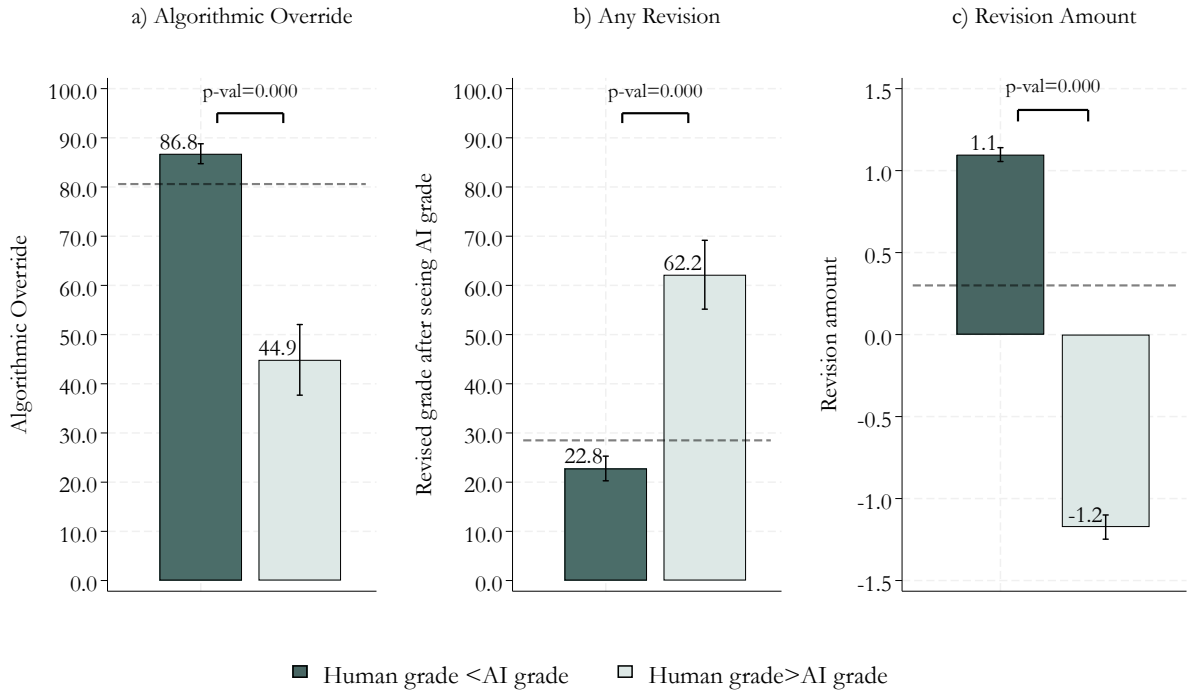
5.3 AI-Assistance

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 algorithmic override as any case where the final human grade is not equal to the algorithmic recommendation. Figure 7 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 (Panel a, dashed line). They override the recommendation more often when the AI grade is above 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 (Panel b, dashed line), but they often do not adjust the grade all the way to the AI grade. Evaluators are significantly more likely to adjust their grade down than up. Specifically, evaluators revise in approximately 22.8% of 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.[32] AI assistance therefore raises the agreement rate from 36% to about 47.7%, an increase of about 33% (depicted in Appendix Figures A.8, A.9, and A.10). 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.

Figure 7: Algorithmic Override and Grade Revisions by Initial Grade Disagreement



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

[32]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.

5.4 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 not 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 6). 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.

5.5 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. We conclude that this is consistent with AI assistance reducing productivity on this task.

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

The results of our policy experiment, presented in Section 5, 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 AI is used as an assistant. Why were these recommendations ignored in 80% of the cases? Are there systematic pattern of behaviors, which help explain why the evaluators deemed the AI-assistant incapable of distinguishing between high- and low-quality candidates based on the input, despite the fact that having simply followed its advice would have led to greater hiring success with potentially much lower effort? In this section, we provide evidence of the central role that LLM-generated essays play in explaining our findings.

6.1 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.

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),[33] humans award a smaller premium to 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.[34] 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.

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, Munyikwa, and Horton, 2025). In contrast, if human graders assign relatively lower grades to LLM-generated essays, it suggests they may hold negative perceptions

[33] 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.

[34] The results are robust to using different cut-offs for classifying the essays as being LLM-generated, see Appendix Figure A.9.

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

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***	0.069***	0.091***	0.075***	-0.084***	-0.062**	-0.184***	-0.147***
	(0.023)	(0.024)	(0.022)	(0.022)	(0.026)	(0.025)	(0.040)	(0.042)
Mean (non-LLM)	0.497	0.497	0.772	0.772	0.314	0.314	-0.557	-0.557
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	1,968	1,968	1,265	1,265	1,265	1,265	1,968	1,968

Notes: Columns 1-6 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 4).

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 difference between final human and AI grades is about 31% higher (Column (8))[35]. 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.12). 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 Human-with-AI-Grade-and-Rationale treatment group.[36] 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.

6.2 LLM-Applications and Downstream Outcomes

In Section 6.1, 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.

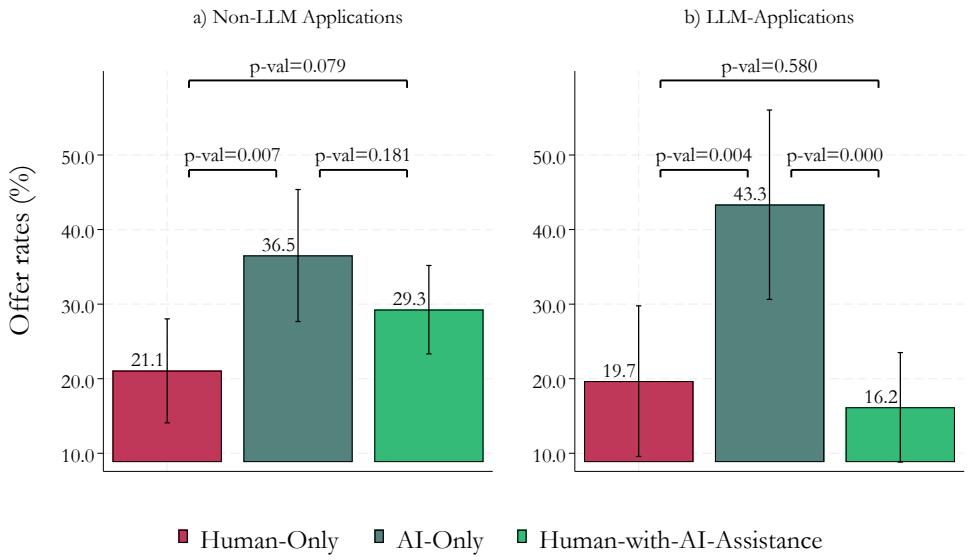
Figure A.16 presents the main findings and Table 5 presents these results in a regression format. In the Human-Only pipeline, the likelihood of receiving a fellowship offer is nearly identical for LLM-generated and non-LLM-generated applications (19.7 vs. 21.1% respectively). However, there is a striking 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 [37]. Columns (1) and (3) replicate the findings shown in Figure A.16, while Columns (2) and (4)

[35]The results are very similar when we use alternative cut-offs for classifying the essays as being LLM-generated. See Appendix Tables A.10 and A.11.

[36]The results are similar when we use the two alternative classification cutoffs, 90% and 95%, see Appendix Figures A.13 and A.14.

[37]The results are qualitatively the same when using the alternative cut-offs to classify applications as being LLM-generated (see Appendix Tables A.14 and A.13)

Figure 8: Offer Rates by Pipeline and LLM-Application



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

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.

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								
$\beta\_AI{=}\beta\_AIAssistance$	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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

7 Conclusion

As GenAI technologies achieve mass adoption, policy makers and firms are reconsidering whether such systems should augment human workers or take over tasks entirely. To inform this debate, we embed GPT-4 into the hiring process of an organization recruiting teachers in Ghana and find that fully automating the screening process increases downstream hiring success by 73% relative to the human-only baseline. However, using GPT-4 as an assistant provides no meaningful improvement to either outcomes or productivity. Evaluators frequently override algorithmic recommendations because many applicants used GenAI tools to prepare their materials, a factor the algorithm did not account for and which evaluators viewed as a negative signal of candidate quality.

While we find evidence in favor of automation, these results should be interpreted with caution. Our findings come from a setting in which GenAI was a relatively new technology, before mass adoption had occurred, so they should not be viewed as static. Labor market conditions, GenAI capabilities, and perceptions of these technologies are all evolving. While we capture some early dynamics over time, the long-term effects remain unclear. For instance, once GenAI reaches widespread adoption, the entire labor markets could shift (Raymond, 2024), prompting changes in both applicant strategies and employer practices. Thus, while our evidence points to the potential for automation using GenAI in economically important tasks, much remains to be understood. A key challenge for research in the coming years will be to find ways to design and train GenAI

algorithms to be human-complementary, in general as well as when it comes to decision making that improves matching in labor markets.

References

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

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

Agan, Amanda Y., Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. 2023. "Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias." *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. 2024. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." March, https://papers.ssrn.com/abstract=4505053.

Angelova, Victoria, Will Dobbie, and Crystal Yang. 2023. "Algorithmic Recommendations and Human Discretion." September, https://papers.ssrn.com/abstract=4589709.

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 Raymond. 2025. "Generative AI at Work*." *The Quarterly Journal of Economics* 140 (2): 889–942. 10.1093/qje/qjae044.

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

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." August. 10.48550/arXiv.2303.10130, arXiv:2303.10130 [econ].

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.

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

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

Insight, Global. 2024. "2025 AI in Hiring Survey Report."Technical report, Insight Global, https://insightglobal.com/2025-ai-in-hiring-report/.

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

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."

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

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

Otis, Nicholas, Rowan Clarke, Solène Delecourt, David Holtz, and Rembrand Koning. 2024. "The Uneven Impact of Generative AI on Entrepreneurial Performance." February. 10.2139/ssrn.4671369.

Ouyang, Long, Jeff Wu, Xu Jiang et al. 2022. "Training language models to follow instructions with human feedback." In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22 27730–27744, Red Hook, NY, USA: Curran Associates Inc., , November.

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].

Raymond, Lindsay. 2024. "The Market Effects of Algorithms | Department of Economics." https://economics.stanford.edu/events/market-effects-algorithms.

Thomas, Huw. 2025. "AI job application rise risks employing incapable staff, boss warns." March, https://www.bbc.com/news/articles/cx29z8lyx71o.

Vafa, Keyon, Ashesh Rambachan, and Sendhil Mullainathan. 2024. "Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function." June. 10.48550/arXiv.2406.01382, arXiv:2406.01382 [cs].

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 Munyikwa, and John Horton. 2025. "Algorithmic Writing Assistance on Jobseekers' Resumes Increases Hires." *Management Science*. 10.1287/mnsc.2024.04528, Publisher: INFORMS.

Online Appendix

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

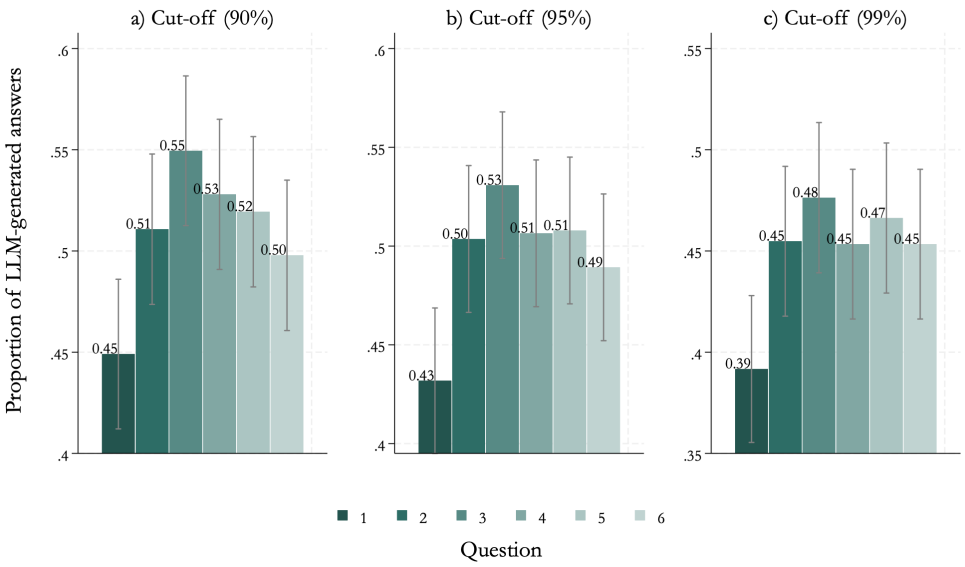
Notes: The Table displays summary statistics for the overall experimental sample. Panel A displays question-level summary statistics from our grading "experiment", and Panel B displays application-level summary statistics from our policy experiment. The outcome variables are defined in Section 4.2.1.

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

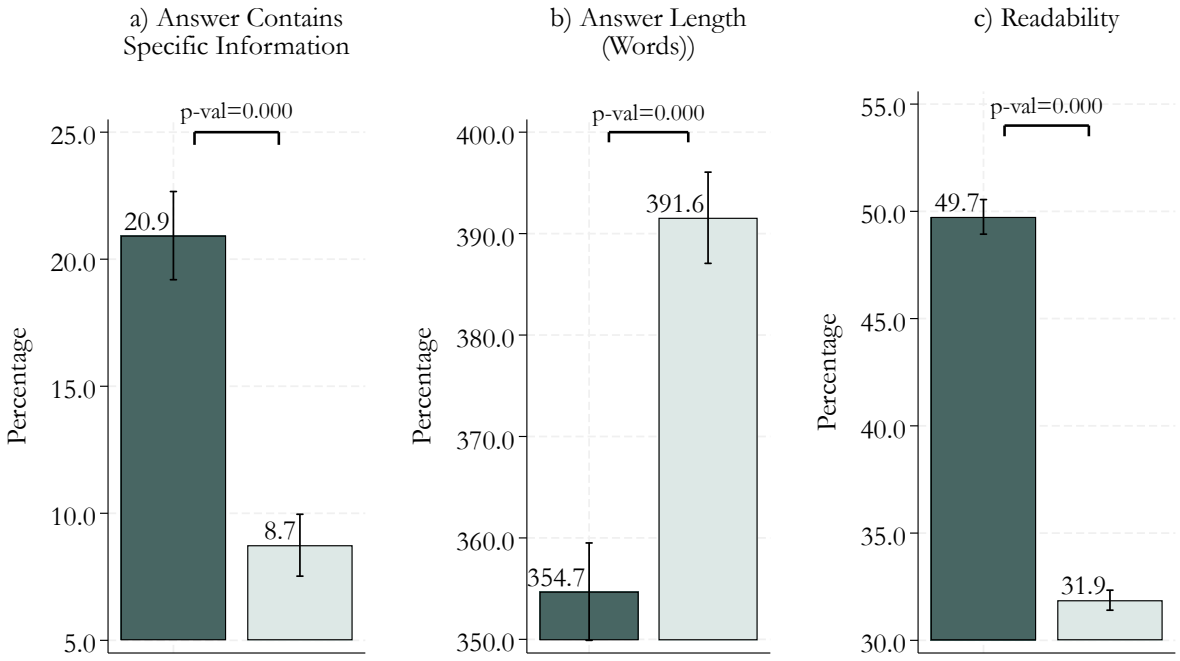
Notes: The figure shows the balance table for our policy experiment. Last two columns (under "Joint") report the F-statistic and the p-value from a joint test of significance of the set of treatment dummies in explaining each row variable in a regression with strata (week) fixed effects included and 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.

Figure A.1: How common are LLM-Generated Essays (by question)?



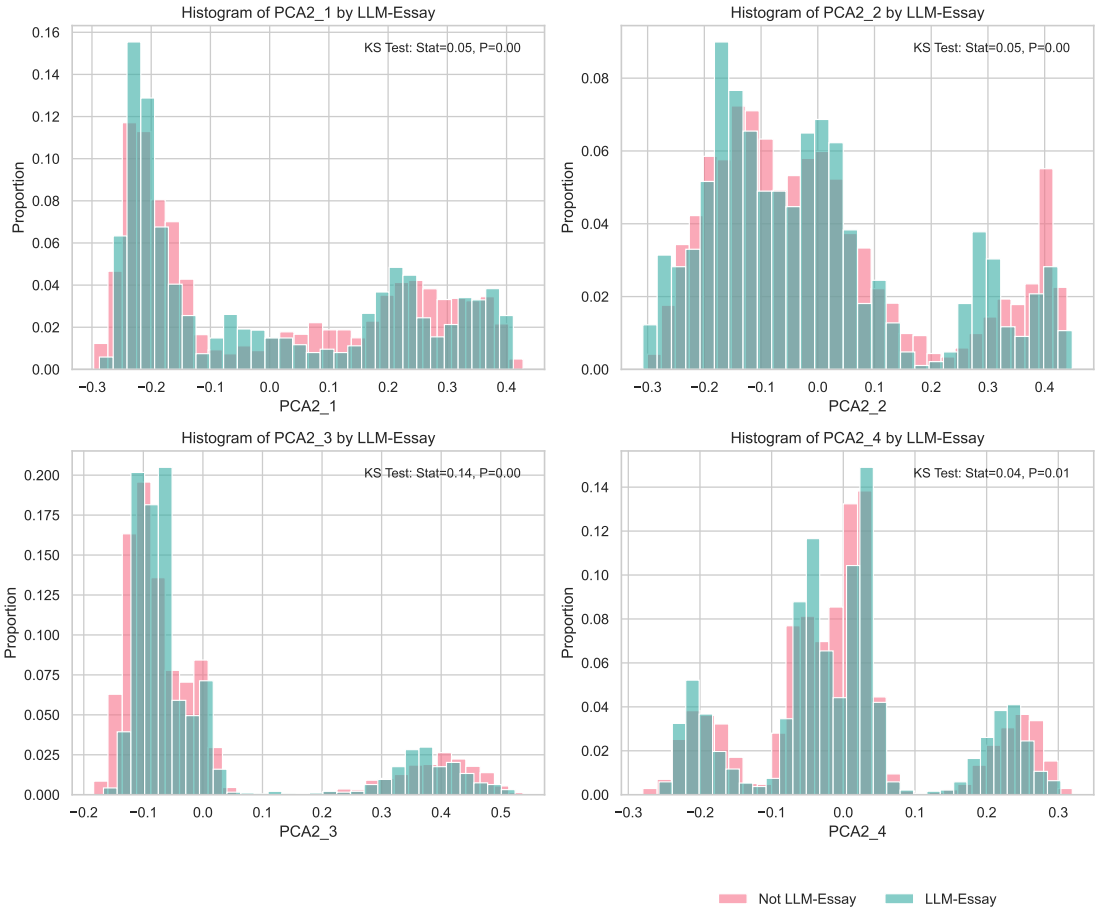
Notes: The figure shows the proportion of answers classified as LLM-generated for each question, using a three different probability cut-offs based on the Pangram Text model's estimates.

Figure A.2: Characteristics of AI-generated answers



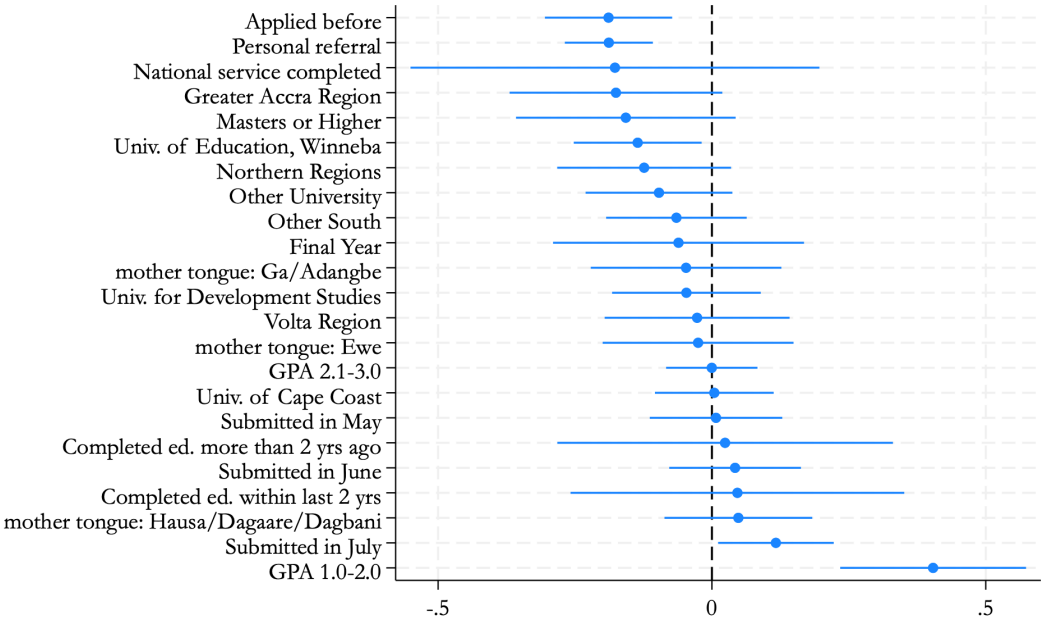
Notes: The figure depicts the characteristics of LLM- and non-LLM-essays. Panel a: Proportion of answers that contain specific information (for example on applicant's gender or university). Panel b: Answer length in words. Panel c: The complexity as measured by the Flesch reading ease (Flesch, 1948), 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.3: Is Semantic Content Different Across LLM and Non-LMM answers?



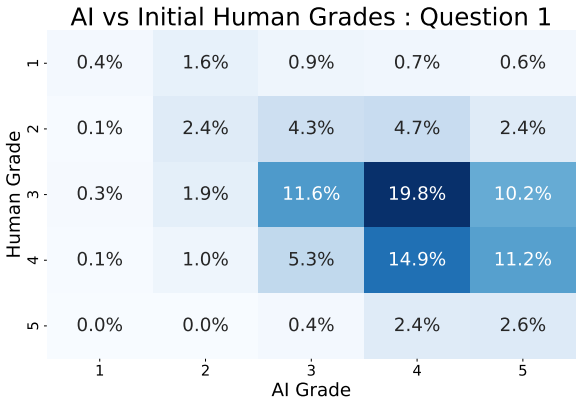
Notes: The figure depicts the distribution of first four principal components (out of 10 that were generated) of the vector embeddings that were generated using "voyage-lite-02-instruct" model from Voyage AI for LLM- and non-LLM-essays, and the Test statistic and the p-value of the Komolgorov-Smirnov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

Figure A.4: What Predicts LLM Usage?

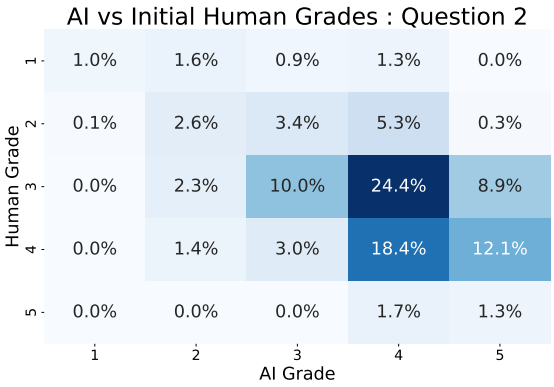


Notes: The figure displays coefficients from an OLS regression at the application level of mean of the question-level likelihood of being LLM-generated on different demographic controls, for a subset of people for whom we have all these controls available.

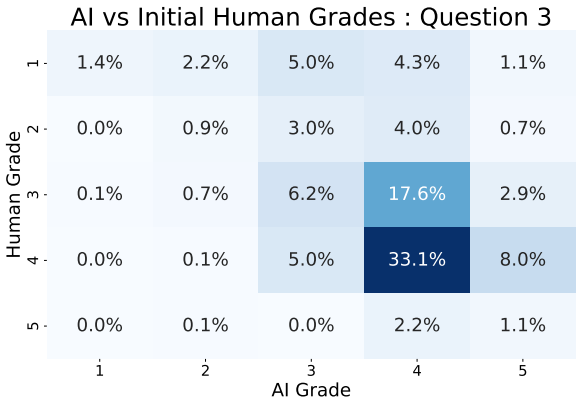
Figure A.5: Initial Human Grades vs. AI Grades by Question



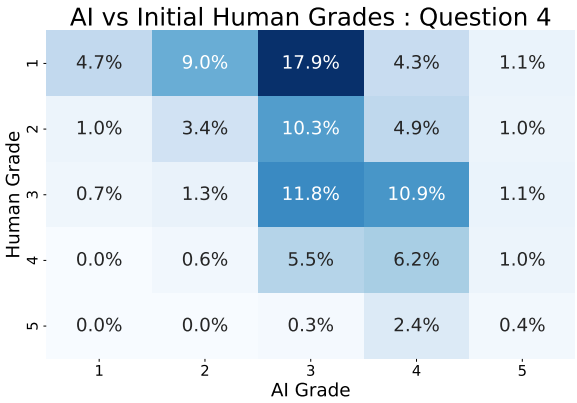
a) Question 1



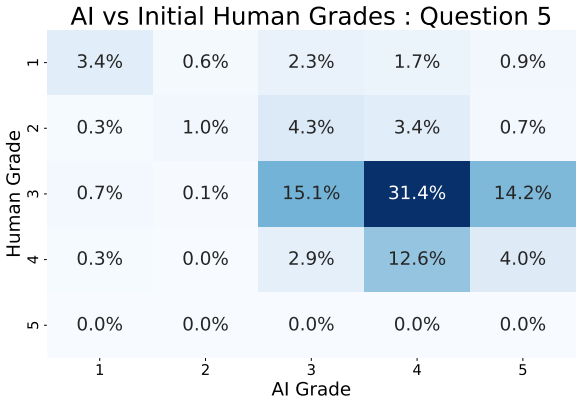
b) Question 2



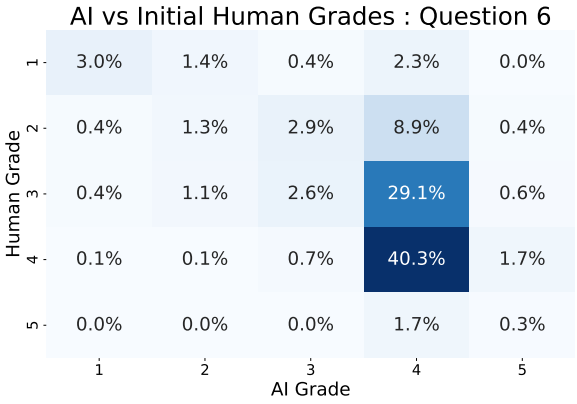
c) Question 3



d) Question 4



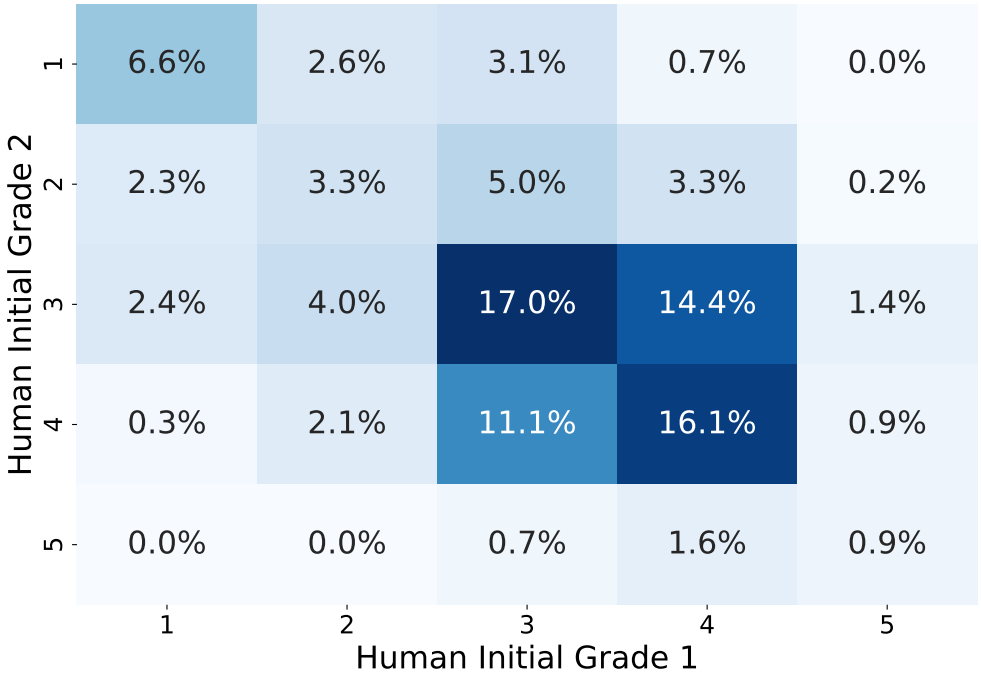
e) Question 5



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.

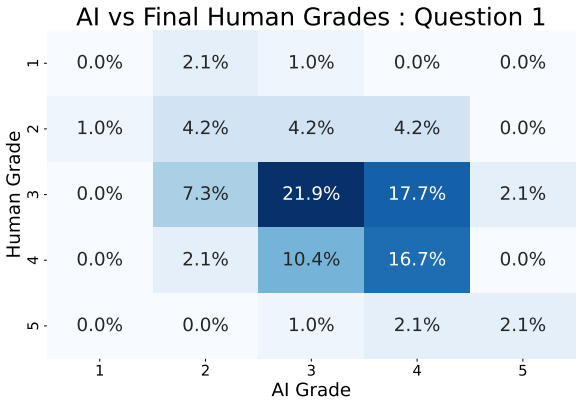
Figure A.6: Initial Human Grades Consistency



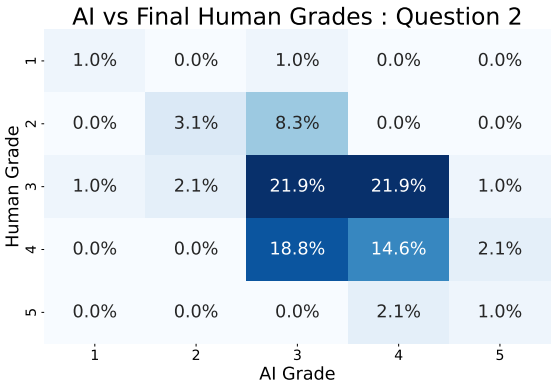
	Number	Percentage
Human grade 1!=Human grade 2	323	56
Human grade 1=Human grade 2	253	44

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.

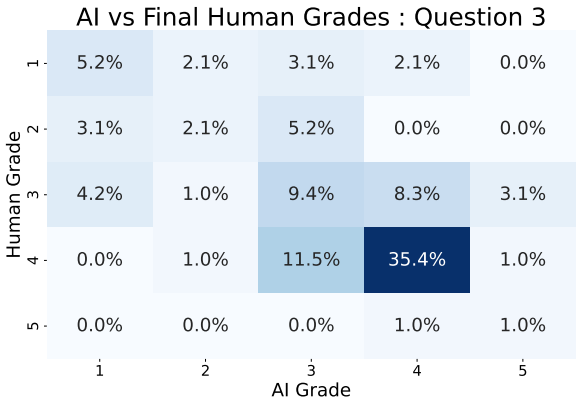
Figure A.7: Initial Human Grade Consistency by Question



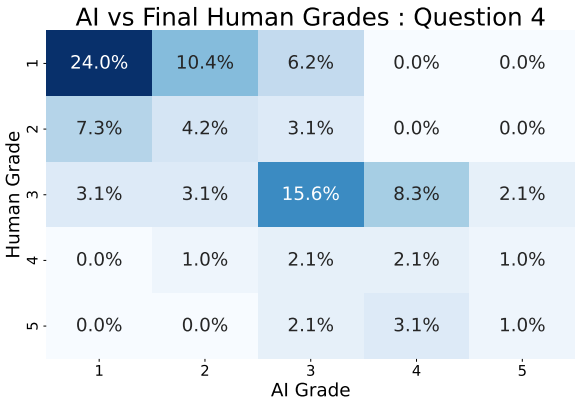
a) Question 1



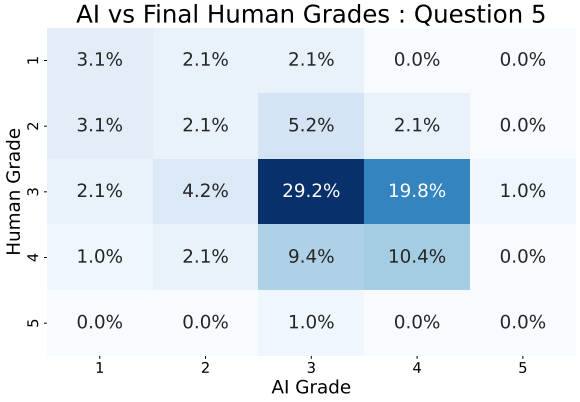
b) Question 2



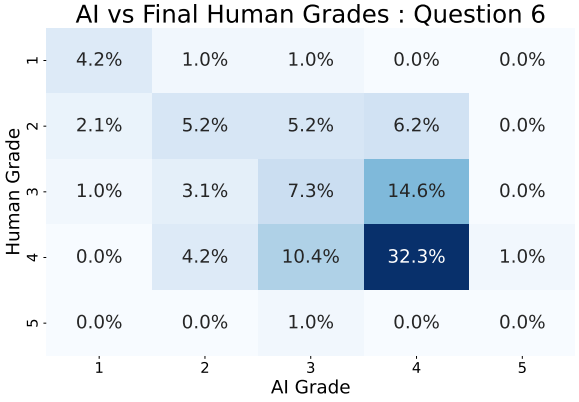
c) Question 3



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.

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)

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

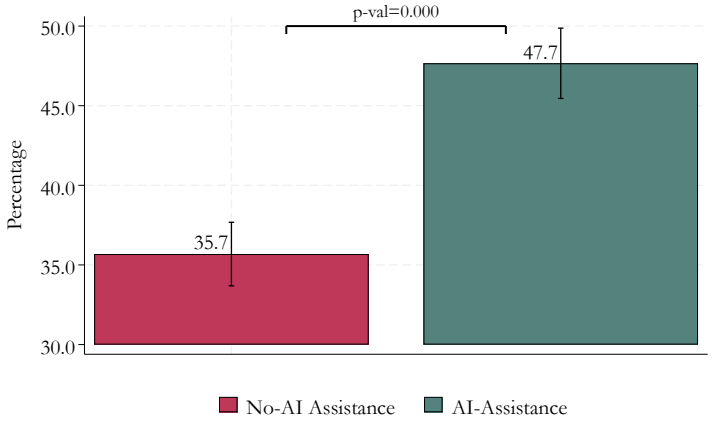
Notes: Columns 1-4 report estimated coefficients from OLS regressions respectively of an indicator variable for whether the applicant was advanced to the assessment centre (column 1), attended the assessment center, conditional on being advanced to the assessment center (column 1), received a job offer (column 2) and was hired, that is accepted the job offer (column 4). 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

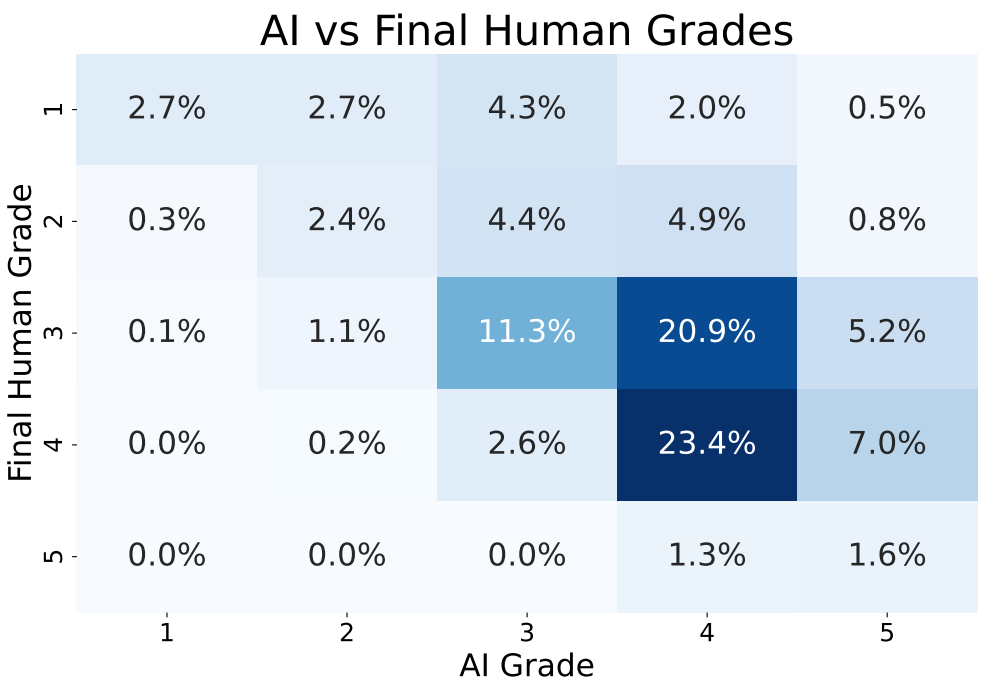
Notes: Columns 1-2 report estimated coefficients from OLS regressions of total in-person assessment grades on initial human total applications grades and the total AI grades, for people who were advanced to, and attended the in-person assessment. All columns include stratum (week) fixed effects; columns 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.8: Average Agreement in Final Grades



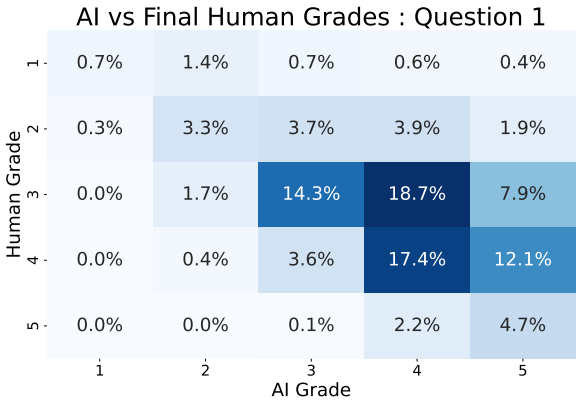
Notes: The figure shows the proportion of questions where the final human grade matched the AI grade, separately by whether the application was assigned AI-assistance. Error bars indicate the 95% confidence intervals. p-values are calculated from a t-test from a regression of a binary indicator for grade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

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

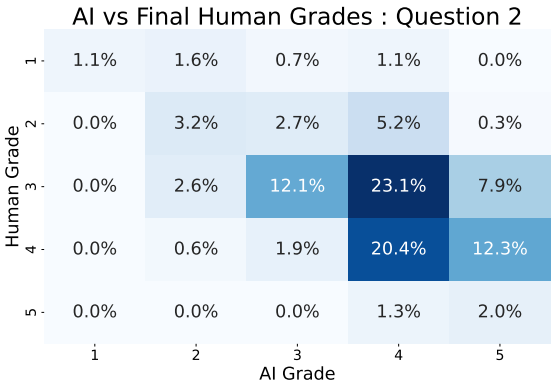


Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement frequencies between final human and AI grades (both ranging from 1 to 5). The diagonal (top-left to bottom-right) indicates complete agreement. Areas below (above) the diagonal represent cases where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

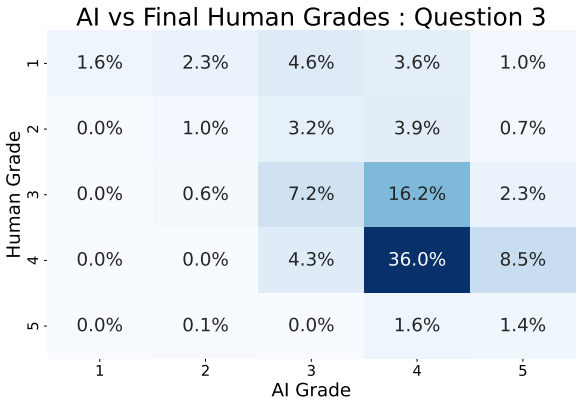
Figure A.10: Final Human Grades vs. AI Grades by Question



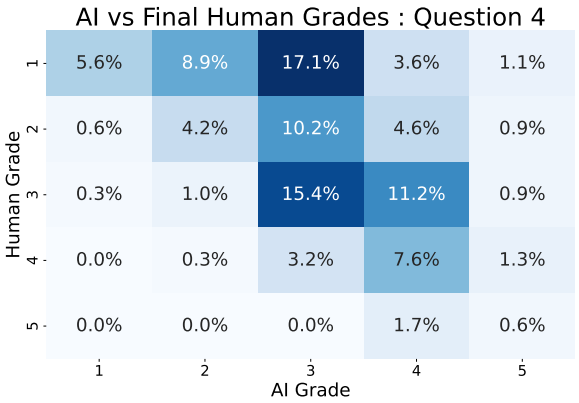
a) Question 1



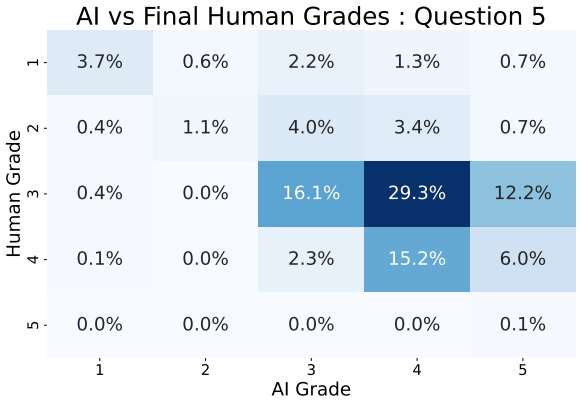
b) Question 2



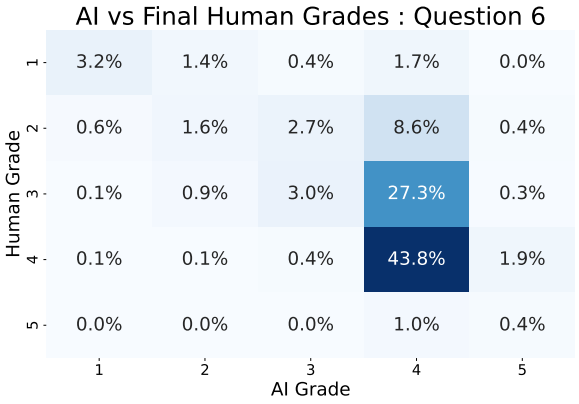
c) Question 3



d) Question 4



e) Question 5



f) Question 6

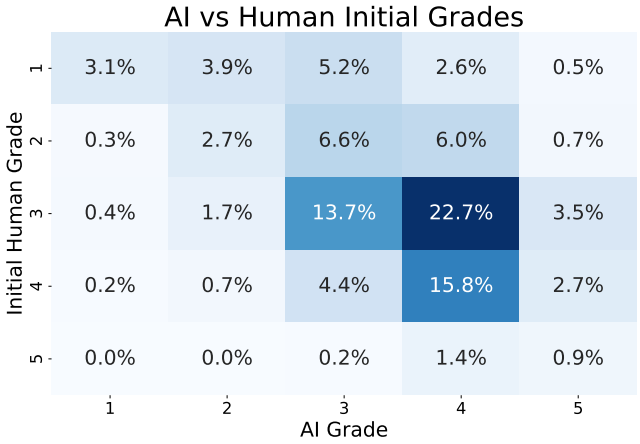
Notes: The matrix depicts the distribution of grades across a 5x5 grid, where cells represent agreement frequencies between final human and AI grades (both ranging from 1 to 5), separately for each question. The diagonal (top-left to bottom-right) indicates complete agreement. Areas below (above) the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

Table A.7: Time Spent on Application

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

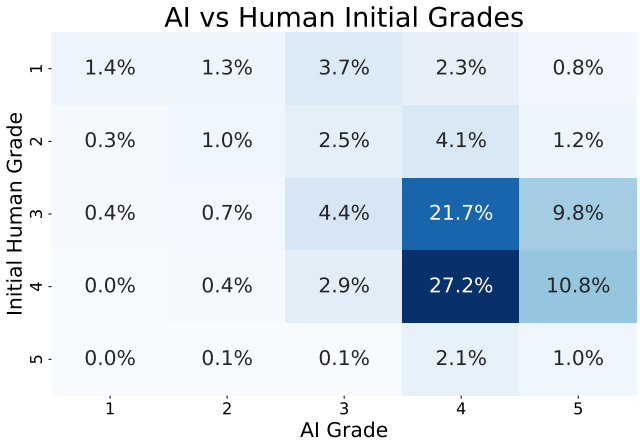
Notes: Columns 1-6 report estimated coefficients from OLS regressions of log of time (in seconds) spent grading questions. Columns 1-3 represent time up to the initial grade, and columns 4-6 represent time up to the final grade. For the group without AI assistance, times to initial and final grades are equal. All columns include stratum (week) fixed effects; columns 2 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 95th percentile on question-level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 %
Grade difference (overall)	-0.66
Grade difference (if difference >0)	-1.04



b) LLM-generated

	Number
Agreement rate (initial human & AI grades)	35 %
Grade difference (overall)	-0.77
Grade difference (if difference >0)	-1.19

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

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

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

Table A.10: Robustness Check (95% Likelihood Cut-Off): Human Graders Override the Algorithm More When Grading LLM-Written Essays

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

Notes: This table reports a robustness check for an alternative probability cut-off (95%) used to classify an essay as LLM-generated. Columns 1-6 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 4).

Table A.11: Robustness Check (90% Likelihood Cut-Off): Human Graders Override the Algorithm More When Grading LLM-Written Essays

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

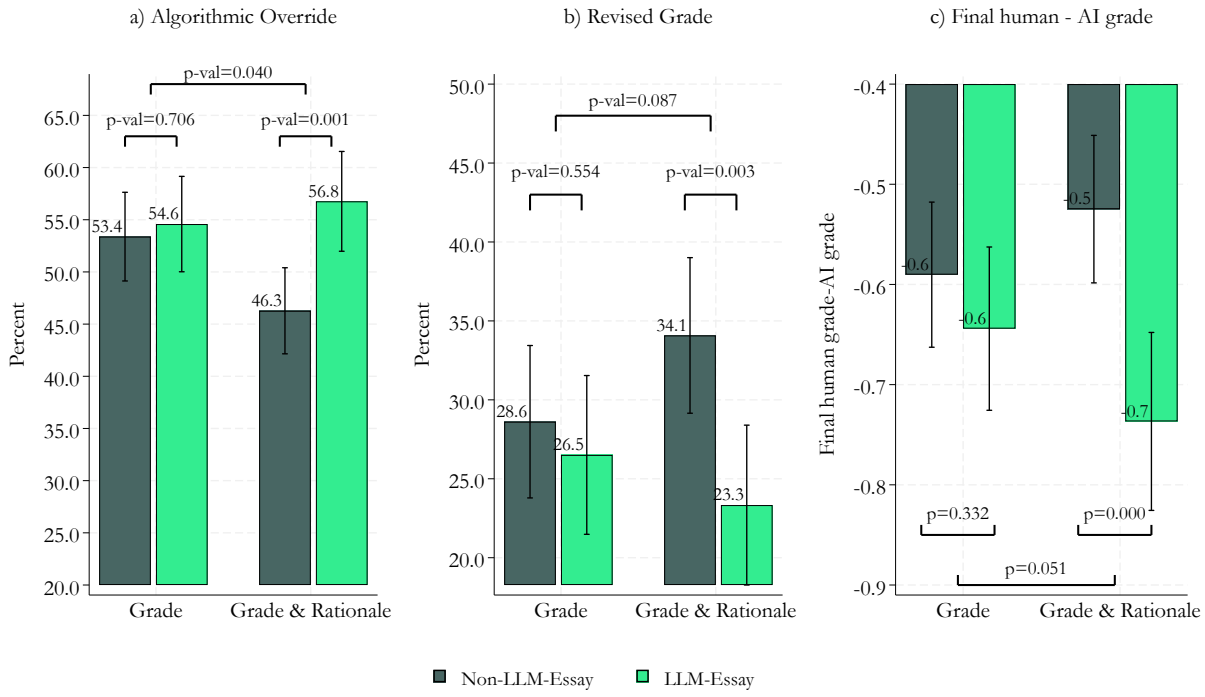
Notes: This table reports a robustness check for an alternative probability cut-off (90%) used to classify an essay as LLM-generated. Columns 1-6 report estimated coefficients from OLS regressions respectively of algorithmic override (final human grade differs from AI grade) overall (Columns (1) and (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 4).

Table A.12: 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

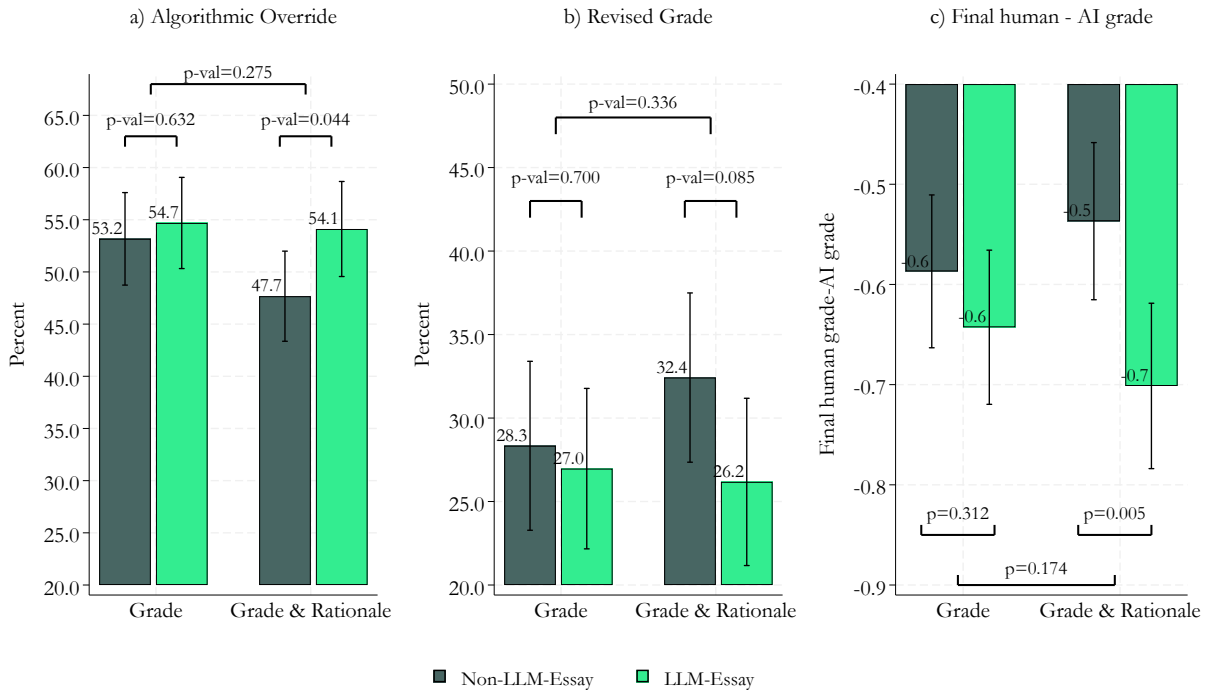
Notes: Columns 1-6 report estimated coefficients from OLS regressions respectively of algorithmic override (final human grade differs from AI grade) overall (Columns (1) and (2)) and when thre is initial grade disagreement (Columns (3) and (4)) and the difference between final human and AI grades (Columns (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



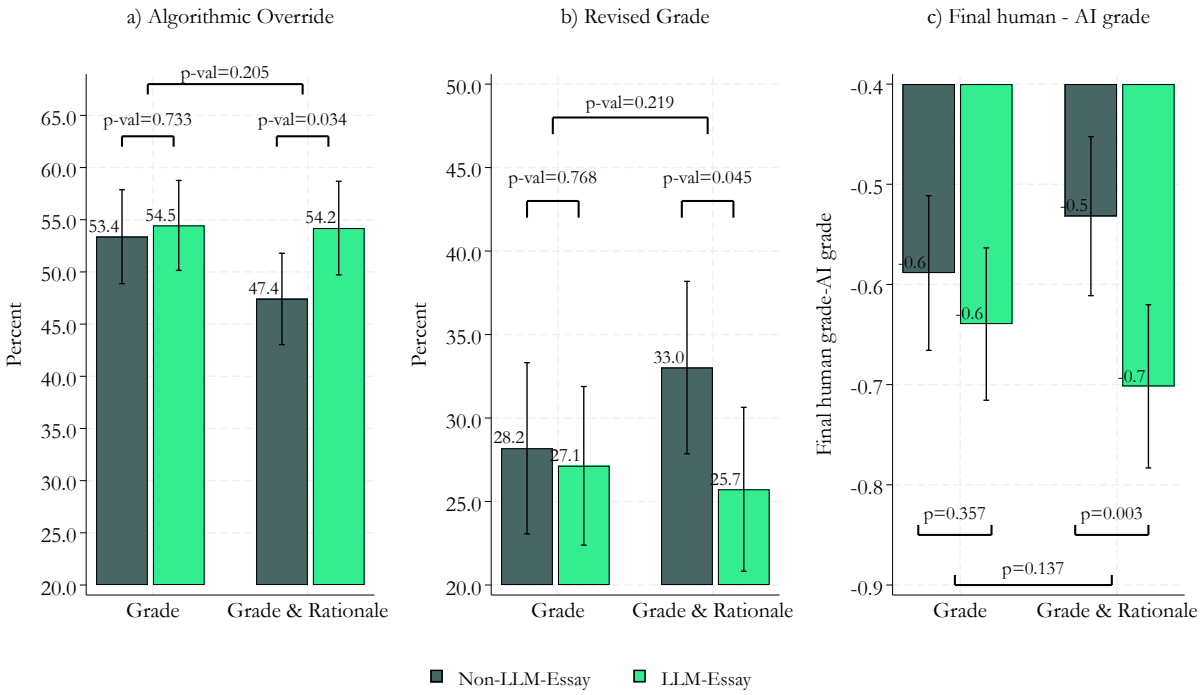
Notes: The figure depicts differences in algorithmic override (Panel a), initial grade revision (Panel b), and the difference between final human and AI grade(Panel c) for the two different types of AI assistance–AI grade & AI grade with rationale. p-values come from t-tests of equality of means.

Figure A.13: Robustness Check (95% cutoff): Algorithmic Override, Grade Revision, and Differences in Final Human and AI Grades by the Type of AI Assistance



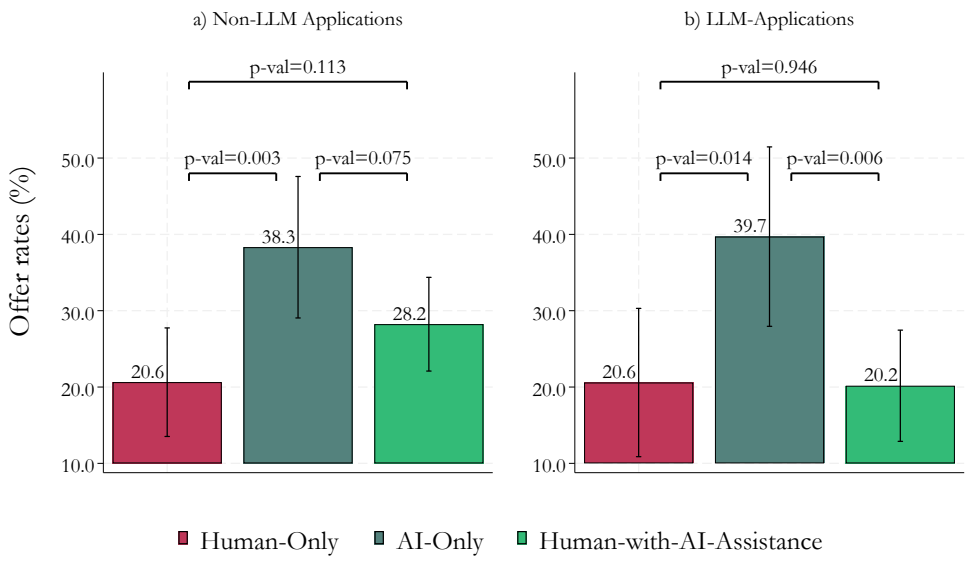
Notes: The figure depicts differences in algorithmic override (Panel a), initial grade revision (Panel b), and the difference between final human and AI grade(Panel c) for the two different types of AI assistance–AI grade & AI grade with rationale. p-values come from t-tests of equality of means.

Figure A.14: Robustness Check (90% cutoff):Algorithmic Override, Grade Revision, and Differences in Final Human and AI Grades by the Type of AI Assistance



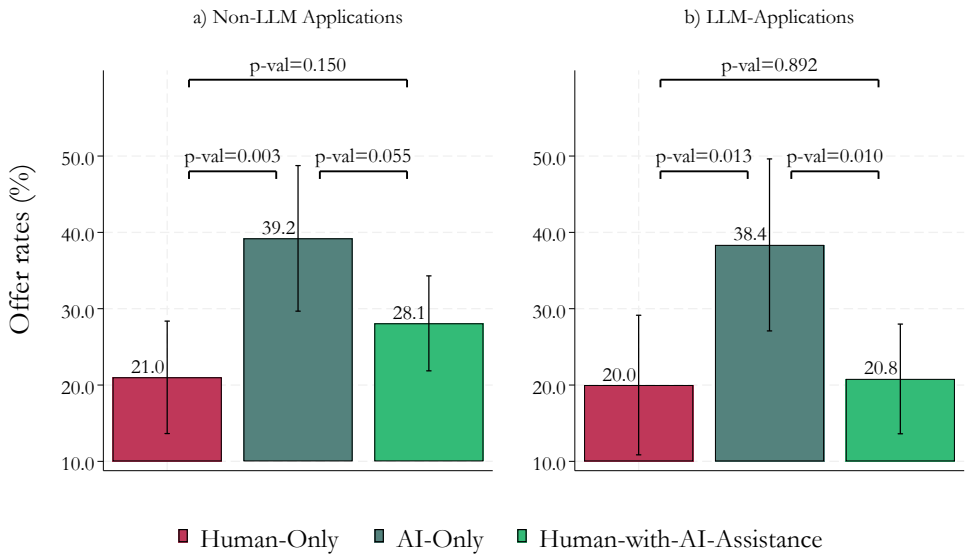
Notes: The figure depicts differences in algorithmic override (Panel a), initial grade revision (Panel b), and the difference between final human and AI grade(Panel c) for the two different types of AI assistance–AI grade & AI grade with rationale. p-values come from t-tests of equality of means.

Figure A.15: Robustness Check (95% cut-off): Offer Rates by Pipeline and LLM-Application



Notes: The figure shows regression coefficients from equation 2 without control variables run separately for a subsample of non-LLM- and LLM-applications, where the outcome variable is a binary indicator of whether a candidate received a job offer. The cranberry bar represents the constant term, i.e. mean offer rates in the Human-Only pipelines, the emerald and mint bars represent the sum of the constant and the respective beta coefficients. Error bars indicate 95% confidence intervals derived from standard errors of the linear combinations of the constant and the beta regression coefficients. We report two sets of p-values from our regression: one from a t-test evaluating whether the beta coefficient is statistically different from zero (lower p-values), and another from a t-test testing for equality of the beta coefficients between AI-Only and Human-with-AI-Assistance pipelines (upper p-values).

Figure A.16: Robustness Check (90% cut-off): Offer Rates by Pipeline and LLM-Application



Notes: The figure shows regression coefficients from equation 2 without control variables run separately for a subsample of non-LLM- and LLM-applications, where the outcome variable is a binary indicator of whether a candidate received a job offer. The cranberry bar represents the constant term, i.e. mean offer rates in the Human-Only pipelines, the emerald and mint bars represent the sum of the constant and the respective beta coefficients. Error bars indicate 95% confidence intervals derived from standard errors of the linear combinations of the constant and the beta regression coefficients. We report two sets of p-values from our regression: one from a t-test evaluating whether the beta coefficient is statistically different from zero (lower p-values), and another from a t-test testing for equality of the beta coefficients between AI-Only and Human-with-AI-Assistance pipelines (upper p-values).

Table A.13: Robustness Check (Cut-off 95%) LLM-Applications and Downstream Outcomes

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

Notes: 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.14: Robustness Check (Cut-off 90%) LLM-Applications and Downstream Outcomes

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

Notes: 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

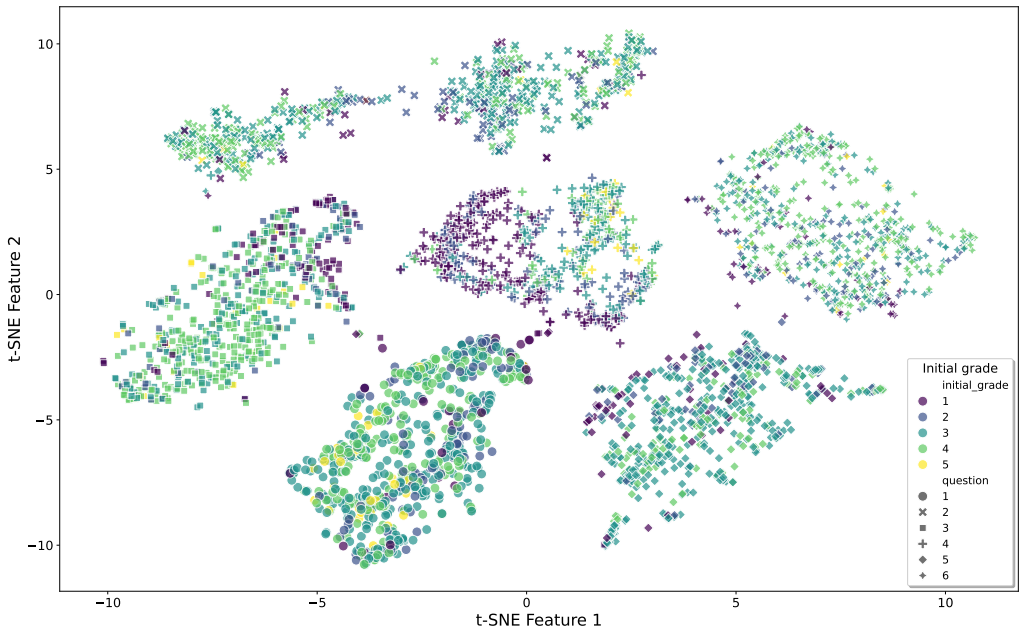
B Signal in Grades

This section presents the details of how we constructed the "semantic signal" variable mentioned in Section 5.

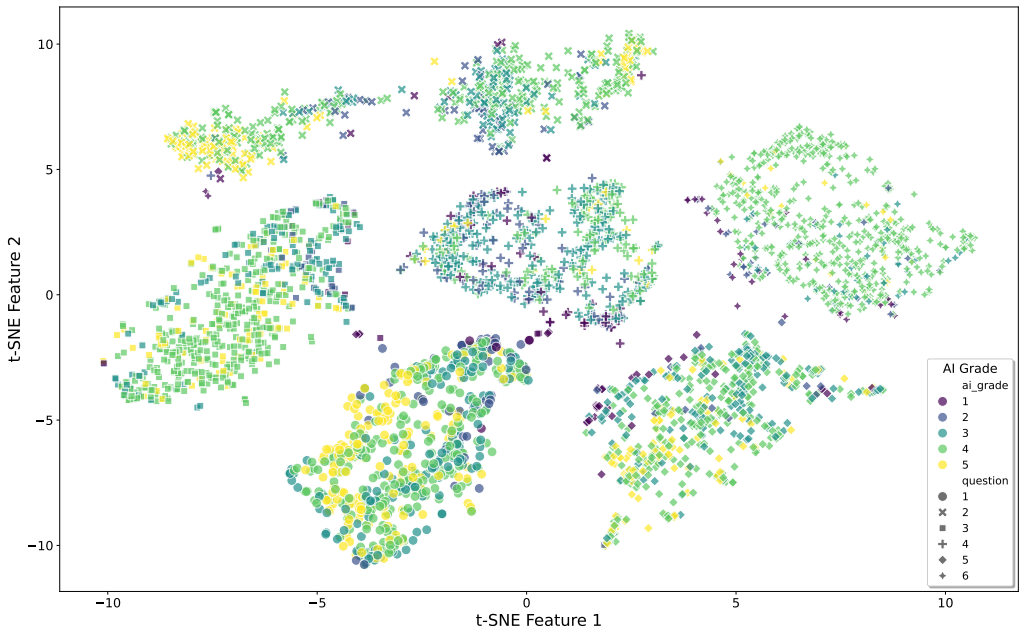
Vector Embeddings of the Essay Answers We first converted each essay answer into vector embeddings using the "voyage-lite-02-instruct" model from Voyage AI. The original 1024-component embedding vectors were first condensed to 50-dimensions using a PCA (Principal Component Analysis) reduction. Next, 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.17. 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.17 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.17: 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 3:

$$\theta_{ijq} = \alpha + \beta_1 SameScore_{ijq} + \beta_2 SameScoreXAIAssistance_{ijq} + \beta_3 SameScoreXAIOnly_{ijq} + \gamma_1 AIAssistance_{ijq} + \gamma_2 AIOnly_{ijq} + X'_{ijq}\lambda + Q_{ijq} + \epsilon_{ijq} \quad (3)$$

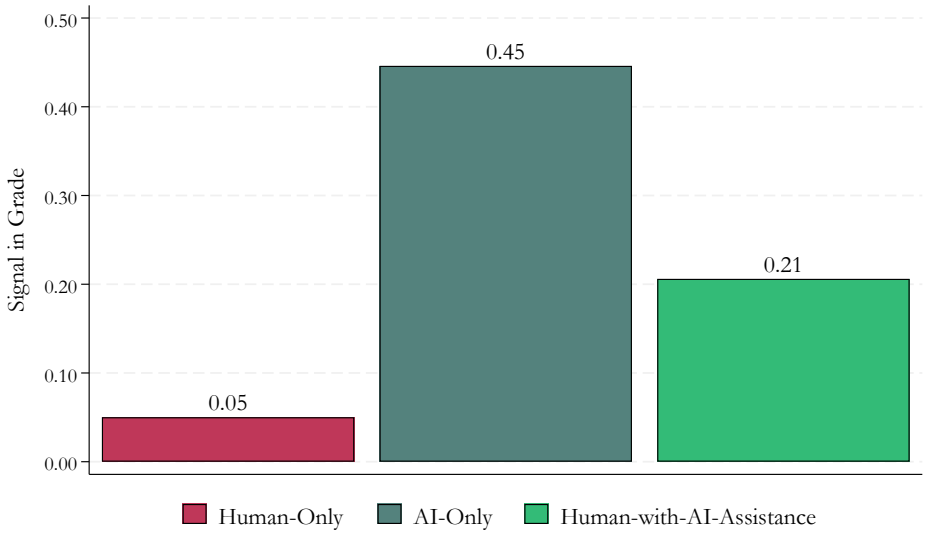
where θ_{ijq} are pairwise similarity scores, Q_{ijq} contains question fixed effects, and X_{ijq} is a vector of control variables that includes grades of texts i and j. Our main coefficients of interest are β_2 and β_3 which tell us how much more similar are texts *within* the same grades for grades 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.18 visualizes the measure of total signal constructed from the coefficients presented of estimating equation 3. Figure B.18 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.

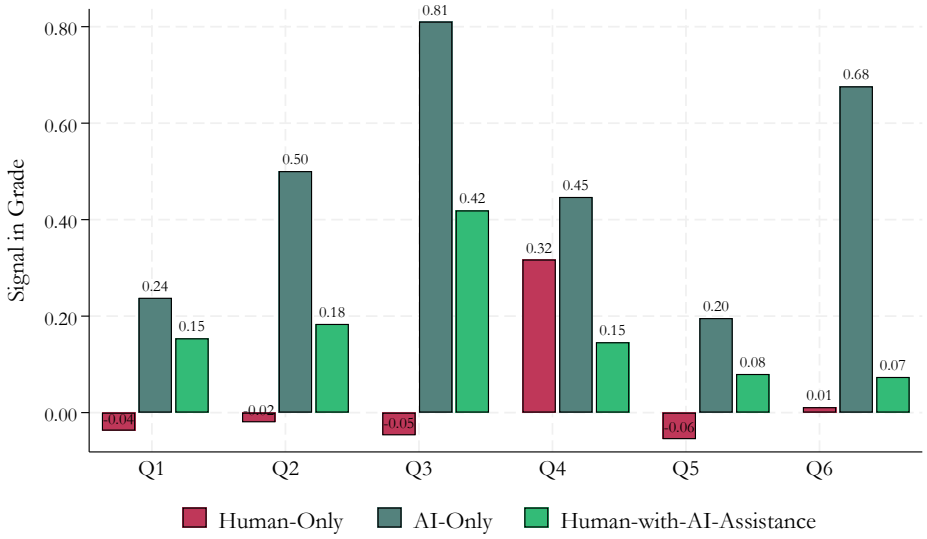
For an alternative measure of the amount of "signal" contained in grades, we train a random forest classifier on an 80% sample of the question answers to predict the grades on the rest of the sample. We then look at the model's performance metrics overall and for each grade. The idea is that when there is more text-based signal contained in each grade, the model will perform better. Table B.15 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

Figure B.18: 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 3 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.

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 performance.

Table B.15: Random Forrest Classifier

Panel A: Initial Grades

Grade	Precision	Recall	F1 Score	Observations	Accuracy
1	0.663	0.538	0.594	106	
2	0.438	0.064	0.111	110	
3	0.477	0.666	0.556	332	
4	0.491	0.502	0.496	265	
5	1.000	0.042	0.080	24	
Weighted Average	0.515	0.501	0.470	837	
Accuracy					0.501

Panel B: AI Grades

Grade	Precision	Recall	F1 Score	Observations	Accuracy
1	0.833	0.192	0.312	26	
2	0.250	0.019	0.034	54	
3	0.470	0.368	0.413	190	
4	0.619	0.898	0.733	440	
5	0.725	0.228	0.347	127	
Weighted Average	0.584	0.597	0.544	837	
Accuracy					0.597

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.

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+++"
```