Voice AI in Firms:
A Natural Field Experiment on Automated Job Interviews

Brian Jabarian Luca Henkel

January 26, 2026

Click here for the most recent version

Abstract

This paper studies whether AI automation can improve organizational outcomes by reducing variance when collecting information. We conducted a large-scale natural field experiment in which 70,000 job applicants were randomly assigned to be interviewed by human recruiters or AI voice agents. In both conditions, human recruiters evaluate the interviews and make hiring decisions. Applicants interviewed by AI agents are 12% more likely to receive job offers, and these gains translate into higher job starts and worker retention, with no decline in the productivity of hired workers. Analyzing interview transcripts reveals that AI voice agents achieve *controlled variance*: their interviews are more structured and consistent while remaining responsive to individual applicants, which is associated with more hiring-relevant information collected. These results demonstrate that automating information collection with AI can enhance decision quality through standardization.

Keywords: Artificial intelligence, interviews, hiring, organizational design, field experiment
JEL Classification: C93, J24, M15, M51, O33

Contact: Brian Jabarian (Corresponding author): Booth School of Business, University of Chicago, CDR, CAAI, CESifo, J-PAL, IGL, JI-LAEE, brian.jabarian@chicagobooth.edu. Luca Henkel: Erasmus University Rotterdam, ERIM, CESifo, JILAEE, IZA, Tinbergen Institute; henkel@ese.eur.nl.
Acknowledgements: We thank Peter Andre, Iav Bojinov, Joshua Deutschmann, Drew Fudenberg, Avi Goldfarb, Mitchell Hoffman, John Horton, Alex Imas, Matthew Jackson, Jon Kleinberg, Hyunjin Kim, Rem Koning, Andrew Koh, David Laibson, Devin Pope, John A. List, Paolo Piacquadio, Raphael Raux, Pëllumb Reshidi, Frederik Schwerter, Frank Schilbach, Benjamin Schoefer, Sofia Shchukina, Richard Thaler, Catherine Tucker, George Wu, and Luigi Zingales for valuable feedback and suggestions. Bernard Shen, Rishane Dassanayake, Ziyue Feng, Marco Di Giacomo, Shubhaankar Gupta, and Andrew James provided outstanding research assistance.
Partnership & support: At PSG Global Solutions, we are deeply grateful to Vivek Padmanabhan, David Koch, David Bethea, and Anand Krishnan for their innovative leadership and unwavering scientific partnership. We also thank Glaiza Acevedo, Pritam Gawai, Rob McIntosh, Anand Narole, Jon-Ray Riveras, Ram Sabaru, James Shen, and Shraddha Bhujbal for their indispensable logistical support. At Google Cloud, we are grateful for the infrastructure and engineering support provided by Karim Atek, Brad Burns, and John Cecala. At Chicago Booth, we thank John Buenger and Michael Wu, Rich Cortez and Yvonne McConico, and Cheryl Danton for their guidance on legal, compliance, ethical, and data governance matters. Finally, we are grateful to Emily Bembeneck and Amy Boonstra for their administrative support.
Funding: Brian Jabarian gratefully acknowledges funding support from the UChicago Booth Center for Applied Artificial Intelligence, the Becker-Friedman Institute Program in Behavioral Economics Research, Google Cloud Research Program, Effective Venture US and the Swiss National Science Foundation Grant #10001768.
Research transparency: The study was approved by the University of Chicago IRB (#IRB24-1894; #IRB25-1002) and pre-registered on the AEA RCT Registry (#15385). See also Appendix K. This study was conducted under the Data-Use Agreement #160600008471, established by Brian Jabarian through the Booth School of Business at the University of Chicago, with PSG Global Solutions, a subsidiary of Teleperformance.
Disclosure: The partner firm supplied the data but had no role in analyses, manuscript preparation, or the decision to publish. Both authors and research assistants declare no financial or personal conflicts of interest during the conduct of this study. The pre-registered data collection was completed on June 7, 2025. Nearly two months later, on August 1, 2025, Brian Jabarian accepted an unpaid role as Chief Economist at PSG Global Solutions (a Teleperformance subsidiary), renewing the existing research partnership over the next five years, providing extended access for future data and field experiments within the same pre-agreed publication rights as in the original DUA #160600008471.

1 Introduction

Important firm outcomes, such as who is hired, promoted, or selected as a client, depend on information collected through screening interactions. Doing so at scale requires delegating screening to multiple workers or repetitions over time. Since workers differ in their approaches and abilities, execution necessarily varies across them and over time, which can introduce noise into decision-relevant signals. We show that AI can improve firm outcomes by reducing variance in information collection. In a large-scale natural field experiment in hiring, interviews conducted by an AI voice agent rather than human recruiters led to higher job offers, job starts, and worker retention, with no decline in the productivity of hires. AI achieves these gains by implementing *controlled variance*: interviews are conducted in a more structured and consistent way, thereby reducing interviewer-driven dispersion while still delivering interviews tailored to individual applicants.

Job interviews exemplify how variance in information collection can undermine hiring outcomes. Interviews are designed to collect information about applicants' skills, motivation, and fit, yet substantial evidence shows that human-led interviews are prone to both between-recruiter and within-recruiter variation, especially when discretion is high.[1] Such variation can distort hiring outcomes by introducing noise unrelated to applicant quality. In contrast, AI voice agents have the technical capability to consistently follow guidelines and conduct highly structured interviews at scale. At the same time, this capability might not reduce undesirable variation in real-world settings, or applicants might respond negatively to AI-led interviews. Therefore, whether AI-led interviews improve hiring outcomes remains an open empirical question.

We provide evidence from a natural field experiment conducted in partnership with a recruitment firm, PSG Global Solutions,[2] involving 70,884 applications for entry-level customer service jobs. Applicants who pass an initial pre-screening are randomly assigned to be interviewed by a human recruiter (*Human interviewer* condition) or by an AI voice agent (*AI interviewer*). In a third condition, they have the choice between the two (*Choice of interviewer*). Importantly, all hiring decisions are made by human recruiters. This feature separates the information collection stage - automated by AI - from the evaluation stage, which remains under human control. We collected comprehensive data at the interview, applicant, recruiter, and firm levels to examine how AI-led interviews change the

[1] For instance, recruiters differ in their degree of bias towards specific applicants (Kausel et al., 2016; Mocanu, 2025; Shukla, 2025) and their judgment is influenced by previous applicants (Radbruch and Schiprowski, 2025). More generally, meta-analyses have shown that human-led interviews exhibit low validity in detecting talent, especially when unstructured (e.g., McDaniel et al., 1994; Levashina et al., 2014).

[2] PSG Global Solutions is a recruitment process outsourcing firm and a subsidiary of Teleperformance, a global provider of business process outsourcing services.

information collected in interviews, how recruiters respond to that information, and how these changes translate into hiring outcomes.

During interviews, recruiters are instructed to follow guidelines that outline the topics to be covered, their recommended sequence, and sample questions. However, these guidelines grant recruiters substantial flexibility. For instance, they may tailor the topic coverage, order, and wording of questions to individual applicants. The AI voice agent uses the same guidelines. After the interview, applicants take a standardized test assessing language and analytical skills. Recruiters then assess applicants' performance in interviews and tests and make a threshold-based hiring decision, determining whether the applicant meets the firm's standard.[3] The firm's performance metrics for hiring quality are the likelihood that a hire starts the job and remains employed for at least a month. In our high-turnover market, this retention rate provides a relevant proxy for worker-firm match quality, as both face low frictions: firms can easily identify and replace low-performing workers, while workers can voluntarily leave without negative consequences.

Overall, our experiment shows that AI voice agents not only match human recruiters in the complex task of conducting job interviews but also deliver evidence of improved outcomes in several dimensions without damaging core operations. We find that while applicants in the *Human Interviewer* condition receive a job offer in 8.70% of cases, this fraction significantly increases to 9.73% in the *AI Interviewer* condition – a 12% higher likelihood of receiving a job offer. Importantly, among all applicants randomized into either treatment, applicants in the *AI Interviewer* condition also have an 18% higher likelihood of starting their job ($p < 0.001$) and a 18% higher likelihood of having an employment spell lasting at least one month ($p < 0.001$). These positive effects also persist when we look at whether applicants are still employed after two (17% higher likelihood), three (16%), or four months (17%). In addition, we find positive effects for applicants from AI-led interviews when we condition our sample on applicants who have accepted their job offer. Among those, applicants in the *AI Interviewer* condition have a 7% higher likelihood of starting their job ($p = 0.003$) relative to applicants in the *Human Interviewer* condition. They also have a 6% higher likelihood of still being employed after one month ($p = 0.025$).

Our evidence further indicates that improved retention does not come at the cost of worker productivity. First, we find no significant differences in the reasons for separation between the hired workers. That is, the fraction of workers who leave involuntarily (41%) versus voluntarily (59%) does not differ between workers hired in the *AI Interviewer* and the *Human Interviewer* condition. Second, for a subset of the workers hired, we observe

[3]Recruiters evaluate both AI-led and human-led interviews. When human-led, they assess the interviews conducted by themselves.

performance measures. We find neither statistically significant nor economically meaningful differences in the average time a worker spends handling customers, the satisfaction scores of the customers they handled, or the quality assurance scores assigned by their employer. As these three measures jointly determine productivity in our work context, there appear to be no differences in productivity between hiring modes.

To understand why offer rates are higher under AI, we analyze interview transcripts using natural language processing. We first confirm that interview performance scores and comments strongly predict offer decisions, even after controlling for standardized test scores.[4] Analyzing interview transcripts shows that AI agents standardize the interview process by implementing a more structured interview. Compared to human recruiters, they adhere more closely to the expected topic order, cover a more consistent number of topics across interviews, and use more standardized wording with greater lexical richness in prompts and follow-ups. This standardization is associated with applicants in AI-led interviews exhibiting higher levels of linguistic features that, in human-led interviews, predict higher offer rates, such as sustained conversational exchange, and lower levels of features associated with lower offers, including backchannel signals and applicant-posed questions. Together, these patterns suggest that a key advantage of AI-led information collection lies in its ability to achieve what we call *controlled variance*: the AI voice agent adapts its flow to each applicant within a standardized framework, reducing interviewer-driven dispersion while preserving within-interview responsiveness. This structured consistency yields richer and more comparable hiring-relevant information. However, our analysis also points to room for improvement: 5% of applicants ended their interview because they were unwilling to speak to an AI, and in 7% of cases, the AI voice agent faced technical difficulties.

Next, we analyze applicants' behavioral responses to the introduction of AI voice agents and find no evidence of backlash. First, applicants accept job offers with similar likelihoods in the *AI Interviewer* condition relative to the *Human Interviewer* condition. Second, the industry's key applicant satisfaction metric – Net Promoter Score, the likelihood of recommending the firm to a friend – is almost identical across treatments. Third, in a detailed candidate experience survey, applicants rate interview quality variables such as perceived stress, comfort, follow-up fluency, and feedback quality similarly between treatments. Differences emerge only in the perceived naturalness of the interaction, with AI-led interviews being perceived as significantly less natural, and in reported gender-based discrimination. Here, switching to AI nearly halves the rate of reported discrimination

[4]Furthermore, 96% of recruiters state in our recruiter survey that they consider interview performance at least as important as test scores in their offer decisions, and 33% consider them more important.

(3.30% vs 5.98%, $p = 0.02$).

Importantly, when given the choice, most applicants prefer an AI interviewer over a human recruiter: in the *Choice of interviewer* condition, 78% choose the AI voice agent. Survey evidence shows generally positive attitudes toward AI in this sample – most respondents expect AI to benefit both themselves and society – and these attitudes predict interviewer choice. However, along the quality dimension, we find evidence of negative sorting into AI: applicants who choose the AI voice agent have significantly lower language and analytical scores than those who choose a human recruiter.

We then analyze how recruiters evaluate applicants. A total of 131 recruiters evaluated applicants, and a core group of 43 handled most of them. Recruiters assign significantly higher interview scores to AI-interviewed applicants than to those they interview themselves. This effect is driven by a shift from low to medium scores, while the share of high scores remains unchanged. Similarly, sentiment analysis of recruiters' qualitative comments accompanying their scores reveals that comments are significantly more positive for AI-interviewed applicants. Aggregating to the recruiter level shows that these differences are widespread among our sample of recruiters: it is not a few outliers but a majority who provide higher scores and extend more offers to applicants interviewed by AI. Interestingly, we find that recruiters weigh the quality signals they receive from the interview and standardized test scores differently in their offer decisions. When evaluating AI-interviewed applicants, human recruiters place less weight on interview scores and more on language scores than when evaluating applicants interviewed by themselves.

Related literature. Our paper contributes to three strands of research. First, we contribute to the literature that studies the impact of employing generative AI tools on economic outcomes. A growing body of evidence shows that these tools can enhance productivity in diverse tasks such as writing, customer support, or software development (Choi and Schwarcz, 2023; Doshi and Hauser, 2023; Noy and Zhang, 2023; Peng et al., 2023; Chen and Chan, 2024; Wiles and Horton, 2024; Brynjolfsson et al., 2025; Dell'Acqua et al., 2025; Kumar et al., 2025; Otis et al., 2025). Importantly, in these studies, humans remain in control of the core labor tasks, while AI tools augment their work through assistance and information provision. In contrast, we examine a setting in which an AI agent replaces humans in an expert task: conducting job interviews. In doing so, the AI agent must autonomously collect and interpret information generated during a natural-language conversation. Thus, we provide causal evidence from a natural setting on the impact of automating a production stage using AI agents, both in terms of economic consequences and human behavioral responses. Our results highlight that a key advantage of

AI-based systems may lie in their ability to implement *controlled variance*, which informs the debate on which tasks or environments AI automation may have the largest effects.[5]

Second, with our hiring setting, we contribute to the literature on the use of AI and algorithms in labor markets. A large body of literature has examined the impact of technologies that help recruiters select applicants prior to job interviews or help evaluate them afterward (see, e.g., Hoffman and Stanton, 2024, for a review). Evidence has shown that algorithmically recommending workers increases match quality and fill rates in online markets (Horton, 2017), adding AI algorithms to the screening increases the success of hiring (Awuah et al., 2025), and optimizing algorithms has significant effects on the quality of applicants selected for interviews (Li et al., 2025) and hires (Dargnies et al., 2025). Moreover, Aka et al. (2025) conduct a field experiment in which applicants are screened either by a resume-scoring algorithm or by completing an AI-led screening, before a human recruiter selects applicants and conducts a human-led interview. They find that applicants screened through AI-led interviews are more likely to pass the subsequent human-led interviews and find jobs. On the evaluation side, allowing AI to override human hiring decisions influences job acceptance rates and worker productivity (Cowgill, 2020), and making AI evaluation scores available to recruiters changes their assessment (Avery et al., 2024). Importantly, these papers examine the stages preceding or following the interview and maintain human control over the interview process. In our setting, AI agents are employed in the interview itself, which is one of the most human-intensive parts of the hiring process. Because we randomize AI employment and link interview data to employment data, we can quantify effects on firm efficiency, recruiter behavior, and applicant responses. Accordingly, we relate to a broader literature in labor economics that studies how recruiters use and decide based on quality signals, see e.g., Hoffman et al. (2018).

Third, we contribute to a growing body of behavioral research investigating how humans, and in particular workers and managers in labor settings, perceive, trust, and interact with AI systems that collect or provide information. On the information collection side, recent work on AI in persuasion finds that AI agents are less effective than humans in debt collection calls (Choi et al., 2025), while being capable of changing beliefs in conspiracy theories (Costello et al., 2024). Furthermore, Chopra and Haaland (2023) and Geiecke and Jaravel (2024) show how AI can be used to effectively conduct qualitative interviews at scale for research data collections. On the information provision side, previ-

[5]Recruiters typically perform two expert tasks: conducting job interviews and evaluating applicants. Our evidence thus offers empirical input to the debate on whether AI shifts human labor to higher expertise (Brynjolfsson and McAfee, 2014; Acemoglu and Restrepo, 2019; Athey et al., 2020; Gruber et al., 2020; Alam et al., 2024; Autor, 2024; Autor and Thompson, 2025), in this case by redirecting recruiter expertise toward evaluation and potentially raising standards in low-entry job assessments.

ous studies have mainly examined settings in which AI provides signals in the form of forecasts or recommendations. For instance, Agarwal et al. (2024, 2025) show that humans under-respond to AI predictions in the contexts of radiology and fact-checking. Angelova et al. (2023) and Stevenson and Doleac (2024) show that judges frequently deviate from or override AI recommendations.[6] We add to this literature field evidence on real-world economic outcomes and rich behavioral data on human-AI interactions.

The remainder of the paper proceeds as follows. In Section 2, we describe the technical background of the AI voice agent, the firm's recruitment process, as well as the experimental design and sample. Section 3 presents the recruitment outcomes and Section 4 transcript-based evidence on underlying mechanisms. Section 5 examines applicant responses to AI, Section 6 analyzes recruiter behavior, and Section 7 estimates the operational implications for the firm with respect to time and cost savings. Section 8 concludes.

2 Background and experimental design

We study the impact of introducing an AI voice agent that collects information from humans through spoken conversations. Specifically, the agent conducts hiring interviews, thus interacting with the applicants in a conversation. The goal is to collect information from this interaction that recruiters can use later to make hiring decisions. In the following, we describe the AI system deployed, the economic environment in which the hiring takes place, as well as the design of the field experiment.

2.1 AI voice agents

Technical architecture. AI voice agents are a specific class of recently developed "generative AI" tools that generate new data after being trained on existing data using machine learning. The purpose of AI voice agents is to communicate with humans through natural language conversations. To enable conversations, the agent generates human-like speech and responds to human speech based on three interacting technological systems.

First, to generate the content of AI-based speech, the agent generates text using a large language model (LLM). LLMs are trained on large amounts of text and are tasked to predict the next word in a sequence based on prior context. This enables them to generate contextually relevant content in natural language. Second, to produce the speech itself, the agent is equipped with a text-to-speech system, which converts the text generated

[6] Moreover, Fumagalli et al. (2022) and Dargnies et al. (2026) study people's perceptions and preferences towards algorithms in hiring.

by the LLM into audible speech using a multi-step process. It first models pronunciation by translating text into phonetic representations. Then, it determines the appropriate prosodic features, such as stress and intonation, and regulates the pace for natural delivery. Finally, a neural voice encoder (vocoder) generates natural-sounding speech. Third, to respond to speech inputs from the human counterpart, the voice agent converts the spoken input into text using an automatic speech recognition system. This process involves acoustic modeling of the input waveform to identify phonetic units, followed by lexical and language modeling to infer the most likely word sequence. In this step, again, a large language model is used to infer content from speech with sufficient accuracy, which is particularly important in cases of noisy or poorly articulated speech. The first system then uses this input to generate a response.

Challenges to implementation. Using AI voice agents for spoken conversation poses several challenges to AI systems, which are particularly pronounced in interviewing. Human language is multifaceted, layered, and complex. Any lapses, misunderstandings, or errors on the AI side will reduce the experience of the applicants. It also makes it more difficult or even impossible for recruiters to evaluate the interview, leading to information loss. Moreover, it is important that the AI remains on topic during the entire conversation. This requires minimizing instances of "hallucinations," where LLM-based tools unpredictably generate coherent but factually incorrect or nonsensical output. Similarly, guardrails need to be in place to prevent the AI from going off-topic. Moreover, the AI needs to be secure against attempts by applicants to game it, e.g., if the applicant is parroting buzzwords or reading from a script.

While most of these challenges are shared with text-based AI tools, additional challenges arise for spoken conversations. Any functioning tool must deal with ambient noise, variations in speech rate, and differences in accents and intonation. In addition, the time to respond – latency – becomes even more important as multiple systems need to work together, potentially increasing latency. At the same time, substantial delays in questions and responses break the flow of a conversation. This, in turn, may decrease comfort and increase stress for the human counterpart.

2.2 Economic environment

Data partner. We partner with a firm that has employed a AI voice agent in their hiring process. The firm is the recruitment process outsourcing (RPO) firm PSG Global Solutions (hereafter referred to as "the firm"), integrated in the $11 billion global business process outsourcing (BPO) firm Teleperformance. The firm specializes in high-volume re-

cruitment for Fortune 500 clients across the healthcare, IT, and industrial sectors, with recruiting centers worldwide.

Job descriptions. Our setting is the Philippines, where the firm recruits customer service representatives for large US-based and European clients. The jobs for which our firm recruits pay between Php 16,000 to Php 25,000 per month ($\approx$ \$280 to \$435).[7] Required skills include English fluency, communication skills, flexibility to work in changing shifts, strong analytical and logical thinking, and problem-solving skills. See Appendix Section C for an example of a detailed job description.

Industry background. The customer service industry in the Philippines is a major sector, estimated to employ more than 1.5 million workers (Hernandez, 2024). The Philippines has become the world's leading provider of call center support, with a large proportion working to assist customers in the US. Factors driving this growth include, on the labor supply side, factors such as (i) a large share of young and comparatively well-educated individuals, particularly in terms of English language proficiency, and (ii) comparatively high wages, as call center jobs offer better pay given their skill requirements compared to alternative employment options within the Philippines. On the demand side, wages are relatively low compared to high-income countries such as the US, and the Philippine accent is close to a typical US-American accent. These factors make it attractive for US-based firms to outsource customer support to call centers in the Philippines.

Generally, the call center industry is characterized by high turnover. Estimates from the US suggest that up to 60% of call center workers leave each year (Buesing et al., 2020), and similar numbers have been suggested for the Philippines (Sallaz, 2019). These high attrition rates lead firms to devote substantial resources to continuous recruitment and training of new workers (Berg et al., 2018).

Together, these factors mean recruiting firms face large volumes of applicants. This has resulted in a highly competitive market in which multiple recruitment firms compete to identify and recruit qualified candidates. Due to competition with other recruiting companies, recruiters have to be quick with qualified candidates, as they often apply to several recruiting firms simultaneously. Interviews are scheduled as soon as possible and hiring decisions are also made quickly. Within the large pool of applications there is a substantial fraction of applicants who are unfit for the job and need to be identified and screened out by the recruiters in the interviews.

[7] As reference, at the time of the experiment, the minimum wage in the Philippines ranged from \$125 to \$260 per month, depending on the region.

2.3 Recruitment process

The firm has established low entry barriers to interview as many applicants as possible. Applicants learn about job openings through job advertisements posted on various platforms such as Indeed, Facebook, the firm's website, its digital campaigns, or through referrals.

Recruitment modes. After learning about a job opening, applicants have two ways to apply, which determine the subsequent recruitment mode. In the *Remote* mode, applicants submit an expression of interest online, which contains contact details and some basic information. In the *Walk-in* mode, applicants come in person to the nearest firm recruitment site. Once the firm receives an application, a profile is created in the recruitment software and the formal recruitment process begins. Appendix Figure A.1 provides an overview of the process, which we will discuss in detail next.

Screening process. The firm's approach is to conduct the primary screening during the interview itself. Hence, little to no screening is performed prior to the interview. In the *Remote* mode, the primary variable for selection is the engagement score of an applicant. The engagement score is generated algorithmically based on the level of detail provided by the applicants in their expression of interest. Applicants whose scores are below a certain threshold are screened out.[8] The rest are invited to an interview. For applicants in the *Walk-in* mode, no engagement score is calculated; instead, all eligible applicants are invited to an interview after expressing their interest in person. Interview invitations are sent by telephone, text messages, and email. See Appendix Section D.2 for the content of the invitations. Applicants are assigned randomly to recruiters using a round-robin scheduling algorithm (Silberschatz et al., 2018).

Job interviews. Interviews are conducted in two modes. In the *Walk-in* mode, interviews take place in person, while in the *Remote* mode, the interview takes place remotely by telephone. A full-length interview takes between 10 and 20 minutes. Recruiters follow structured interview guidelines designed to ensure a standardized interview process. There is a maximum of 14 topics that can be covered in each interview (see Appendix Table B.2 for details), and questions are a mix of verification and open-ended questions. Interviews start with questions about the suitability of applicants for the position, such as their current location, willingness to commute, and flexibility with respect to the work schedule. Recruiters then ask applicants about their career goals and motivations, before shifting to questions about previous work experience and their education level. Toward

[8] See Appendix Section D.1 for more details.

the end of the interview, recruiters provide additional details about the position. They also provide applicants with the opportunity to ask questions about the position and the recruitment process. The guidelines allow recruiters significant flexibility to adapt their approach. For instance, recruiters are asked to tailor their questions and follow-up questions to the applicant's background to assess aspects like gaps in employment or transitions between jobs. We provide two full-length examples of interviews in Appendix Section F.

Standardized language and analytical tests. If an interview is successfully completed, applicants are invited to a standardized test. The test takes about 30 minutes and contains a language and a quantitative skill component, each featuring adaptive questions. The language component assesses applicants' writing and reading capabilities in English through classic language testing tasks. Scores are based on the CEFR framework (six categories from A1-beginner to C2-proficient). The quantitative skill component consists of three individual parts: attention to detail, verbal reasoning, and numerical ability. Performance is aggregated across the three individual parts into a score from 0 to 100. Completing the test is mandatory to advance to the hiring decision stage.

Review and hiring decisions. Recruiters make hiring decisions based on applicants' interview and test score performance. To do so, they review the (i) interview transcript, (ii) audio recording, and (iii) standardized test scores. Interview performance is assessed according to four main categories. First, recruiters consider the applicant's level of experience in customer service roles. Second, they assess how proficient the applicant is in communicating in English. Third, they evaluate the risk of attrition, that is, the likelihood that the applicant will not work in the prospective job for an extended period. Fourth, they determine whether the applicant's salary expectations align with the offered wage range. Recruiters rate interviews on a three-point scale and provide a short justification in an open-ended text format. For details on the scoring, see Appendix Table B.3. Recruiters make threshold-based hiring decisions, determining whether an applicant is suitable to be hired or not. Although monthly hiring targets exist for each application site, the high volume and continuous flow of applicants as well as sustained demand from client firms ensure that individual hiring decisions have essentially no external effects on other applicants. If a recruiter judges an applicant to be suitable for the job, they assess whether their location and qualifications match a job opening supplied by the client account. If a match is found, an email with the job offer is sent to the applicant. If not, the applicant is kept in the system and may be contacted at a later stage.

Onboarding. If an applicant accepts an offer, they are forwarded to the respective client company. Depending on the client and the job profile, the applicants undergo additional validation and medical checks. Once passed, the applicant begins the onboarding process with a job training period. A key challenge at this step is that usually a substantial fraction of applicants accept the offer but do not show up for training, are absent during training, or do not pass the additional validation checks employed by the client.

Employment. After completing training, applicants begin regular employment with the client. The key employment outcome for the firm is job match quality, i.e., the compatibility between the skills and preferences of the worker and the requirements and conditions of the position. Because match quality is generally not observable, we use worker retention, i.e., whether a worker remains employed after a given period, as a proxy (as is standard in the literature, see e.g., Jovanovic, 1979; Autor and Scarborough, 2008; Hoffman et al., 2018). The hiring firm's main objective is to maximize the retention rate among applicants, as the client firms compensate them for hires who remain employed for at least one month. We argue that the retention rate is a suitable measure of match quality in our context for the following reasons. The business-process outsourcing (BPO) market we study features minimal frictions on both worker and firm side. Because jobs are standardized and performance is quantifiable and continuously tracked, it is easy for firms to identify poor matches quickly, and hard for workers to "hide" performance. At the same time, workers can be replaced at short notice. These factors enable firms to easily fire workers for lack of job performance, which indeed is a prevalent feature in our context: according to the company's records, more than 30% of all involuntary separations are due to failed performance review. Workers, on the other hand, can voluntarily separate from the firm quickly and without incurring negative consequences. In fact, about half of all separations are initiated by the worker in our setting. Due to these factors, we argue that short-term retention is a sufficient statistic for match quality.

2.4 Experimental design

Our pre-registered experiment[9] aims to test the causal impact of automating job interviews with AI voice agents. Accordingly, our treatment variation concerns the interview stage, where we vary who conducts the interview. Once an applicant qualifies for an interview, they are randomized into one of three experimental conditions: *Human Interviewer*, *AI Interviewer*, and *Choice of Interviewer*.

[9]See Appendix Section K for details on the mapping between pre-registration and paper.

Human Interviewer. In the *Human Interviewer* condition, applicants are interviewed by a human recruiter. In the *Remote* mode, human recruiters interview applicants remotely via phone; in the *Walk-in* mode, they conduct the interview in person at the nearest application center of the firm.

AI Interviewer. In the *AI Interviewer* condition, applicants are instead interviewed by an AI voice agent. The agent is prompted to follow the same structured interview guidelines that is given to the human interviewers. The AI voice agent conducts the interview in interview modes – *Remote* and *Walk-in* – via phone. While in *Remote* the phone call takes place remotely, in *Walk-in* the phone call takes place at the application center in which the applicant applied. That is, in both modes, the external environment in which the interview takes place is the same across the *Human* and *AI Interviewer* conditions. When the call starts, the AI voice agent immediately discloses its artificial identity to avoid any deception, according to firm compliance, and explicitly states that a human recruiter will review the interview, evaluate it, and make the hiring decision, not the AI itself.

Choice of Interviewer. Lastly, in the *Choice of Interviewer* condition, applicants can choose whether a human or an AI voice agent will interview them. They are offered the choice upon receiving the interview invitation (in both *Remote* and *Walk-in*). If they do not choose within a certain time frame, the AI voice agent calls them. The AI agent briefly explains the process, provides them with insights into how the AI agent works, and then asks them to make the choice.[10]

Assessment across treatments. Importantly, the evaluation of interviews and subsequent hiring decisions is done by humans, irrespective of treatment condition. That is, regardless of whether the AI voice agent or a human recruiter conducted the interview, a human recruiter reviews the audio, transcript, and test scores of an applicant and then makes a hiring decision. Each applicant is assigned a unique recruiter. In particular, when a human recruiter conducts the interview, the same recruiter also later evaluates the applicant. When an AI voice agent conducts the interview, the same round-robin scheduling algorithm that assigns recruiters to applicants who face a human interviewer also assigns recruiters to evaluate applicants who face the AI agent. This process means that recruiters know whether an applicant was interviewed by *AI* or not. However, in the *Remote* mode, recruiters do not observe whether the applicant was randomly assigned or self-selected for an interview with the AI voice agent or a human. In contrast, in the *Walk-in*, since

[10]The firm implemented this feature because while most applicants have experience with human recruiters, they have none with AI voice agents. This design provides applicants with initial exposure to the AI agent, potentially alleviating some apprehension about being interviewed by an AI.

recruiters themselves ask the applicants assigned to the *Choice of Interviewer* condition to make a choice, they know if the candidates self-selected for an interview with the AI voice agent or with a human recruiter. Importantly, all recruiters evaluate applications of all three treatment conditions. They are further instructed to apply the same assessment criteria for interviews and hiring, regardless of treatment or mode of interview.

By varying only who conducts the interview, the experimental design isolates the direct effect of automating the interview stage. This feature allows us to analyze the performance of AI in collecting information from humans in natural conversations, as well as the human response to AI before, during, and after the interview.

2.5 Experimental sample

Our full sample consists of 70,884 applications that the firm received from March 7 to June 7, 2025. In total, 17,621 applications are in the *Walk-in* mode, 53,263 in the *Remote* mode. Applications were received for 48 different job postings and 41 different client accounts, which operate in the technology, insurance, telecommunications, retail, finance, healthcare, and transportation (See Appendix Table B.1 for the full breakdown). All applications for the job postings were part of the experiment. Applications were processed by 26 different application sites in 19 cities. The sites are distributed across several regions in the Philippines.[11] Most of the applications are being processed by sites located in Metro Manila and the Central Visayas region.

Randomized sample. Of the 70,884 applications, 67,056 were found to be eligible and were therefore randomized to one of the three treatment conditions described above. As pre-registered, we will use this sample for our analysis. In total, 40,103 applications (59.81%) were randomized into the *AI Interviewer* condition (10,421 in *Walk-in*, 29,682 in *Remote*), 13,557 applications (20.22%) were randomized into the *Human Interviewer* condition (3,478 in *Walk-in*, 10,079 in *Remote*), and 13,396 (19.98%) into the *Choice of Interviewer* condition (3,469 in *Walk-in*, 9,927 in *Remote*). In Appendix Table B.4, we provide evidence that randomization was successful, as pre-treatment variables are balanced between the three treatments. In total, 6,319 applications received job offers, and we can match 4,294 applications with employee data.

Applicant characteristics. Of all applications, 64,556 were submitted by unique individuals, which means that 5341 applicants submitted more than one application during the experiment. The majority of 58% of applicants learn about the job through digital job

[11] See Appendix Figure A.2 for information on the regions and the distribution of applications per region.

postings, 19% of the applications happen through referrals, and the rest through other sources such as word of mouth. In total, 60% of applicants were female. Most applicants were between 20 and 30 years old and had some prior experience in customer service jobs.

Applicant survey. To collect more detailed data on applicants' beliefs and experiences, the firm invited applicants to participate in a customer experience survey. Invitations were sent by email. The survey contained five blocks of questions; see Appendix G for the instructions. Applicants were randomly assigned a short or long version of the survey. The long version took about 10 minutes and contained a total of 23 questions within the five blocks. Applicants were compensated with $2 for answering all questions in the survey, a payment that implies an hourly wage six times higher than the minimum wage. The short version took approximately 2 minutes to complete, was offered without compensation, and contained only a subset of questions (11 in total). The survey was sent to 19,200 applicants, of whom 2,764 completed it. This implies a completion rate of 14%.

Recruiters. Our sample of applications was assessed by a total of 131 recruiters. On average, each recruiter had an average of 512 applicants assigned to them throughout the experiment (Median = 121, SD = 1,153). However, this average masks substantial heterogeneity in the number of applications assigned to recruiters. A core team of 43 recruiters is assigned 90% of all applications.

Recruiter survey. To collect additional data on the beliefs and opinions of the recruiters, the firm conducted a firm-wide survey after the experiment but before the results were disclosed to them.. The survey lasted about five minutes and contained two blocks. The first asked recruiters to predict the impact of AI, the second asked about their interview and evaluation experiences; only recruiters who had previously evaluated AI-led interviews received this block. See Appendix H for the instructions. The survey was completed by 173 recruiters. Of these, 133 evaluated AI-led interviews and 98 evaluated applicants who participated in the experiment. These 98 recruiters evaluated 82% of the applicants in our experiment.

3 Main results on employment outcomes

In this section, we compare key recruitment process outcomes between the *AI Interviewer* and *Human Interviewer* conditions. We consider the likelihood that an applicant receives a job offer, successfully starts the job, and is employed one to four months after starting.

Furthermore, we assess the productivity of hired applicants on the job and the separation reasons among those who leave.

Recruiters' forecasts. To benchmark our findings, in our recruiter survey, we asked the firm's recruiters to forecast the impact of introducing AI-led interviews. Overall, recruiters expected that applicants interviewed by the AI voice agent to perform worse. In total, 36% of the recruiters expected applicants to receive lower, 49% equal, and 15% higher offer rates. Similarly, for retention rates, 48% of the recruiters expect workers hired through AI to have lower, 39% equal, and 13% higher retention. We find similar fractions when asking recruiters to predict workers' productivity. Moreover, 61% expected AI-led interviews to be of lower quality.[12]

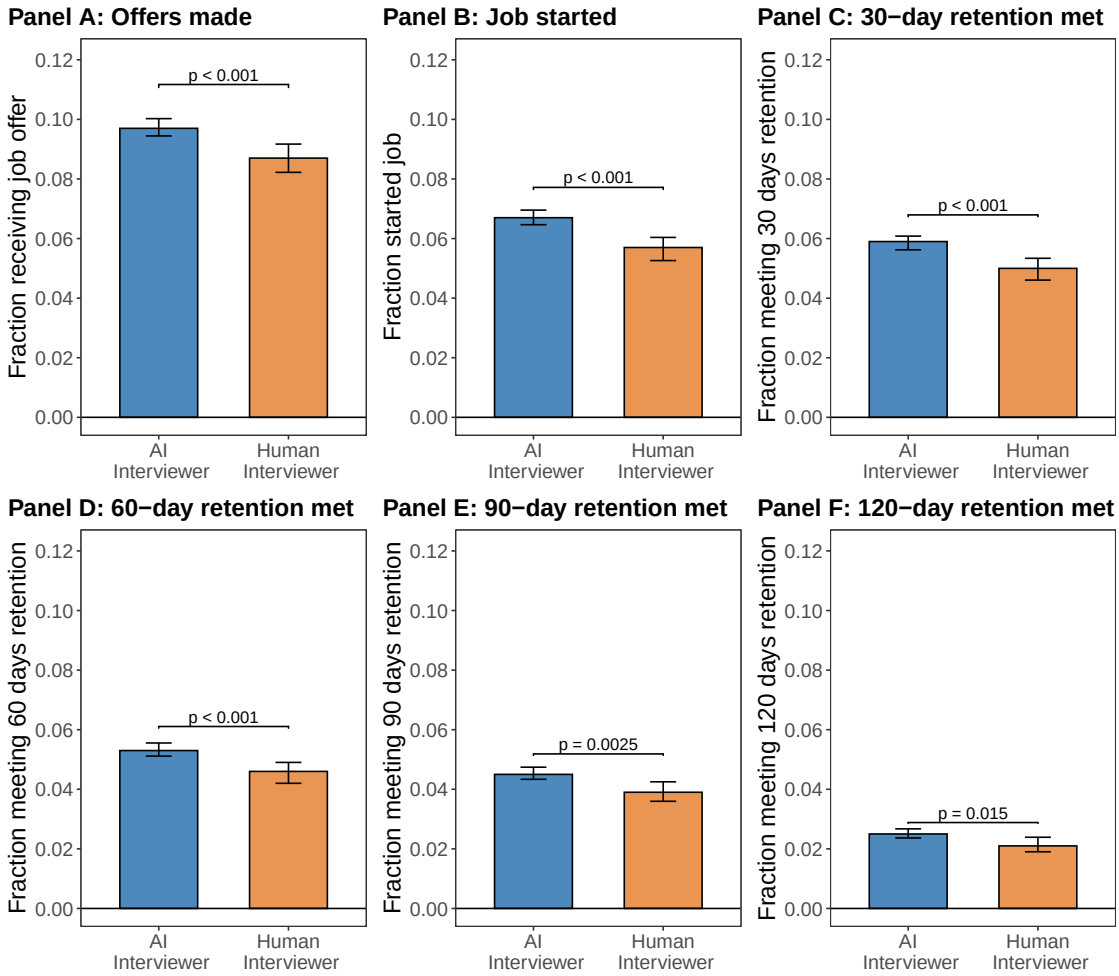
3.1 Results on offer rates and match quality

For the analysis, we consider two samples: first, the unconditional sample, where we compare the outcomes between all applicants who were randomized into either the *Human Interviewer* or *AI Interviewer* condition. Second, the conditional sample, where we condition on applicants who have accepted their job offer.

Sample: all applicants. Figure 1 displays the treatment effects for the unconditional sample. The likelihood of receiving a job offer in the *Human Interviewer* is 8.70% (1,179 out of 13,557 applicants). In contrast, in the *AI Interviewer* condition, the likelihood is 9.73% (3,904 out of 40,103 applicants). Consequently, applications that are interviewed by the AI voice agent have a 1.03 percentage point or 12% higher likelihood of receiving a job offer. This difference is significant ($p < 0.001$, two-sample proportion test). Moving to the job outcomes, we find that 5.65% (766) and 6.71% (2,691) of applicants have started their job in the *Human Interviewer* and *AI Interviewer* condition, respectively. Again, this unconditional difference is significant ($p < 0.001$, two-sample proportion test). Then, we compare the likelihood that applicants are still working at their job one month after starting. In total, 4.97% of applicants (674) in *Human Interviewer*, and 5.85% of applicants (2,347) in *AI Interviewer* are still working. Hence, applicants who were interviewed by the AI voice agent have an 18% (0.88 percentage points) higher unconditional retention rate ($p < 0.001$, two-sample proportion test). Moving to longer-term employment (Panels D-F), we find that the positive effect persists, as we find that the likelihood that applicants are still employed is 17% higher after two months, 16% higher after three months, and 17% higher

[12]Forecasts may be driven by genuine predictions about the AI's technical capabilities but also influenced by factors such as incumbent bias or motivated reasoning, i.e., recruiters downweight the AI's capabilities because it threatens their expertise.

Figure 1: Treatment effect on key recruiting outcomes in the unconditional sample



Notes: The figure displays the recruiting outcomes of applicants. Each panel displays the fraction of applicants who realize the specific outcome. Fractions are displayed separately for the *Human Interviewer* condition, in which applicants are interviewed by a human, and for the *AI Interviewer* condition, in which applicants are interviewed by an AI voice agent. Bars indicate 95% confidence intervals; p-values calculated from a two-sample proportion test.

after four months ($p < 0.001$, $p = 0.003$, and $p = 0.015$, respectively, two-sample proportion tests).

Sample: only applicants who accepted an offer. Next, we examine the job outcomes of applicants in *Human Interviewer* and *AI Interviewer*, conditional on the applicants who have accepted a job offer. That is, we restrict our sample to those applicants and repeat the previous analysis (see Appendix Figure A.3 for a graphical presentation). In total, 8.14% of applicants (1,104) in *Human Interviewer* accept a job offer, while this fraction is 8.99% (3,604) in the *AI Interviewer* treatment. Among them, 68.84% in *Human Interviewer* and 73.36% in the *AI Interviewer* condition successfully start their job, and 60.60% and 64.32% are still employed after one month. Accordingly, applicants in *AI Interviewer* who have

accepted their job offer have significantly higher fractions of job starters and employees staying for at least one month ($p = 0.003$ and $p = 0.025$, respectively, two-sample proportion test). We find similarly positive effects when looking at longer employment time horizons: After two months, 55.62% in *Human Interviewer* and 59.13% in *AI Interviewer* are still employed ($p = 0.038$). After three months, these numbers are 47.92% and 50.31% ($p = 0.16$), and after four months, 26.09% and 27.86% ($p = 0.25$). Accordingly, retention rates are consistently higher in the *AI Interviewer* condition and improvement in terms of percentage ratio is stable over time (one month: 6%, two months: 6%, three months: 5%, four months: 7%), although later differences are no longer statistically significant due to the shrinking sample size.

Robustness. Our results are robust to the inclusion of additional pre-treatment controls and fixed effects. Specifically, we regress our three key recruitment outcomes on treatment status. We control for an applicant's gender, source of application (i.e., referral, online job posting, etc.), their pre-treatment engagement score, and whether the application is from an applicant who submitted more than one application to any of the firm's job postings during the period from six months before the experiment began until its conclusion. Furthermore, we include fixed effects for the week in which the application was received, the recruiter who was assigned to the application, the city of the application site that received the application, and the specific job posting that the application targeted. Across all outcomes, whether an application led to an offer, a successful job start, and an employment spell of at least one, two, three or four months, we find very similar treatment effects with and without added controls and fixed effects. For details, see Appendix Table B.5 (sample of all applicants) and B.6 (sample of applicants who accepted an offer).

3.2 Results on separation reasons and productivity

Separation reasons. In addition to the retention rate as our main job outcome variable, we also observe, for employees who left their job after starting, whether they left voluntarily (i.e., initiated by the employee) or involuntarily (i.e., initiated by the employer). The primary reasons for voluntary leaves are that employees took another job, returned to school or university, or had family obligations or childcare responsibilities. The primary reasons for involuntary leaves are excessive absenteeism or failed performance checks. Comparing rates between treatments among those employees who left their jobs, we find that 58.25% of employees hired from the *Human Interviewer* condition separate voluntarily, while 58.81% of employees hired from the *AI Interviewer* condition do so (thus, 41.75% and 41.19%, respectively, leave involuntarily). Hence, there is no difference in leaving

reasons between conditions ($p = 0.95$, two-sample proportion test).

Job performance. For a subset of employees, we also observe their job performance. Specifically, we observe three key performance variables that are commonly considered to jointly determine a customer service workers' productivity (see e.g., Brynjolfsson et al., 2025). First, the *average handle time*, which is the average time (in minutes) a worker spends handling a customer request from start to finish. Second, the *quality assurance score* (between 0 and 100), which is the employer's assessment of a worker's quality in handling customer requests based on a review of selected interactions. Third, the *customer satisfaction score* (between 0 and 100)[13], which measures customer feedback elicited after a request. Productive workers manage to achieve low average handle times while maintaining high customer satisfaction and quality assurance scores.

For these three productivity measures, we compare the performance of workers hired in the *AI Interviewer* condition to those hired in the *Human Interviewer* condition. We estimate an OLS regression controlling for client, worker, and month fixed effects. Table 1 shows the results. Across all three performance variables, we find no significant differences between workers hired in the *AI Interviewer* condition and those hired in the *Human Interviewer* condition.[14] While workers in the *AI Interviewer* condition take slightly longer to handle requests, customers are also slightly more likely to be satisfied. The differences thus do not show a consistent directional tendency across the variables and are economically very small. This result suggests that the positive effects on job offers, starters, and retention are not accompanied by decreased employee productivity.

4 Signals from interview transcripts

Why are applicants in the *AI Interviewer* condition more likely to receive offers? We investigate a key potential mechanism: AI voice agents conduct interviews more consistently than human recruiters, resulting in a more standardized interview process. This standardization is associated with applicants exhibiting more informative linguistic features. That is, signals that more precisely reveal applicants' types and are therefore more relevant to recruiters' hiring decisions. To test this mechanism, we analyze interview content using transcripts.

[13] Depending on the client, different scales are used. We perform a linear transformation to create a common 0-100 scale.

[14] We do not observe all three variables for every worker. Our results are similar if we focus only on workers for whom we observe all three variables.

Table 1: Job performance of applicants hired

	Dependent variable:								
	Average handle time (minutes)			Quality assurance (0-100)			Customer satisfaction (0-100)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
AI Interviewer	1.519	1.429	1.379	-0.293	-0.295	-0.263	1.618	1.697*	1.461
	(1.027)	(1.009)	(1.007)	(0.535)	(0.537)	(0.519)	(0.990)	(0.994)	(0.969)
DV mean	23.992	23.992	23.992	86.265	86.265	86.265	66.946	66.946	66.946
Observations	7,560	7,560	7,560	4,512	4,512	4,512	3,877	3,877	3,877
R^2	0.295	0.301	0.304	0.361	0.362	0.375	0.450	0.450	0.467
Client FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Worker FE	–	Yes	Yes	–	Yes	Yes	–	Yes	Yes
Month & tenure FE	–	–	Yes	–	–	Yes	–	–	Yes

Notes: The table shows OLS estimates on the difference in job performance of workers depending on the treatment condition. *AI interviewer* is an indicator equal to one if the interview leading to the hiring of the worker was AI-led, and zero if it was human-led. The dependent variable in columns (1)-(3) denotes the time (in minutes) a worker spent handling a customer service request, averaged across a month among all requests handled by the worker. The dependent variable in (4)-(6) is a monthly quality assurance score (from 0 to 100) by the employer about the worker, where higher scores indicate better performance by the worker. The dependent variable in (7)-(9) is satisfaction reported by customers for requests handled by a worker (from 0 to 100), averaged across a month. Observations for this regression are at the worker-month level, and all standard errors are clustered at the worker level. Significance levels: * $p < 0.1$,** $p < 0.05$, *** $p < 0.01$.

Interview transcript data. The firm shared with us raw verbatim transcripts. In total, we have transcripts available for 34,109 applications in the *Human Interviewer* and *AI Interviewer* condition, which is a subset of all interviews conducted. Using LLMs and standard prompt engineering techniques, we first anonymized and labeled the transcripts. We then construct three distinct sets of variables. First, a general classification of interview types. Second, a set of variables capturing recruiter behavior in interviews. Third, variables that capture applicants' language and behavior. We describe the transcript variables briefly in the main text, and provide a detailed description of all transcript variables in Appendix Section E.

4.1 Interview content and its relevance for job offers

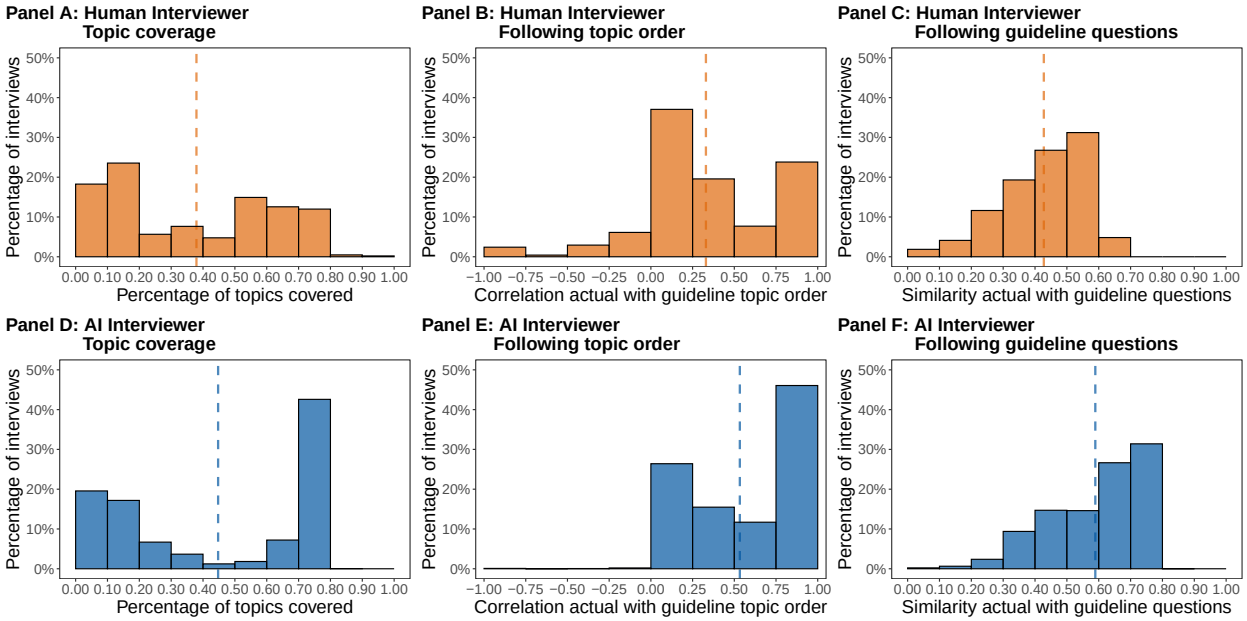
Relevance of interview content. We start our analysis by assessing the relevance of job interviews for offer decisions. Anecdotally, the firm views job interviews as a crucial step in the hiring process, with recruiters placing significant weight on applicants' interview performance when deciding whether to extend a job offer. This is supported by recruiters' survey responses: when asked about the relative importance of interview performance compared with test scores in determining offers, 33% of recruiters say interview performance is more important, 63% say they are equally important, and only 4% consider test scores more important.

To empirically test the relevance of interviews in hiring decisions, we examine the baseline predictive power of interview variables for offer decisions. Accordingly, we focus on applications in the *Human Interviewer* condition. Our outcome variable is whether an application led to a job offer. We use two variables that capture or assess interview content to predict offer decisions. First, the numeric score (1,2,3) with which recruiters rate applicants' interview performance, and second, the open-ended text assessment with which recruiters describe the applicants' performance. For open-ended text, we use natural language processing to classify text assessments by sentiment (negative = -1, neutral = 0, positive = 1).

We find that both variables strongly predict offer decisions. Each increase in interview score is associated with a 25 percentage-point increase in offer likelihood. Conditional on test score performance and our other controls and fixed effects, the increase remains highly significant at 15 percentage points. The sentiment of the recruiter text assessment is similarly highly predictive of offer decisions in the *Human Interviewer* condition. For details, see Appendix Table B.7.

Interview type classification. Given that interview content matters for offers, which types of interviews occur? We classify each transcript into one of ten mutually exclusive interview types using an LLM (see Appendix E.1 for details). We find that the largest share of interviews is classified as *comprehensive interviews*, characterized by an organic opening and closing, high applicant engagement, and at least eight distinct covered topics. In total, 42% of interviews conducted by the AI voice agent and 39% of interviews conducted by a human recruiter are classified in this way.[15] In total, 7% of AI and 25% of human interviews are *Screen-outs*, in which the interview is ended by the recruiter because the applicant does not meet certain requirements. Instances of *Disengaged interactions*, where applicants are unresponsive, distracted, or disinterested, account for 12% and 9% of interviews, respectively, whereas instances of applicants being unavailable account for 14% and 9% of interviews. Lastly, we identify two AI-specific categories. In 7% of AI-led interviews, the interview is aborted due to a technical failure of the AI voice agent, and in 5% of interviews, the applicant explicitly expresses the unwillingness to continue speaking with the AI voice agent. As a summary, Appendix Figure A.6 plots the distribution of the interview types.

Figure 2: Recruiter distribution topic coverage



Notes: The figure displays the distribution of several recruiter variables inferred from transcript data. "Topic coverage" denotes the percentage share of how many of the 14 guideline topics are covered during an interview. "Following topic order" is defined as the Kendall's τ correlation between the actual and guideline topic sequences. "Following guideline questions" denotes the similarity between the questions asked and the guideline questions. Distributions are displayed separately for the *Human Interviewer* condition, in which applicants are interviewed by a human, and for the *AI Interviewer* condition, in which applicants are interviewed by an AI voice agent. The dashed lines display mean values.

4.2 Recruiter behavior in interviews

Variables capturing recruiter behavior. As described in Section 2.3, recruiters are tasked with following structured interview guidelines to ensure a standardized interview process. Specifically, they have a list of 14 topics, a recommended topic order, and representative questions and conversation starters for each topic. To measure the extent to which human recruiters and the AI voice agent implemented structured, standardized interviews, we construct the following variables for each interview: (1) the *percentage of topics covered*, i.e., how many out of all topics the recruiter covered, (2) the *correlation of the actual with guideline topic order* (using Kendall's τ), measuring how closely recruiters followed the guideline topic order, (3) the *similarity of the actual with guideline questions*, measuring how closely recruiters phrased questions following the guidelines, and (4) recruiters *vocabulary richness*, capturing the richness of language used by the recruiter. For details on the variable construction, see Appendix Section E.2.

[15]This difference is significant at the 1% level and robust to the inclusion of controls and fixed effects (see Appendix Table B.10).

Treatment differences in recruiter behavior. Figure 2 provides the distributions of the first three variables separately for both treatments. Starting with Panel A and D, which displays the *percentage of topics covered*, we find that an average AI-led interview covers 45% of all possible topics, while the average human-led interview covers 38%, a significant difference ($p < 0.001$, two-sample t-test). Importantly, the distribution of topic coverage is more distinctly bimodal in the *AI Interviewer* condition, with a concentration around either a high coverage or a very low coverage. Moreover, the variance of topics covered is significantly lower relative to the *Human Interviewer* condition ($p < 0.001$, Levene test).

We next turn to the *correlation of the actual with guideline topic order*, displayed in Panels B and E. While the correlation between actual and guideline order is $\tau = 0.53$ in *AI Interviewer*, it is $\tau = 0.33$ in *Human Interviewer* ($p < 0.001$, two-sample t-test). Hence, as expected in an interactive conversation with probing and follow-ups, both AI and human recruiters do not mechanically implement the topic order in every instance and instead frequently deviate from it. However, AI implements it more consistently across interviews, leading to significantly lower variance in the correlation between the actual and guideline order of topics ($p < 0.001$, Levene test).

Third, in Panels C and F, we compare the *similarity of the actual with guideline questions* asked by the recruiters across conditions. The average similarity per interview is 0.59 in *AI Interviewer* and 0.43 in *Human Interviewer*, again a significant difference ($p < 0.001$, two-sample t-test). As with topic order, neither recruiter mechanically reads from a script but tailors questions and follow-ups to individual applicants and their responses, leading to correlations of less than one in each interview. Yet, the AI voice agent appears to adhere more closely to the guidelines.

The results from these measures suggest that the AI voice agent conducts interviews more consistently. But does this consistency come at the expense of simpler, more mechanical language? To address this question, we examine the *vocabulary richness* of the language used by recruiters. The AI voice agent achieves significantly higher richness scores than human recruiters (7.64 compared to 6.66, $p < 0.001$, two-sample t-test). Moreover, the distribution of vocabulary richness for the AI voice agent is more tightly centered around high scores, whereas that of the human recruiters varies substantially across interviews ($p = 0.004$, Levene test). Appendix Figure A.4 plots the distributions.

Robustness. So far, we have compared the transcript content with the firm's interview guidelines and found that the AI voice agent follows them more closely. The benchmark in that analysis is thus the firm standard. As a complementary test, we examine within each condition how similar recruiters' questions are to the *average question* of each topic –

that is, to the typical language used for that topic within the same condition. Specifically, we compute the distance between each recruiter's question embedding and the centroid embedding of all recruiters for that topic, where higher values indicate a greater deviation from the average. Across nearly all topics, we find that the distance is significantly lower in the *AI Interviewer* than in the *Human Interviewer* (see Appendix Table B.8 for details).

Heterogeneity in recruiter behavior. The previous results were at the transcript level, comparing the average interview performance of AI and human recruiters. A complementary analysis is to compare their respective performance across interviews. That is, for each human recruiter, we calculate the average value of our four recruiter behavior variables across all interviews they conducted during the experiment. We find substantial heterogeneity in behavior. Some recruiters achieve substantially higher topic coverage and similarity with the guideline topic order and questions, as well as higher vocabulary richness than others. For instance, some recruiters cover on average 60% of topics in their interviews, while others cover less than 40%. Importantly, when comparing the distribution of recruiter averages with the average of the AI voice agent, we find that the latter achieves a higher similarity with the guideline topic order and questions than 83% and 100% of human recruiters, respectively. For the percentage of topic coverage and vocabulary richness, the AI voice agent achieves higher scores than 61% and 64% of human recruiters.[16] For the full distribution, see Appendix Figure A.5.

Taken together, these results suggest that AI voice agents implement structured interviews through *controlled variance*. They do not mechanically follow a fixed script but purposefully vary questions and wording across interviews, thereby providing a personalized interview experience to applicants. Yet, implementing these variations, AI remains significantly more consistent than human recruiters, both in adhering to firm guidelines and relative to the average recruiter behavior within each topic.

4.3 Applicant behavior in interviews

Given the AI voice agents' more structured interview approach, how does it impact the responses of applicants? To assess the quality of applicant responses, we construct variables that capture key linguistic features of applicants' responses.

Linguistic features of applicants. For each interview, we construct the following eight variables using standard natural language processing methods: (1) vocabulary richness,

[16]For comparison, the 25% human recruiters scoring highest within the respective variable score on average a topic coverage of 61%, a similarity with the guideline topic order and questions of 0.61 and 0.49, respectively, and a vocabulary richness of 8.86.

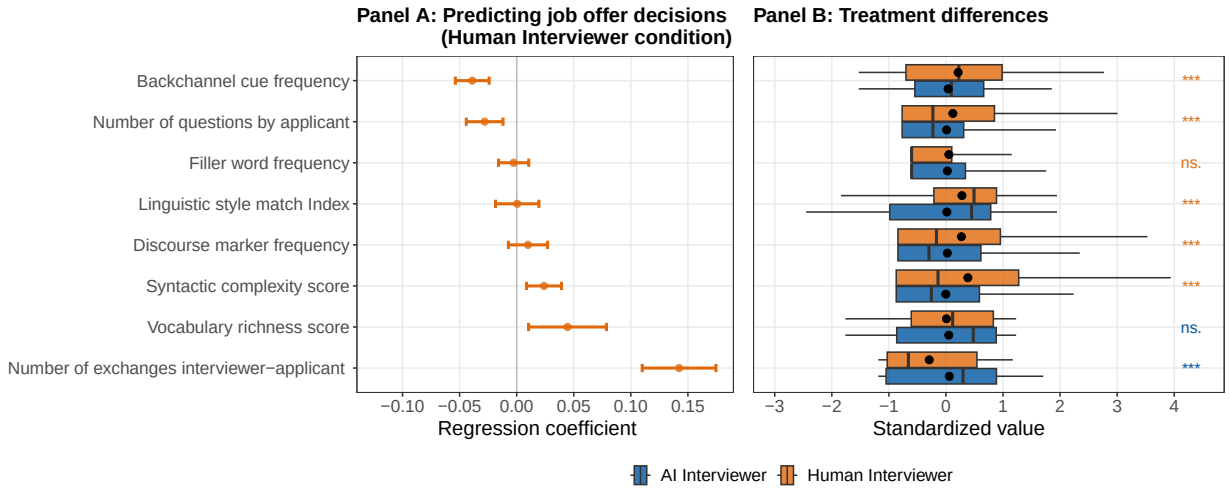
(2) syntactic complexity, (3) frequency of discourse markers (sequential, causal, and clarifying words), (4) frequency of filler words, and (5) frequency of backchannel cues (short cues indicating attention or agreement). We also record (6) the number of exchanges between the interviewer and the applicant and (7) the number of questions posed by the applicant. Finally, we construct (8) an index of linguistic style matching between the interviewer and the applicant. Appendix E.3 provides details on the construction of each variable.

Treatment differences in linguistic features. We develop a two-step approach to investigate how applicants' language differs between AI-led and human-led interviews. In the first step, we identify which linguistic features displayed by applicants are positively, and which are negatively valued by human recruiters. A feature is defined as positively (negatively) valued if it positively (negatively) predicts job-offer decisions. Accordingly, we estimate a joint regression of job-offer decisions on all linguistic variables within the *Human Interviewer* condition. Results are reported in Panel A of Figure 3. We find that the number of exchanges between the interviewer and the applicant, as well as the richness of applicants' vocabulary and the syntactic complexity, are significantly positive predictors. In contrast, the frequency with which applicants use backchannel cues and the number of questions they pose are negative predictors.

In the second step, we compare how the linguistic features differ between treatments. As shown in Panel B of Figure 3, the average scores differ significantly for six of the eight features, and we reject the equality of distributions in every instance ($p < 0.01$, Kolmogorov–Smirnov test, Bonferroni corrected). Importantly, features that positively predict job offers are higher in the *AI Interviewer*, while those that negatively predict job offers – or have no significant predictive power – are higher in the *Human Interviewer* condition. These patterns suggest that AI-led interviews elicit more of the linguistic features that recruiters positively value (e.g., number of exchanges) and reduce those associated with lower hiring odds, such as backchannel cues and applicant-posed questions.

These results indicate that the AI voice agent conducts interviews differently from human recruiters, along dimensions associated with higher offer rates. Recruiters value comprehensive interviews that cover a greater number of topics and contain specific linguistic features. AI-led interviews are more likely to meet these criteria, producing more comprehensive conversations with broader topic coverage and higher scores on linguistic features that, in human-led interviews, predict higher job offers.

Figure 3: Transcript linguistic feature analysis



Notes: Panel A displays the coefficients of an OLS regression of job offer decisions on the displayed variables measuring the linguistic content of applicants' responses (all standardized) in the *Human Interviewer* condition. Error bars indicate the 95% confidence interval, obtained using standard errors clustered at the applicant level. For details on the variables, see Appendix Table E.3. Panel B displays the distribution of the variables split by treatment using box plots. Each box represents the interquartile range (25th-75th percentiles), with the horizontal line indicating the median. Whiskers extend to 1.5 times the interquartile range, and dots represent means. Stars indicate statistically significant differences in means between the two treatments (* $p < 0.1$,** $p < 0.05$, *** $p < 0.01$), based on two-sided t-tests with Bonferroni correction.

5 Applicant responses to voice AI interviews

In this section, we examine applicants' attitudes towards AI-led interviews in greater detail. Specifically, we analyze their decisions to accept job offers, their perceived interview experience, and their choices in the *Choice of interviewer* condition.

5.1 Applicants' response to job offers

After completing an AI-led interview, applicants may update their understanding of the hiring company or the prospective job, which in turn influences their decision about whether to accept a job offer. For instance, the use of AI voice agents as interviewers may affect applicants' perceptions of match quality.

In total, of the 6,319 job offers made, applicants accepted 5,854, implying a 92.64% acceptance rate. In *Human Interviewer*, the acceptance rate is 93.64%. The rate is slightly lower in *AI Interviewer* with 92.14%. However, we cannot reject the equality of acceptance rates between the two treatments at the 5% level ($p = 0.086$, two-sample proportion test). In the *Choice of interviewer* condition, 92.73% of applicants who chose the human interviewer accepted an offer, compared to 92.57% of those who chose the AI interviewer. We

find no significant differences in acceptance rates when comparing the *Human Interviewer* condition with applicants who selected the human interviewer, or the *AI Interviewer* condition with those who selected the AI voice agent (Human: $p = 0.55$; AI: $p = 0.67$; two-sample test of proportions). Similarly, we find no significant difference in acceptance rates between applicants who chose the human interviewer and those who chose the AI voice agent ($p = 0.92$, two-sample test of proportions). Accordingly, among applicants who receive job offers, there appears to be no negative reaction to the AI voice agent, as evidenced by no significant increase in offer refusals.

5.2 Applicants' interview experience

Next, we analyze applicants' attitudes and perceptions of their interview experiences. These were measured using the customer experience survey administered to applicants after the interview.

Net promoter score (NPS). We start by comparing the NPS – how applicants rate their propensity to recommend the recruiting firm to a friend – which is the central feedback metric used by the firm and the industry more generally. We find that the average rating for applicants in the *AI Interviewer* condition is 8.97 on a 1-10 scale, whereas it is 8.84 in the *Human Interviewer* condition. Accordingly, AI voice agents yield a small and insignificant increase in the NPL score ($p = 0.25$, t-test).

Perceived recruiter quality. Turning to how applicants assess recruiter quality, we find no significant differences between the *AI Interviewer* and *Human Interviewer* conditions in how applicants rate the recruiter's knowledge about the firm ($p = 0.58$, t-test) or the role ($p = 0.11$, t-test), nor in the extent to which they felt their time was valued by the recruiter ($p = 0.24$, t-test). The only notable difference is in the perceived relevance of the interview questions, with applicants rating the AI voice agent as asking slightly more relevant questions ($p = 0.044$, t-test). The index of perceived recruiter quality, which averages the individual items, also shows no difference between treatments ($p = 0.92$, t-test). Thus, applicants evaluate recruiter quality similarly across conditions.

Perceived interview quality. We find that applicants rate interviews conducted with the AI voice agent as slightly less stressful and more comfortable, but the differences are very small and far from significant at conventional levels (stressful: $p = 0.65$, comfortable: $p = 0.69$, t-test). In terms of naturalness, applicants rate the interview experience with the AI voice agent as significantly less natural ($p = 0.014$, t-test). This results in a higher index of perceived interview quality for human-led interviews than for AI-led interviews

($p = 0.076$, t-test). When asked about the follow-up flow and the frequency of feedback, the applicants in both treatments rate their experience similarly (follow-up flow: $p = 0.78$, frequency of feedback: $p = 0.25$, t-test). These results suggest that the AI voice agent can deliver interview-quality speech that applicants perceive as similar to that of humans.

Fairness and discrimination perception. When asked to rate the fairness of their interview, applicants rate it similarly across both treatments ($p = 0.68$, t-test). For the question of whether subjects feel discriminated by the recruiter based on their gender, we find a significant difference: while 3.30% of applicants answering the survey in *AI Interviewer* report feeling discriminated, 5.98% do so in the *Human Interviewer* condition, a significant difference ($p = 0.020$, two-sample test of proportion). Accordingly, reported discrimination nearly halves with the AI voice agent. Note, however, that these reports constitute a relatively small sample (62 out of 1818 respondents in *AI Interviewer* and 22 out of 346 respondents in *Human Interviewer*), as is common in discrimination data.

Open-ended feedback. At the end of the survey, the applicants were invited to share additional feedback about their interview experience in an open-ended text response. In total, 9.60% applicants of the *Human Interviewer* and *AI Interviewer* condition provided a response. We use two complementary approaches to analyze the responses. First, we use sentiment analysis to classify whether applicants express negative, positive, or neutral sentiment in their text.[17] We find that of those applicants who provide a response, in *AI Interviewer* 71% of the responses have a positive, 14% a negative, and the remaining 14% a neutral sentiment. In contrast, in the *Human Interviewer* condition, 52%, 30%, and 19% of responses exhibit positive, negative, and neutral sentiment, respectively. Accordingly, the likelihood that applicants express a positive interview experience in *AI Interviewer* is higher ($p = 0.005$, two-sample test of proportion). Second, we use *gemini-2.5-flash* to categorize responses into 13 distinct categories. For details on the category definitions and example responses, see Appendix Table B.16. We find that 45% of responses in *AI Interviewer* and 19% in *Human Interviewer* can be categorized as mentioning a comfortable and positive interview experience. In total, 10% of responses in *AI Interviewer* mention problems with the audio, questions in general, or the AI voice agent specifically. In *Human Interviewer*, a total of 13% mention problems. For the full results, see the Appendix Table A.10. The categorization results are thus in line with the sentiment analyses. However, because there may be selection into who provides an open-ended response, these results should be interpreted with caution.

[17] The analysis was conducted on the whole sample, with the instructions blind to treatment status.

Overall, we conclude that applicants generally rate their experience similarly between human recruiters and AI voice agent. Notable exceptions are that the interaction with the AI voice agent is rated as less natural and has fewer applicants reporting gender discrimination, but more positive open-ended responses.

Robustness. A concern when interpreting applicants' survey responses is that their responses may be biased by demand effects. Although the survey is administered by a separate unit to which recruiters do not have access, and this was communicated to applicants in the survey, applicants may still believe that they can influence the process through their responses. For example, this may lead them to provide overly positive answers. Although such misreporting would bias the level of survey responses, it is less likely to influence our relative treatment comparison, as it would require an interaction between misreporting and treatment. To empirically assess the extent of bias, we randomly assigned applicants to either receive the survey invitation immediately after their interview or after a final decision on their application. This allows us to test whether applicants' responses differ. We find no evidence that applicants strategically change their responses, as we cannot reject the null hypothesis of equal means for our survey variables. See Appendix Table B.17 for details.

5.3 Applicants' choices of interviewers

Choices. In the *Walk-in* mode, out of the 3,469 applicants that were randomized into the *Choice of Interviewer* treatment, 3,420 (98.59%) made a choice between the human recruiter and the AI voice agent. Of those, 2,370 (69.30%) chose the AI voice agent as interviewer. In the *Remote* mode, out of the 9,927 applicants that were randomized into the *Choice of Interviewer* treatment, 9,659 (97.30%) made a choice between the human interviewer and the AI voice agent. The remaining applicants did not respond to the text interview invitation or the follow-up call. Of those making a choice, 7,885 (81.63%) chose the AI voice agent as the interviewer. Accordingly, in both modes, most applicants prefer to interview with an AI voice agent instead of a human recruiter. Aggregating across modes, we have that 78.41% of applicants choose the AI voice agent. Choice patterns are relatively stable throughout the experiment, as Figure A.9 in the Appendix shows.

Predicting choices. What explains the high fraction of applicants who prefer AI over humans? A primary reason may be convenience: Interviews with the AI voice agent can be scheduled at the applicant's preferred time, including immediately after receiving the interview invitation. In fact, as shown in Section 7.1, AI-led interviews are completed much

faster than human-led interviews. A second reason may be applicants' attitudes towards AI. Our survey evidence indicates that the sample generally perceives AI as having a major impact on the workplace and that this impact will generally be positive. Among all survey respondents, 48% believe that AI will have a major impact on themselves personally, 34% believe that the impact is minor, and 18% believe that AI will have no impact. A total of 47% of the survey respondents think that the impact of AI on them in the workplace will be positive, 34% think that the positive and negative impact will be roughly balanced, and only 19% think AI will have a negative impact on them.[18] We obtain similar results when looking at how the respondents assess the impact of AI on workers generally.

Importantly, for those survey takers in the *Choice of Interviewer* condition, among those believing in a positive impact, 77% choose the AI Voice agent, among those believing in a balanced impact, 72% choose the AI Voice agent, and among those believing in a negative impact, 65% choose the AI Voice agent. When regressing the interview choice on the survey item, we find that the item predicts the choices, albeit only significantly so once controls and fixed effects are added. For detailed regression results, see Appendix Table B.9.[19]

Sorting. Next, we analyze sorting effects: whether applicants differ in their quality when they can choose a recruiter compared to when they are assigned one. We use applicants' test scores as our measure of applicant quality, as they provide a signal about quality independent of interview performance. Three potential sorting effects could be present. First, positive AI sorting means that high-quality applicants opt for the AI interviewer, whereas lower-quality applicants opt for the human interviewer. Second, negative AI sorting means the reverse pattern. Third, the choice and quality of the applicants might not be correlated.

We start by analyzing the association between applicants' choices in *Choice of Interviewer* and their test scores. Applicants who selected the AI voice agent scored an average of 3.14 (out of 6) on the language test and 47.54 (out of 100) on the analytical test. In contrast, applicants who chose the human interviewer obtained higher test scores of 3.37 and 49.77 on the language and analytical tests, respectively. These differences are statistically significant ($p < 0.001$ and $p = 0.001$, respectively, t-tests). The differences are robust to

[18] Perhaps unsurprisingly, recruiters are more pessimistic about the impact of AI on them. We asked the same questions in the recruiter survey and found that 68% of recruiters believe AI will have a significant personal impact, whereas only 12% believe the impact will be generally positive.

[19] One might worry that applicants' perception of the directional impact of AI on themselves is affected by their experience with the AI interviewer. Comparing applicants' responses in *AI Interviewer* with those in *Human Interviewer*, we find no difference ($p = 0.58$, t-test).

the inclusion of controls and fixed effects; for details, see Appendix Table B.11. Similarly, we reject the equality of distributions (language: $p < 0.001$, analytical: $p = 0.010$, Kolmogorov–Smirnov test). For the distributions, see Appendix Figure A.7.

As a next step, we analyze whether there exist sorting patterns conditional on receiving the same interviewer. That is, we compare applicants who got assigned the AI voice agent in *AI Interviewer* with those who chose it in *Choice of Interviewer*, and applicants who got assigned the human recruiter in *Human Interviewer* with those who chose the human recruiter in *Choice of Interviewer*. This type of analysis is robust to a potential confounding factor: it may be that experiencing the AI voice agent, rather than the human recruiter, affects test performance. For example, applicants may increase or decrease their effort after interacting with the AI voice agent. More plausibly, the influence could arise from differential attrition, as applicants are more or less likely to drop out of the recruitment process after completing the interview with the AI voice agent and before taking the test.

We regress on test scores a dummy that is equal to one if the applicant has chosen the respective interviewer and zero if the applicant got assigned the interviewer. We find that applicants who choose AI have lower test scores than those assigned to it, whereas applicants who choose the human interviewer have higher test scores. However, three of the four differences are no longer statistically significant after controlling for and including fixed effects. Accordingly, the sorting effects are slightly weaker when comparing performance in the choice condition with that in the assigned conditions. For details, see the Appendix Table B.12. Together, our results suggest the presence of negative AI sorting.

5.4 Gender differences

An important concern with the introduction of AI tools is their potential to exacerbate existing inequalities, particularly between genders (see e.g., Chen, 2023; Otis et al., 2024). Accordingly, we analyze treatment heterogeneity by gender.

To do so, we regress a gender dummy on the likelihood of receiving a job offer (details in Appendix Table B.13). Without controls, we find a sizable gender difference in offer decisions in the *Human Interviewer* condition: female applicants have a 20% lower likelihood of receiving job offers (7.94% versus 9.86%). However, this difference substantially diminishes with the inclusion of controls and fixed effects and disappears entirely when applicants' test scores are additionally added. Although we do not have sufficient pre-treatment individual-level characteristics to conclusively determine the origin of this difference, these results suggest that it is primarily driven by differences in applicant characteristics rather than by direct gender discrimination by recruiters. Importantly, the gen-

der difference in job offers does not significantly change in the *AI Interviewer* condition. Here, female applicants have a 17% lower likelihood of receiving job offers (8.98% versus 10.87%), and the difference similarly vanishes once controls, fixed effects, and test scores are added.

Overall, these results show that the observed reduction in *perceived* discrimination documented in the survey is not coupled with increases in gender differences. Accordingly, in our context, the introduction of AI voice agents does not increase existing gender inequalities. Interestingly, in the *Choice of the interviewer* condition, we observe a small but statistically significant gender difference. While male applicants choose the AI voice agent in 76% of the cases, female applicants choose it in 80% of cases ($p < 0.001$, two-sample test of proportions).

6 Recruiter response to voice AI interviews

In this section, we analyze how human recruiters react to AI-led interviews when making offer decisions.

6.1 Analyzing recruiters' offer decisions

We begin by analyzing how recruiters score interviews and how they use available interview signals and the intended test scores. We focus our analysis on the *Walk-in* mode because, in the *Remote* mode, it was not mandatory for recruiters to log their interview scores in the system.[20]

Interview scores. For a total of 15,303 applications (88% of all applications in *Walk-in*), we observe the interview score (1,2,3) with which a recruiter rated an applicant's interview performance.[21] For a subset of 10,779 applications, we also observe their justification for the score in an open-ended text format.

On average, applicants receive a score of 1.90 in the *Human Interviewer* condition and a score of 2.01 in *AI Interviewer*, a significant difference ($p < 0.001$, two-sample t-test). Accordingly, recruiters rate the interviews conducted by the AI Voice agent as higher than the interviews they conduct themselves. In Appendix Figure A.8, we plot the distribution of scores across the two conditions. We find that higher scores in *AI Interviewer* are driven

[20] Appendix Table B.15 replicates the main result of this section using the full sample.

[21] Availability is balanced across *Human* and *AI Interviewer* condition ($p = 0.64$, two-sample test of proportions).

exclusively by a higher frequency of score 2 rather than score 1, whereas the frequency of score 3 remains similar.

Interview score justification. When providing the interview score, recruiters also submit a short justification for their score in an open-ended text format. Using natural language processing, we classify their justifications according to their sentiments. We find that 31% of the justifications in the *AI Interviewer* condition are categorized as positive sentiment, while only 24% are so in the *Human Interviewer* condition. This difference is statistically significant ($p < 0.001$, two-sample test of proportions). In contrast, the fraction of justifications that have negative sentiments is 28% in the former and 38% in the latter. At the same time, the fraction of neutral sentiment justifications is roughly similar across conditions (*AI Interviewer*: 41%; *Human Interviewer*: 38%). Accordingly, recruiters' comments show a positive response to the interview content produced by the AI voice agent, consistent with such interviews receiving higher interview scores.

6.2 Determinants of offer decisions

To determine whether to extend an offer to an applicant, recruiters have three signals about an applicant's quality available to them: (i) interview performance, (ii) standardized quantitative test score, and (iii) standardized language test score. We are interested in how much weight recruiters put on each when making decisions and, in particular, whether the weights differ across treatments.

Applicant test performance. To start, we compare applicants' performance in both tests across treatments. In the quantitative test, applicants in *Human Interviewer* achieve an average test score of 48.58 (out of 100), while applicants in *AI Interviewer* achieve an average test score of 48.13, a small and non-significant difference ($p = 0.27$, two-sample t-test). The distributions similarly do not significantly differ ($p = 0.53$, Kolmogorov–Smirnov test). In the language test, the applicants score an average of 3.24 (out of 6) and 3.15 in *Human Interviewer* and *AI Interviewer*, respectively. The differences in averages and distributions are significant ($p < 0.001$, two-sample t-test; $p = 0.027$, Kolmogorov–Smirnov test). However, the magnitude is modest, with *Human Interviewer* applicants generating a 2.57% higher score compared to *AI Interviewer* applicants.

Recruiter weights on signals. In Table 2, we regress the recruiters' decisions whether to extend a job offer to an applicant on interview, analytical, and language scores. For comparability, we standardize all three variables. We find that all three signals have a significant positive influence on the likelihood that a recruiter extends a job offer to an appli-

Table 2: Predicting job offer decisions of recruiters

	Dependent variable: Job Offer Made	
	(1)	(2)
Interview score (std.)	0.091^{***}	0.100^{***}
	(0.010)	(0.010)
Language test score (std.)	0.108^{***}	0.115^{***}
	(0.010)	(0.010)
Analytical test score (std.)	0.034^{***}	0.028^{***}
	(0.011)	(0.010)
Interview score (std.) $\times$ *AI Interviewer*	-0.047^{***}	-0.029^{**}
	(0.012)	(0.012)
Language test score (std.) $\times$ *AI Interviewer*	0.028^{**}	0.022^{*}
	(0.012)	(0.011)
Analytical test score (std.) $\times$ *AI Interviewer*	-0.003	0.001
	(0.012)	(0.012)
Mean DV in Human Interviewer	0.38	0.38
Controls and fixed effects	–	Yes
Observations	9,965	9,864
R^2	0.118	0.218

Notes: The table shows OLS estimates predicting job offer decisions of recruiters using standardized test scores and interview scores. The dependent variable is a dummy equal to one if an application led to a job offer. Test scores are standardized. "*AI Interviewer*" is a dummy equal to one if the application was in the *AI Interviewer* condition, and zero if the application was in the *Human Interviewer* condition. Controls include an applicant's gender, source of application, pre-treatment engagement score, and whether they have applied before to any of the firm's job postings. Fixed effects include week, recruiter, application side, and job posting fixed effects. An observation is an application. Standard errors in parentheses are clustered at the applicant level. Significance levels: * $p < 0.1$,** $p < 0.05$, *** $p < 0.01$.

cant in the *Human Interviewer* condition. In terms of magnitude, we find that almost equal weight is put on the language and interview score, while less weight is put on the analytical score. Specifically, a one standard deviation increase in interview score is associated with an 8.9 percentage point higher likelihood of offer, keeping performance in language and analytical tests constant, controlling for baseline characteristics and including fixed effects. In contrast, a one standard deviation increase in the language and analytical score is associated with an increase of 10.8 and 3.5 percentage points, respectively.

Importantly, we find a significant interaction when comparing the influence of each

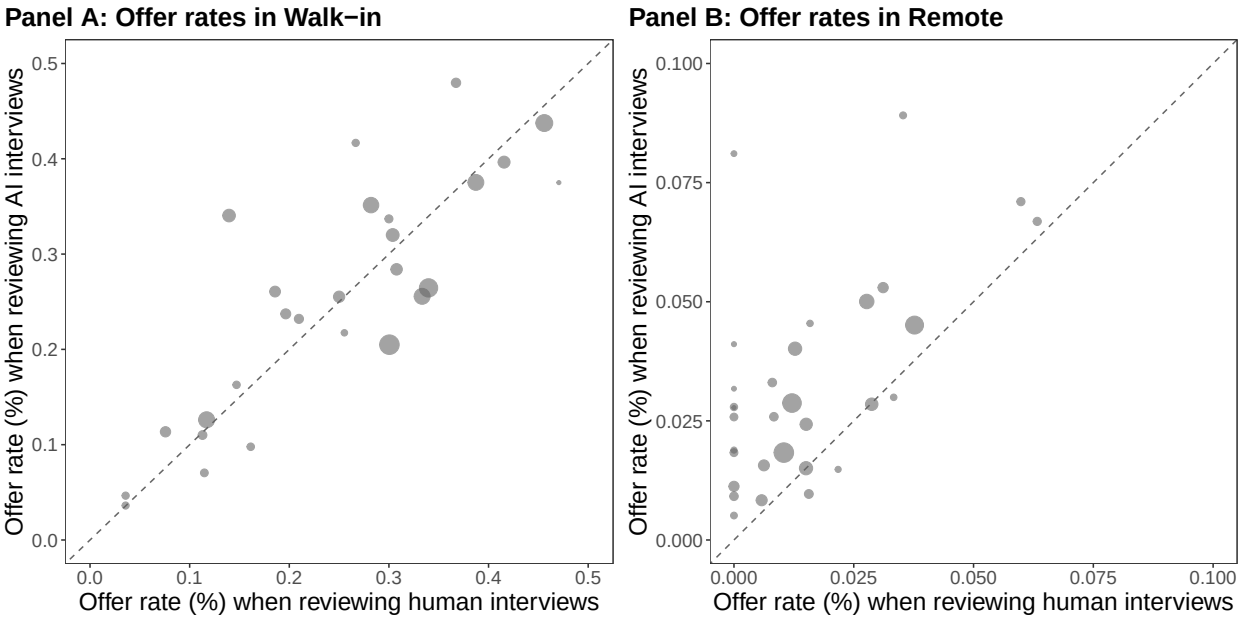
signal in *Human Interviewer* relative to *AI Interviewer*. The language score is significantly more predictive of offers in *AI Interviewer*, while the interview score is significantly less predictive. This suggests that interviewers place more weight on the independent quality signal coming from the language test than on the interview when the signal from the interview comes from the AI and not from themselves conducting the interview.

Mechanism. To further investigate whether the interaction effect reflects recruiters' differential weighting of signals rather than other influences, we use their responses from the recruiter survey. Specifically, we use their response to the question of how important they consider the interview and test score in making offer decisions. As documented previously, there is a sufficient degree of heterogeneity in their responses. We exploit this heterogeneity by repeating the analysis of Table 2 separately for the sample of recruiters who state in the recruiter survey that the interview score is equally or less important than the test scores in determining their offer decisions, and recruiters who state that interview scores are more important. We would expect the previously documented interaction effect – namely, that interview scores are less predictive of offer decisions in the *AI Interviewer* condition relative to the *Human Interviewer* condition – to be stronger among recruiters who consider interview performance more important. We indeed find that the interaction of treatment with interview score is more negative (and only statistically significant) for recruiters who consider the interview as more important than the standardized scores. For details, see Appendix Table B.14. This suggests that recruiters indeed weigh the signals from AI-led interviews differently.

6.3 Recruiter heterogeneity analysis

We now turn to analyzing the heterogeneity in behavior between recruiters. We focus on recruiters who were assigned to review at least 25 applications in *Human Interviewer* and at least 25 applications in *AI Interviewer*. This leaves us with 61 recruiters from a total of 131 from the full sample.

Figure 4: Average offer rate of recruiters across treatments



Notes: The unit of observation is a recruiter. Dot size indicates the total number of interviews assigned to each recruiter for review.

Offer rates across recruiters. We are interested in two types of heterogeneity on the recruiter level. First, how much do offer rates differ between recruiters (level difference)? Second, how much do offer rates differ between treatments across recruiters (slope difference)? To answer these questions, we calculate the average offer rate of each recruiter, separately for the *Human Interviewer* and *AI Interviewer* conditions. In Figure 4, we then plot as a scatter plot the resulting average offer rates across conditions for each recruiter.

We find sizable variation in average offer rates among recruiters in both the *Walk-in* (Panel A) and *Remote* mode (Panel B). That is, recruiters differ in their general propensity to extend offers to applicants. These differences could be caused by recruiter-specific traits, regional differences in the application pool, or characteristics of the job opening for the respective client firm. Importantly, however, offer rates are highly correlated across treatments (Full sample: $\rho = 0.87$, $p < 0.001$, *Walk-in:* $\rho = 0.83$, $p < 0.001$, *Remote:* $\rho = 0.47$, $p = 0.007$). That is, recruiters who have, on average, a high offer rate when reviewing human interviews also have a high offer rate when reviewing interviews from the AI voice agent.

In terms of aggregate offer behavior, we find that 69% of recruiters have a higher average offer rate when reviewing interviews conducted by the AI Voice agent relative to reviewing human interviews, while 31% of recruiters have the reverse. Accordingly, our

main effect of AI applications receiving higher offer rates is not driven by a small number of recruiters differentiating strongly between AI and human interviews. Instead, it appears to be driven by the majority of recruiters.

Role of recruiter experience. Next, we investigate whether recruiters' experience handling AI and human interviews matters for their decision-making. Note that the firm has piloted the roll-out of the AI-led interviews in the weeks prior to the start of the experiment. Hence, most recruiters have been exposed to and worked with AI-led interviews. Accordingly, the process was not entirely new to them. Nevertheless, our recruiters differ substantially in the number of interviews they conduct over the course of the experiment. This allows us to analyze the correlation between the number of interviews recruiters are assigned to review and their propensity to extend offers.

We start by investigating whether a general experience effect exists, i.e., the total number of interviews assigned to be reviewed is correlated with the likelihood of extending an offer. We find a significant association: regressing offer rates on total number reveals that, on average, reviewing 10 additional interviews is associated with a -0.06 percentage point lower offer rate. For details, see Appendix Table B.18 column (1). Note that we are not claiming that experiencing more interviews has a causal effect on offers or that it is the effect of exposure to interviews. There are likely recruiter-specific factors that influence both offer rates and the number of interviews reviewed. For instance, there are likely skill differences among recruiters in how fast and efficiently they handle applications.

Instead, we are mainly interested in the association of experience with offer rates *differs* between the *Human Interviewer* and *AI Interviewer* conditions. Accordingly, we interact the number of interviews reviewed with a treatment dummy. We find a significant and positive interaction effect ($p = 0.02$, column (2) of Appendix Table B.18). Hence, the negative association of the number of applications with offer rates is less pronounced in the *AI Interviewer* condition relative to the *Human Interviewer* condition.

7 Organizational returns to AI automation

In this section, we analyze the effects of treatment on operational outcomes. First, we examine how adoption of the AI voice agent affects *time-to-hire*, shifting the queuing bottleneck from scheduling and interview completion to the evaluation stage. Second, we analyze *cost efficiency*, comparing the costs of implementing the AI agent with the costs of human recruiters.

7.1 The impact of AI adoption on recruitment process length

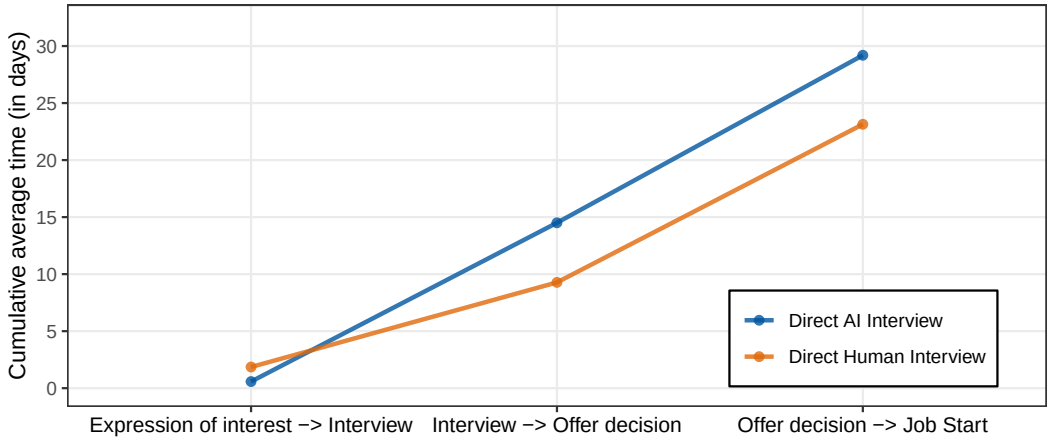
Time-to-hire variables, AI adoption and decomposition. We analyze time differences between three key stages of the recruitment process, which combined make up *time-to-hire*: stage (1) from receiving applicants' *Expression of interest* to an *Interview* taking place (scheduling time), stage (2) from *Interview* to *Offer decision* (evaluation time), and stage (3) from *Offer* to *Job Start* (onboarding time). These time-stamps are available only for the *Remote* mode, so we focus on this mode.

Adopting the AI voice agent is expected to affect the first stages of time-to-hire. As the agent is available 24/7, applicants can self-schedule the interview promptly, plausibly shortening the time from expression of interest to interview. However, after the interview, human recruiters may take longer to evaluate AI-led interviews because they did not conduct the conversation themselves, potentially lengthening the interview-to-offer interval. By contrast, the offer-to-job-start interval is largely standardized and runs independently of interview modality; any time differences here likely reflect applicant characteristics (e.g., responsiveness, document submission speed) or client-side processing rather than interviewer or recruiter differences.

Results. We analyze the timing of successful hires, that is, those who ultimately started their job. Figure 5 displays the results. We find that the median time (average time) from *Expression of interest* to *Interview* is 0.32 days (0.58) in *AI Interviewer* and 0.51 (1.86) in *Human Interviewer*. Both the difference in medians ($p < 0.001$, two-sample Wilcoxon signed-rank test) and averages are significant ($p = 0.094$, two-sample t-test), albeit the latter only at the 10% level. From the first stage to the second stage (*Interview* to *Offer decision*), applications in *AI Interviewer* take a median (average) time of 7.24 days (13.92), while in *Human Interviewer*, the time is 2.62 (7.42). Accordingly, the time is much smaller in *Human Interviewer* compared to *AI Interviewer* ($p = 0.005$, two-sample Wilcoxon signed-rank test; $p = 0.005$, two-sample t-test). Lastly, regarding the time from *Offer* to *Job Start*, this takes AI-led interview applicants a median (average) time of 13.00 (14.69) days, while for Human-led interview applicants it takes 12.00 (13.86) days. Hence, as expected, we find no difference in speed across both conditions ($p = 0.55$, two-sample Wilcoxon signed-rank test; $p = 0.75$, two-sample t-test).

Analyzing the overall time from *Expression of interest* to *Job Start* (all three stages combined), we find that *AI Interviewer* applicants take a median (average) time of 24.00 days (29.19), while in *Human Interviewer*, the time is 20.00 (23.14). These differences are significant at the 5% (10%) level ($p = 0.033$, two-sample Wilcoxon signed-rank test; $p = 0.096$, two-sample t-test). Accordingly, the time saved by employing AI voice agents in inter-

Figure 5: Difference in time to next recruitment step per treatment



Notes: The figure displays the time (in days) it takes applications to reach different stages of the recruitment process, split by treatment condition. "Expression of interest → Interview" denotes the time it takes from an applicant's expression of interest to an interview taking place. "Interview → Offer decision" denotes the time it takes from an interview taking place to a recruiter making an offer decision. "Offer decision → Job Start" denotes the time it takes from a recruiter making an offer decision to the applicant starting their job.

views is offset by human recruiters taking longer to review them, leading to a slightly longer overall recruitment process.[22]

7.2 Cost benchmarks for AI and human interviews

We assess the cost-effectiveness of AI voice interviewers versus human interviewers by analyzing three cost scenarios that differ in their underlying assumptions about the cost structure. Below, we provide a summary of the scenarios and their respective results. In Appendix I, we provide the full details.

Our cost calculation focuses on marginal interview costs, deployment costs, AI system failures, and variations in labor-market costs. Specifically, we compare the per-interview costs of AI-led and human-led interviews based on the partner firm's cost structure. We calibrate three labor market wage levels (based on industry estimates) and three AI pricing scenarios (guided by vendor pricing) to construct nine distinct settings.

Static cost comparison. We start by comparing costs under the assumption that interview costs are static and deterministic. We find that AI interviewing is cost-effective in eight out of the nine settings (see Panel A of Appendix Table I.1 for details). In low-wage environments, AI requires between approximately 8,500 and 24,000 interviews to break

[22]This result reflects a classic queue optimization problem (Hassin and Haviv, 2003), now in the context of AI-led interviews, where time saved upfront must be balanced against longer downstream evaluations by humans.

even at low-to-mid AI prices and is never cost-competitive at high AI prices. In mid-wage environments, AI breaks even after roughly 4,500 to 7,000 interviews. In high-wage environments, AI becomes cost-effective after only about 2,000 to 3,000 interviews, even under conservative AI pricing. Thus, holding interview quality fixed, relative factor prices alone imply that AI adoption scales most rapidly where human labor is expensive.

Dynamic costs with early-stage AI failures. Next, we examine a dynamic cost model in which AI costs evolve over time. We calculate one-year break-even thresholds under a conservative effective failure rate of 5%. Relative to the static benchmark, incorporating these dynamics increases break-even interview counts by approximately 7–57%, depending on the wage environment and AI price tier (see Panel B of Appendix Table I.1). The smallest increases arise in high-wage environments (around 7–11%), while the largest occur when human wages are moderate, and AI prices are high (up to 57%). Despite these increases, AI remains cost-effective in all mid- and high-wage environments, and in low-wage environments at low and mid AI prices, breaking even after roughly 5,000 interviews in mid-wage markets and 2,000–3,300 interviews in high-wage markets, while remaining unattractive in low-wage, high-AI-price scenarios.

Stochastic cost variation. Finally, we allow interview costs to vary stochastically, capturing factors such as volatility in labor costs and AI vendor pricing. Using Monte Carlo simulations, we find that volatility in human labor costs further strengthens AI's cost advantage. Increased variability in human interview costs reduces the median break-even scale for AI adoption, while higher correlations between human and AI cost shocks attenuate this advantage. Across all nine settings, higher wage environments shift the entire break-even frontier inward, implying faster recovery of fixed AI deployment costs.

8 Conclusion

In this paper, we examine the economic and behavioral consequences of substituting human interviewers with AI voice agents. Our large-scale natural field experiment provides empirical evidence on both the capabilities and challenges of deploying AI voice agents. Specifically, we provide causal evidence that automating interviews increases offer rates and retention rates without changing the composition of exits and worker productivity, indicating improvements in match quality.

Analyzing the content of the interviews, our results suggest that these gains arise from a mechanism we label *controlled variance*: AI interviewers execute the firm's structured interview protocol with greater consistency across interviews while still adapting to each

interview through responsive follow-ups. This consistency reduces interviewer dispersion in information collection, yielding interviews that are both more standardized and more informative for hiring decisions. As a result, AI agents replicate the information-collection and screening functions traditionally handled by human recruiters at least as effectively, and in some dimensions, outperform humans.

At the same time, AI automation reshapes behavior within the hiring process. Recruiters discount AI signals, i.e., the information collected through AI-led interviews. While they score interview performance higher, they place greater weight on test scores in their hiring decisions. On the applicant side, a majority of applicants choose to be interviewed by AI agents. Yet this choice reveals negative sorting where lower-ability candidates are more likely to choose AI.

Altogether, our findings show both the returns of AI as an information-collection technology and its infrastructural challenges to seize these returns. Firms can increase the efficiency of screening by interviews through automated standardization, but realizing the full returns from automation requires complementary adaptation in how humans rely on AI signals for decision-making. The benefits of automation appear highest in high-volume, high-turnover environments where tasks are repetitive, outcomes are rapidly observable, and, importantly, where variance in human performance imposes costs. In contrast, settings that depend on tacit knowledge, relational inference, or screening of highly specialized skills may benefit more from human screening.

While our results suggest firm-level efficiency gains, their broader implications depend on how such emerging screening technologies diffuse and interact across markets. As AI agents increasingly perform standardized tasks that collect information from humans through natural language, the composition of recruiter work and the demand for complementary human skills are likely to shift. Potentially, this will compress variation in hiring outcomes while reallocating human effort toward more cognitively complex tasks, at least temporarily. The welfare impact of AI in hiring thus extends beyond the direct productivity improvements we document: it hinges on how substitution and complementarity unfold across tasks, firms, and sectors as automation propagates through the economy. We see understanding empirically this broader reallocation of economic roles between humans and AI agents, within firms and across markets, as key to designing efficient and welfare-enhancing institutions under AI-induced automation.

References

Acemoglu, Daron, and Pascual Restrepo (2019). "Automation and new tasks: How technology displaces and reinstates labor." *Journal of Economic Perspectives* 33 (2): 3–30. [6]

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz (2024). "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." *Working Paper*, [7]

Agarwal, Nikhil, Alex Moehring, and Alexander Wolitzky (2025). "Designing Human-AI Collaboration: A Sufficient-Statistic Approach." *Working Paper*, [7]

Aka, Ada, Emil Palikot, Ali Ansari, and Nima Yazdani (2025). "Better Together: Quantifying the Benefits of AI-Assisted Recruitment." *Working Paper*, [6]

Alam, Md Ferdous, Austin Lentsch, Nomi Yu, Sylvia Barmack, Suhin Kim, Daron Acemoglu, John Hart, Simon Johnson, and Faez Ahmed (2024). "From Automation to Augmentation: Redefining Engineering Design and Manufacturing in the Age of NextGen-AI." *An MIT Exploration of Generative AI*, [6]

"American Trends Panel Wave 119" (2022). Pew Research Center. [92]

Angelova, Victoria, Will Dobbie, and Crystal Yang (2023). "Algorithmic Recommendations and Human Discretion." *Working Paper*, [7]

Athey, Susan C., Kevin A. Bryan, and Joshua S. Gans (2020). "The Allocation of Decision Authority to Human and Artificial Intelligence." *AEA Papers and Proceedings* 110: 80–84. [6]

Autor, David (2024). "Applying AI to rebuild middle class jobs." *Working Paper*, [6]

Autor, David, and Neil Thompson (2025). "Expertise." *Journal of the European Economic Association* 23 (4): 1203–71. [6]

Autor, David H., and David Scarborough (2008). "Does Job Testing Harm Minority Workers? Evidence from Retail Establishments*." *Quarterly Journal of Economics* 123 (1): 219–77. [12]

Avery, Mallory, Andreas Leibbrandt, and Joseph Vecci (2024). "Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech." *Working Paper*, [6]

Awuah, Kobbina, Urša Krenk, and David Yanagizawa-Drott (2025). "Automation with Generative AI? Evidence from a Teacher Hiring Pipeline." *Working Paper*, [6]

Berg, Jeff, Avinash Das, Vinay Gupta, and Paul Kline (2018). "Smarter Call-Center Coaching for the Digital World." Technical Report. New York: McKinsey & Company. [9]

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin

Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners." *Working Paper*, [72]

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond (2025). "Generative AI at Work." *Quarterly Journal of Economics* 140 (2): 889–942. [5, 19]

Brynjolfsson, Erik, and Andrew McAfee (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & company. [6]

Buesing, Eric, Vinay Gupta, Sarah Higgins, and Raelyn Jacobson (2020). "Customer Care: The Future Talent Factory." Technical Report. New York: McKinsey & Company. [9]

Chen, Zenan, and Jason Chan (2024). "Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise." *Management Science* 70 (12): 9101–17. [5]

Chen, Zhisheng (2023). "Ethics and discrimination in artificial intelligence-enabled recruitment practices." *Humanities and social sciences communications* 10 (1): 1–12. [31]

Choi, James J, Dong Huang, Zhishu Yang, and Qi Zhang (2025). "How Good Is AI at Twisting Arms? Experiments in Debt Collection." *Working Paper*, [6]

Choi, Jonathan H., and Daniel B. Schwarcz (2023). "AI Assistance in Legal Analysis: An Empirical Study." *Working Paper*, [5]

Chopra, Felix, and Ingar Haaland (2023). "Conducting Qualitative Interviews with AI." *Working Paper*, [6]

Costello, Thomas H., Gordon Pennycook, and David G. Rand (2024). "Durably Reducing Conspiracy Beliefs through Dialogues with AI." *Science* 385 (6714): eadq1814. [6]

Cowgill, Bo (2020). "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening." *Working Paper*, [6]

Dargnies, Marie-Pierre, Rustamdjan Hakimov, and Dorothea Kübler (2025). "Behavioral Measures Improve AI Hiring: A Field Experiment." *Working Paper*, [6]

Dargnies, Marie-Pierre, Rustamdjan Hakimov, and Dorothea Kübler (2026). "Aversion to Hiring Algorithms: Transparency, Gender Profiling, and Self-Confidence." *Management Science* 72 (1): 285–301. [7]

Dell'Acqua, Fabrizio, Edward III McFowland, Ethan Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani (2025). "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." *Working Paper*, [5]

Doshi, Anil R., and Oliver P. Hauser (2023). "Generative Artificial Intelligence Enhances Creativity but Reduces the Diversity of Novel Content." *Working Paper*, [5]

Fumagalli, Elena, Sarah Rezaei, and Anna Salomons (2022). "OK Computer: Worker Perceptions of Algorithmic Recruitment." *Research Policy* 51 (2): 104420. [7]

Geiecke, Friedrich, and Xavier Jaravel (2024). "Conversations at Scale: Robust AI-led Interviews with a Simple Open-Source Platform." *Working Paper*, [6]

Gruber, Jonathan, Benjamin R Handel, Samuel H Kina, and Jonathan T Kolstad (2020). "Managing intelligence: Skilled experts and AI in markets for complex products." *Working Paper*, [6]

Hassin, Refael, and Moshe Haviv (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*. vol. 59, Springer Science & Business Media. [39]

Hernandez, Edrin (2024). "Analyzing BPO Employment Statistics Philippines: Trends and Insights." Technical Report. Magellan Solutions. [9]

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li (2018). "Discretion in Hiring." *The Quarterly Journal of Economics* 133 (2): 765–800. [6, 12]

Hoffman, Mitchell, and Christopher Stanton (2024). "People, Practices, and Productivity: A Review of New Advances in Personnel Economics." *Working Paper*, [6]

Horton, John J. (2017). "The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment." *Journal of Labor Economics* 35 (2): 345–85. [6]

Jovanovic, Boyan (1979). "Job Matching and the Theory of Turnover." *Journal of Political Economy* 87 (5, Part 1): 972–90. [12]

Kausel, Edgar E., Satoris S. Culbertson, and Hector P. Madrid (2016). "Overconfidence in Personnel Selection: When and Why Unstructured Interview Information Can Hurt Hiring Decisions." *Organizational Behavior and Human Decision Processes* 137: 27–44. [2]

Kumar, Harsh, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman (2025). "Math Education With Large Language Models: Peril or Promise?" In *Artificial Intelligence in Education*. vol. 15880, Lecture Notes in Computer Science Springer, Cham. [5]

Levashina, Julia, Christopher J. Hartwell, Frederick P. Morgeson, and Michael A. Campion (2014). "The Structured Employment Interview: Narrative and Quantitative Review of the Research Literature." *Personnel Psychology* 67 (1): 241–93. [2]

Li, Danielle, Lindsey Raymond, and Peter Bergman (2025). "Hiring as Exploration." *Review of Economic Studies*, [6]

McDaniel, Michael A., Deborah L. Whetzel, Frank L. Schmidt, and Steven D. Maurer (1994). "The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis." *Journal of Applied Psychology* 79 (4): 599–616. [2]

Mocanu, Tatiana (2025). "Designing Gender Equity: Evidence from Hiring Practices." *Working Paper*, [2]

Noy, Shakked, and Whitney Zhang (2023). "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence." *Science* 381 (6654): 187–92. [5]

Otis, Nicholas G, Solène Delecourt, Katelyn Cranney, and Rembrand Koning (2024). "Global evidence on gender gaps and generative AI." *Working Paper*, [31]

Otis, Nicholas G., Rowan Clarke, Solène Delecourt, David Holtz, and Rembrand Koning (2025). "The Uneven Impact of Generative AI on Entrepreneurial Performance." *Working Paper*, [5]

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer (2023). "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot." *Working Paper*, [5]

Radbruch, Jonas, and Amelie Schiprowski (2025). "Interview Sequences and the Formation of Subjective Assessments." *Review of Economic Studies* 92 (2): 1226–56. [2]

Sallaz, Jeffrey J. (2019). *Lives on the Line: How the Philippines Became the World's Call Center Capital*. New York, NY: Oxford University Press. [9]

Shukla, Soumitra (2025). "Making the Elite: Class Discrimination at Multinationals." *Working Paper*, [2]

Silberschatz, Abraham, Peter B. Galvin, and Greg Gagne (2018). *Operating System Concepts*. 10th. Hoboken, NJ: John Wiley & Sons. See Chapter 5, "CPU Scheduling," for the round-robin algorithm. [10]

Stevenson, Megan T., and Jennifer L. Doleac (2024). "Algorithmic Risk Assessment in the Hands of Humans." *American Economic Journal: Economic Policy* 16 (4): 382–414. [7]

Vatsal, Shubham, and Harsh Dubey (2024). "A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks." *arXiv preprint arXiv:2407.12994*, [72]

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *arXiv preprint arXiv:2201.11903*, [72]

Wiles, Emma, and John J Horton (2024). "More, but Worse: The Impact of AI Writing Assistance on the Supply and Quality of Job Posts." *Working Paper*, [5]

Appendix

A Additional figures

Figure A.1: Recruitment process



Figure A.2: Distribution of applications by recruitment center location

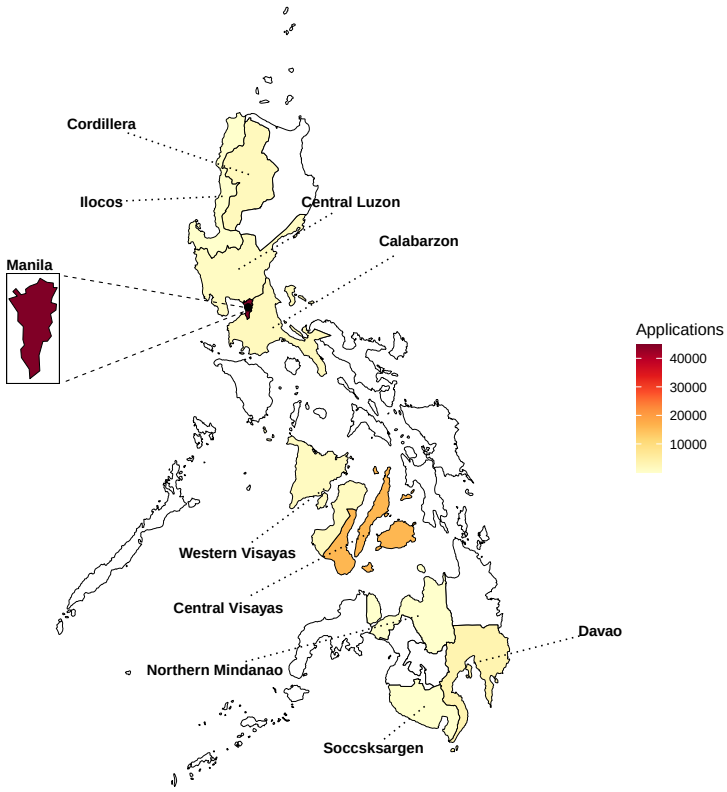
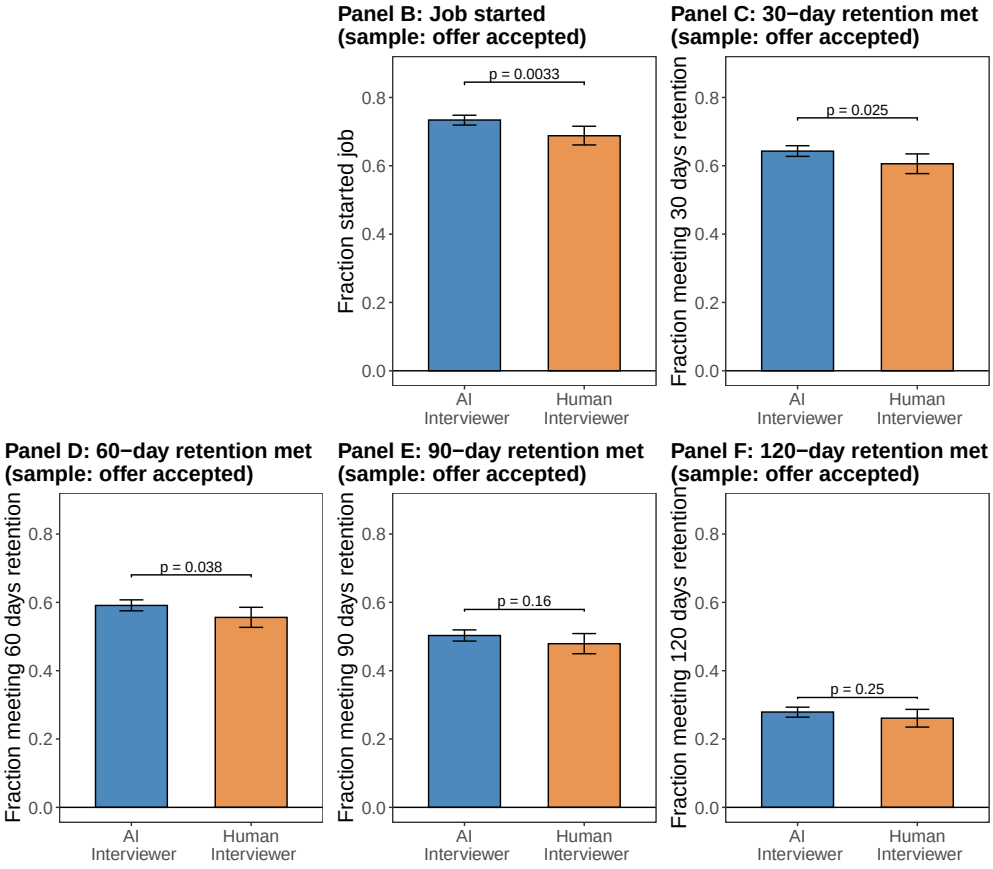
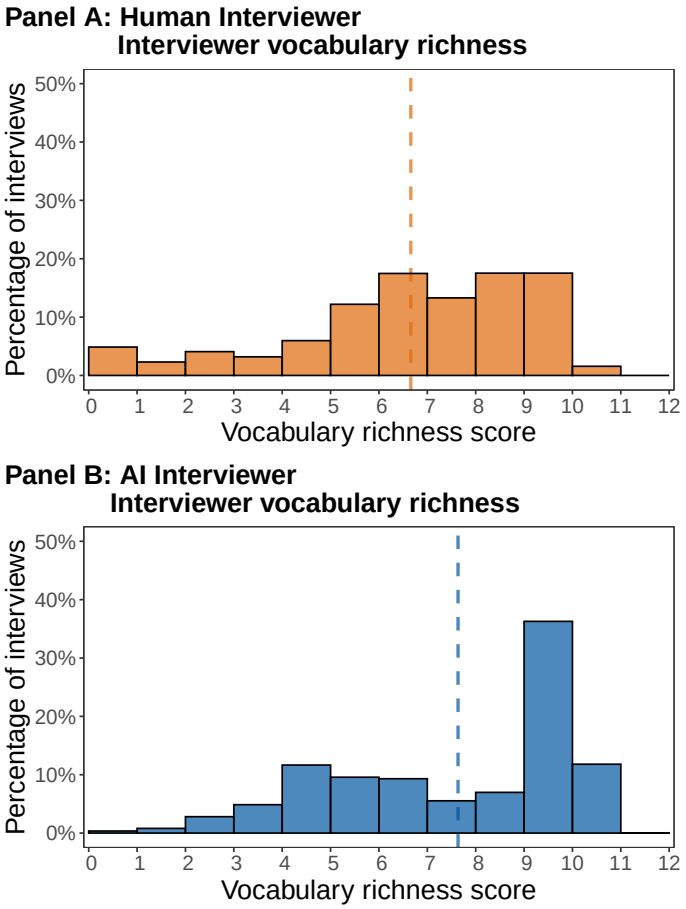


Figure A.3: Treatment effect on key recruiting outcomes in the conditional sample



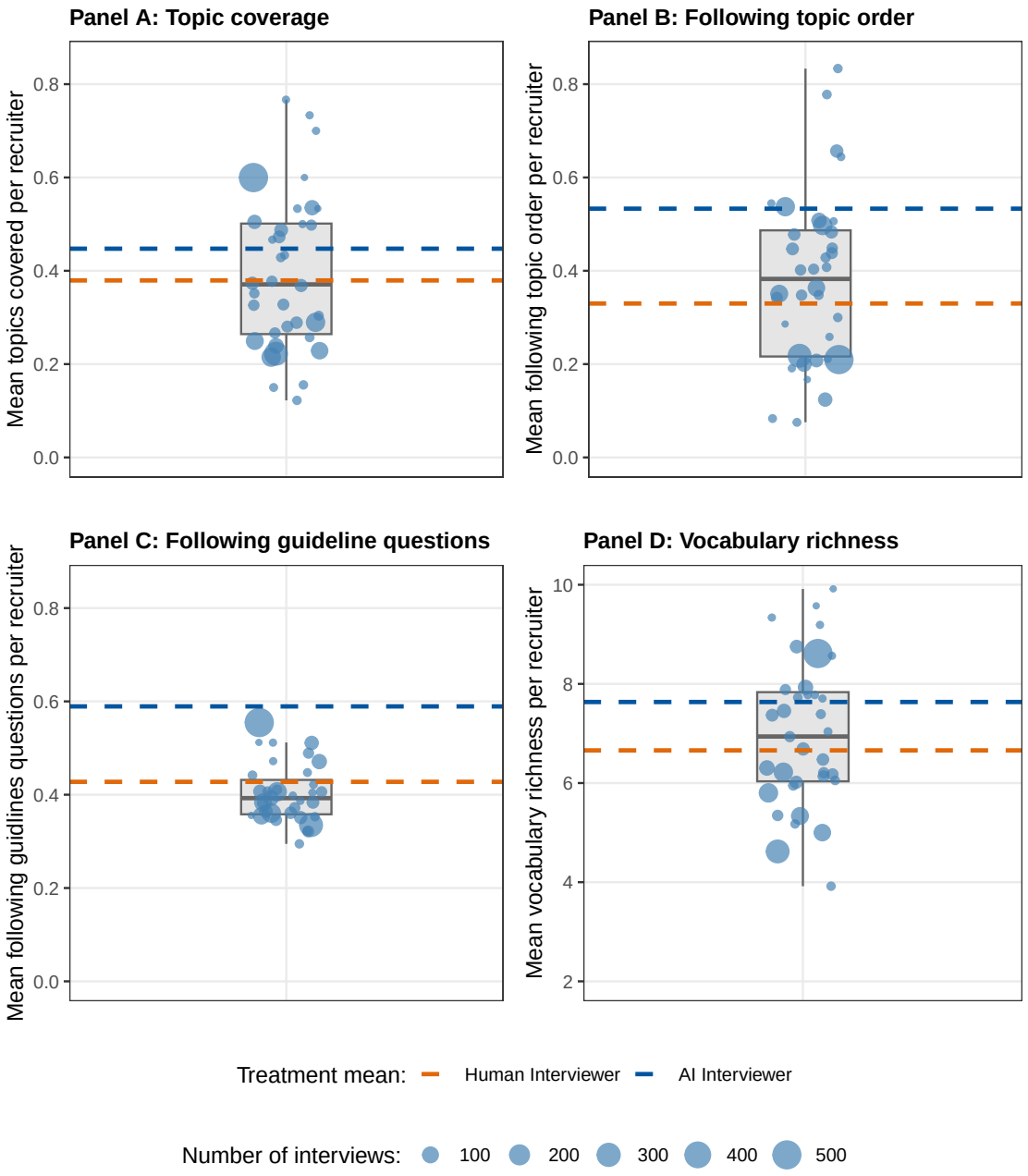
Notes: The figure displays the recruiting outcomes of applicants, conditional on those applicants who accepted their job offer. Each panel displays the fraction of applicants who realize the specific outcome. Fractions are displayed separately for the *Human Interviewer* condition, in which applicants are interviewed by a human, and for the *AI Interviewer* condition, in which applicants are interviewed by an AI voice agent. Bars indicate 95% confidence intervals; p-values calculated from a two-sample proportion test.

Figure A.4: Treatment differences in interviewers' language in interviews



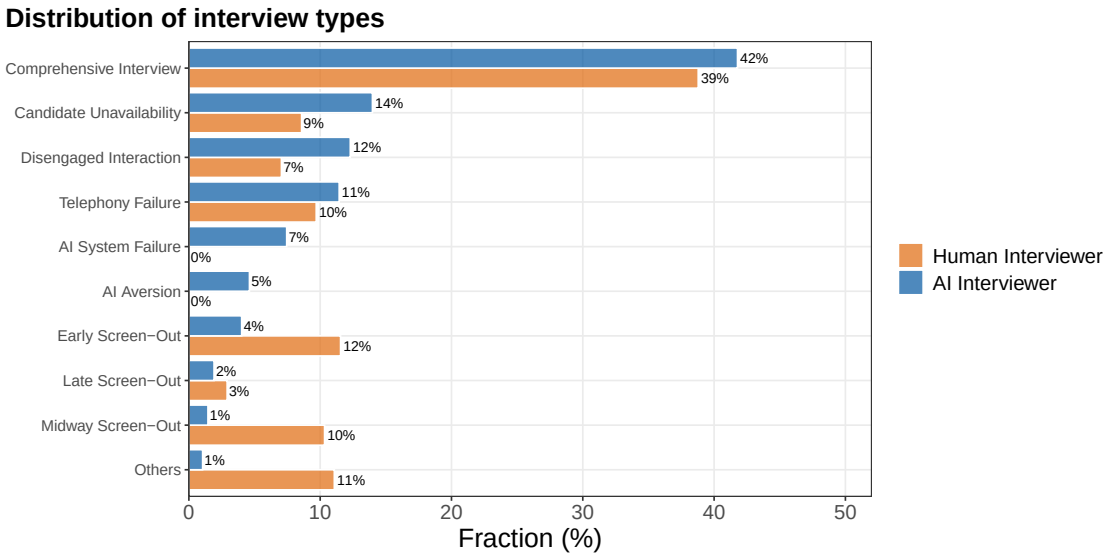
Notes: The figure displays the distribution of the vocabulary richness score of recruiters separately for both treatments.

Figure A.5: Distribution of interviewers' language averaged per interviewer across interviews



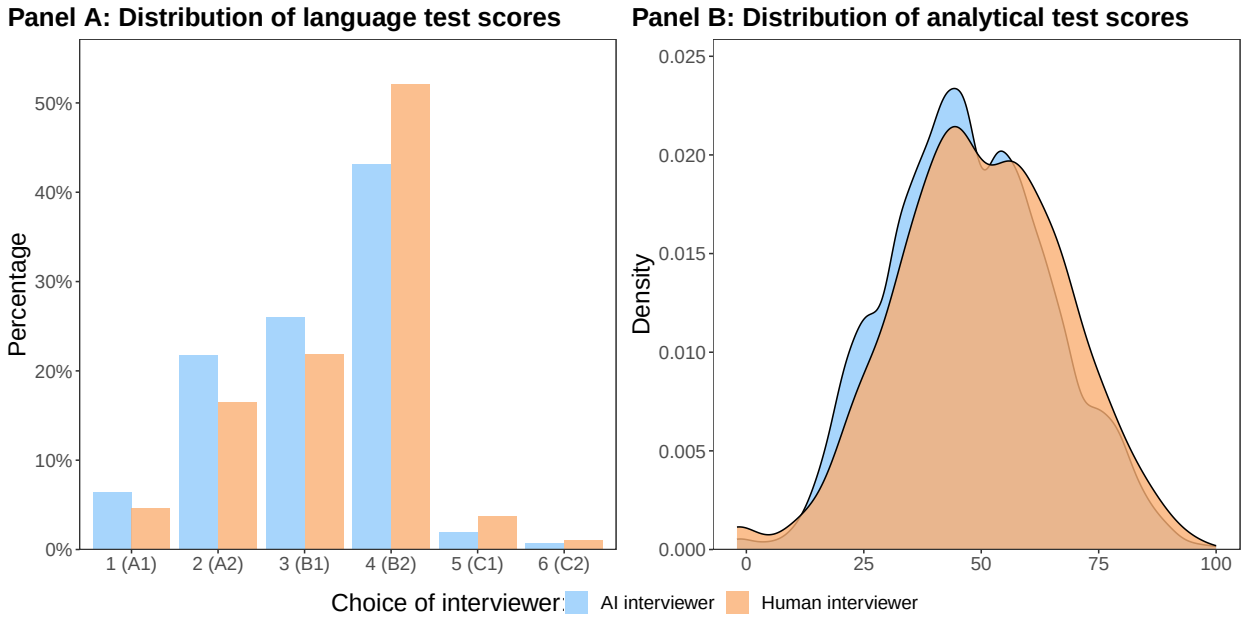
Notes: This figure shows the distribution of interviewer-level averages of four measures of interviewer language behavior. Each panel plots one variable at the recruiter level: (A) unique topic coverage, (B) topic order correlation, (C) topic–question similarity, and (D) vocabulary root richness. Each dot corresponds to a single human recruiter, with the size of the dot proportional to the number of interviews conducted by that recruiter in the *Human Interviewer condition*. The box plots summarize the distribution across recruiters: the box spans the interquartile range (25th to 75th percentile), the horizontal line indicates the median, and whiskers extend to 1.5 times the interquartile range. Dashed horizontal lines show the mean of the corresponding variable at the interview level in each treatment group. For definitions and construction of the variables, see Appendix Table E.2.

Figure A.6: Transcript variables per treatment



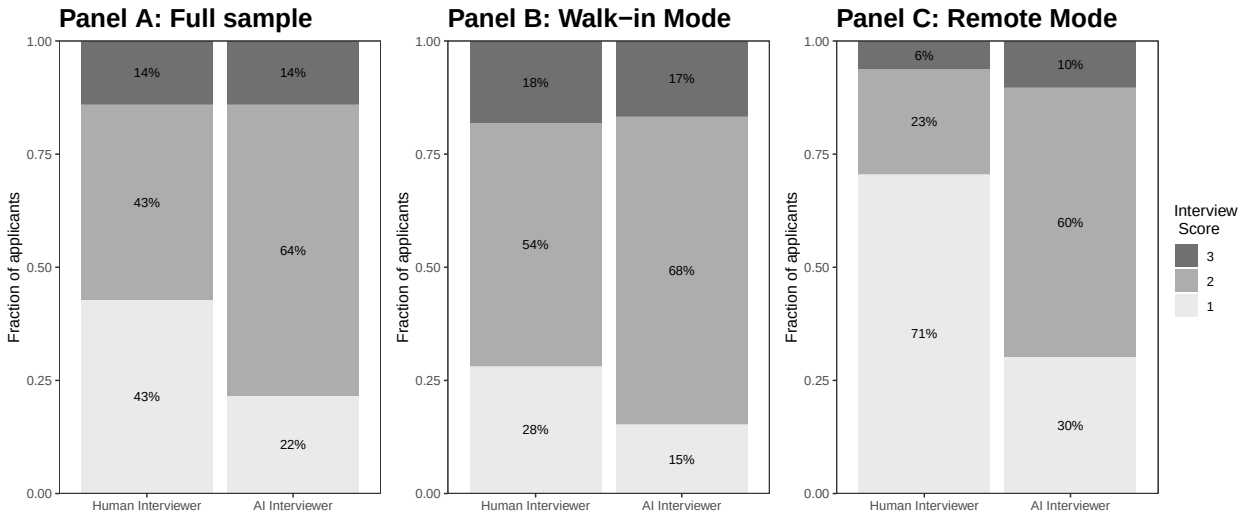
Notes: The figure displays the fraction of interview transcripts that fall in each of a number of interview types. Appendix Table E.1 provides the details about each interview type's definition.

Figure A.7: Distribution of standardized test scores depending on applicants' choice of interviewer



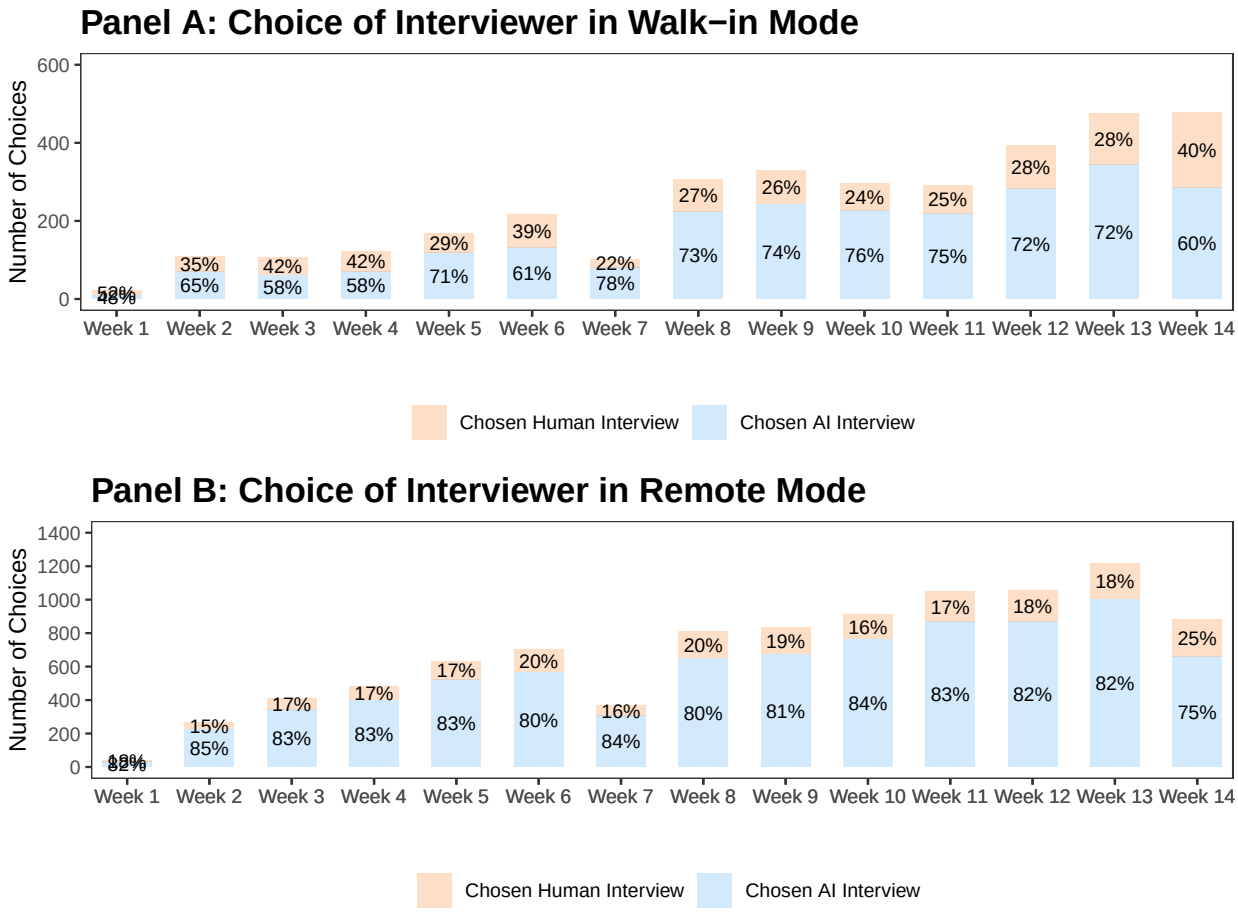
Notes: The figure displays the distribution of test score results of applicants, split by their interviewer choice in the *Choice of interviewer* condition. Panel A displays the results from the language test, which assesses applicants' writing and reading capabilities in English. Scores are based on the CEFR framework (A1 to C2). Panel B displays the results from the analytical test, which assesses in three parts applicants' attention to detail, verbal reasoning, and numerical ability. Scores are aggregated from each of the three parts and range from 0 to 100.

Figure A.8: Distribution of interview scores across treatments



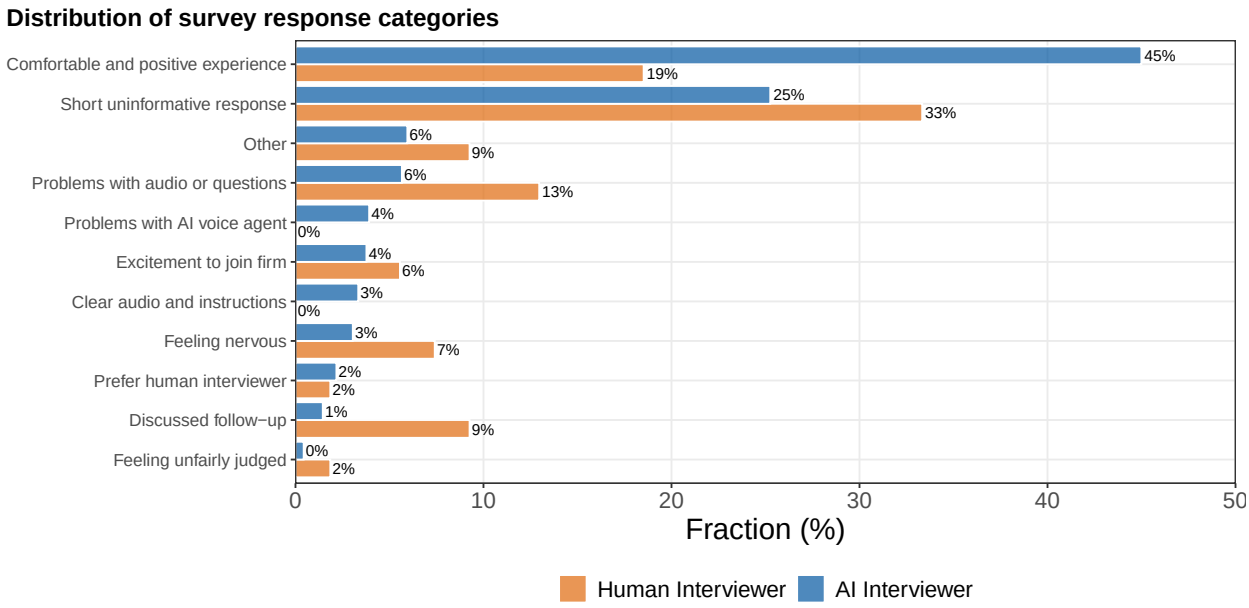
Notes: The figure displays the distribution of the interview score with which recruiters assess each interview. The score is 1-poor, 2-medium, 3-good. For details on the scoring, see Appendix Table B.3. Panel A displays the full sample results. Panel B displays the *Walk-in mode* subsample, in which applicants approached the firm at one of the firm's recruitment centers. Panel C displays the *Remote mode* subsample, in which applicants approached the firm online.

Figure A.9: Choices of interviewer in the *Choice of interviewer* condition over time



Notes: The figure displays applicants' interviewer choice in the *Choice of interviewer* condition over the course of the experiment. In the condition, applicants were given the choice between a human interviewer and the AI voice agent after being invited to the job interview. The experiment ran from March 7 to June 7, 2025. In the *Walk-in mode* (Panel A), applicants approached the firm at one of the firm's recruitment centers. In the *Remote mode* (Panel B), they approached the firm online.

Figure A.10: Distribution of open-ended survey responses



Notes: The figure displays the fraction of survey responses for each response category. Responses are from the customer experience survey fielded to applicants. Applicants responded in an open-ended text field whether they had any additional feedback to share about their interview experience. Responses were classified using an LLM. For the definition of the response categories and example responses, see Appendix Table B.16.

B Additional tables

Table B.1: Client firm industries

Client firm sector	Percentage (%) $(n = 41)$
Technology	22.00
Finance	14.60
Transportation	14.60
Tourism	9.80
Healthcare	7.30
Insurance	7.30
Telecommunication	7.30
Hospitality	4.90
Manufacturing	4.90
Retail	4.90
Energy	2.40

Notes: Distribution of client firm industries for which the company is hiring customer service representatives.

Table B.3: Interview performance scoring system

Score	Label	Assessment	Applicant engagement	Outcome predictions
3	Good	Clear, concise communication; strong problem-solving and critical thinking; solid grasp of role and firm	Shows keen interest; asks relevant questions	High probability of accepting offer and high job show rates and performance
2	Medium	Adequate communication; basic problem-solving; satisfactory but improvable grasp of role and firm	Moderate interest; engages intermittently	Uncertain acceptance; average job show rates and performance
1	Poor	Incoherent or disorganised answers; weak problem-solving; little grasp of role or firm	Low interest; few or irrelevant questions	Low acceptance likelihood; high likelihood of low job show rates and performance

Table B.2: Voice AI and Human Interview structure (14 core topics)

Topic	Applicability	Example of question	Key signal(s)
Introduction	All	"Hi [name], thanks for applying to [Employer]. Is now still a good time to chat about your experience?"	Professionalism; applicant readiness
Source verification	All	"Out of curiosity, where did you see this opening advertised?"	Channel efficacy; genuine interest
Location and commute	On-site roles	"Where are you based and roughly how long would the drive to our [city] site take?"	Commute feasibility; punctuality risk
Motivation & attrition risks	All	"What attracted you to this opportunity and how does it align with your longer-term goals?"	Engagement; values alignment; Availability
Education verification	All	"Let's talk education—what's the highest level you finished, and do you foresee returning to school?"	Job readiness; upskilling intent
Compensation expectations	All	"The role pays between [range]. Where do your salary expectations sit?"	Pay realism; negotiation stance
Employment history	$\geq$1 prior job	"Walk me through your recent call-center roles—volumes handled, key results, and why you moved on."	Experience depth; performance flags
Re-hire eligibility check	Former employees	"Have you worked for [Employer] before? If so, where and who was your supervisor?"	Prior standing; boomerang potential
Availability	All	"If selected, when could you start? Are you deep in any other interview processes?"	Speed-to-hire; offer risk
Data verification	All	"For our records, could you confirm the best phone, Viber, and an emergency contact?"	Contact accuracy; compliance
Needs assessment	Remote	"Do you have reliable internet and a laptop/PC at home for assessments?"	Tech readiness
Profiling	All	"Have you ever worked for [firm] before?"	Experience depth
Further procedure	All	[Explanation of further procedure]	Tech readiness
Wrap-up & referrals	All	"That's everything from my side—any questions for me? And do you know anyone else who might thrive here?"	Applicant curiosity; referral leads

Table B.4: Treatment balance tests

Variable	*Human Interviewer* (1)	*AI Interviewer* (2)	*Choice of Interviewer* (3)	H_0: (1) = (2) p-value (4)	H_0: (1) = (3) p-value (5)	H_0: (2) = (3) p-value (6)
Panel A: Full sample						
Gender (Women=1)	0.60	0.60	0.61	0.21	0.50	0.04
Source is referral	0.19	0.19	0.19	0.96	0.38	0.26
Source is digital ad	0.59	0.59	0.59	0.47	0.85	0.34
Mode is Walk-in	0.26	0.26	0.26	0.45	0.66	0.85
Initial engagement score	37.06	37.12	37.21	0.80	0.62	0.73
Observations	13,557	40,103	13,396			
Panel B: *Walk-in mode*						
Gender (Women=1)	0.56	0.56	0.56	1.00	0.84	0.86
Source is referral	0.12	0.11	0.11	0.54	0.62	0.36
Source is digital ad	0.05	0.06	0.06	0.74	0.70	0.55
Observations	3,478	10,421	3,469			
Panel C: *Remote mode*						
Gender (Women=1)	0.62	0.61	0.62	0.15	0.48	0.02
Source is referral	0.22	0.22	0.22	0.66	0.63	0.30
Source is digital ad	0.77	0.77	0.78	0.66	0.66	0.32
Initial engagement score	50.48	50.42	50.48	0.75	0.99	0.73
Observations	10,079	29,682	9,927			

Notes: Columns (1) - (3) display mean values of variables for the three treatments. “Source is referral” and “Source is digital ad” are binary variables equal to one if the applicant applied through a referral or a digital job advertisement, respectively. Columns (4) - (6) display p-values obtained using pairwise t-tests (variable “Initial engagement score”) or proportion tests (all other variables).

Table B.5: Treatment effect on key recruiting outcomes among all applicants

Panel A:

	Dependent variable:								
	Received job offer			Started job			Employed after one month		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
AI Interviewer	0.0104***	0.0101***	0.0101***	0.0106***	0.0102***	0.0102***	0.0088***	0.0085***	0.0085***
	(0.0028)	(0.0026)	(0.0037)	(0.0023)	(0.0023)	(0.0032)	(0.0022)	(0.0021)	(0.0031)
Mean DV *Human Interviewer*	0.0870	0.0870	0.0870	0.0565	0.0571	0.0571	0.0497	0.0503	0.0503
Controls and fixed effects	–	Yes	Yes	–	Yes	Yes	–	Yes	Yes
Clustering	App.	App.	Rec.	App.	App.	Rec.	App.	App.	Rec.
Observations	53,660	52,367	52,367	53,660	52,367	52,367	53,660	52,367	52,367
R^2	0.0002	0.1683	0.1683	0.0004	0.1098	0.1098	0.0003	0.0983	0.0983

Panel B:

	Dependent variable:								
	Employed after two months			Employed after three months			Employed after four months		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
AI Interviewer	0.0078***	0.0074***	0.0074**	0.0061***	0.0058***	0.0058**	0.0037**	0.0038***	0.0038
	(0.0021)	(0.0021)	(0.0031)	(0.0020)	(0.0019)	(0.0029)	(0.0015)	(0.0015)	(0.0024)
Mean DV *Human Interviewer*	0.0455	0.0461	0.0461	0.0392	0.0397	0.0397	0.0215	0.0216	0.0216
Controls and fixed effects	–	Yes	Yes	–	Yes	Yes	–	Yes	Yes
Clustering	App.	App.	Rec.	App.	App.	Rec.	App.	App.	Rec.
Observations	53,660	52,367	52,367	53,660	52,367	52,367	53,660	52,367	52,367
R^2	0.0002	0.0885	0.0885	0.0002	0.0791	0.0791	0.0001	0.0639	0.0639

Notes: The table shows OLS estimates analyzing the treatment effect of receiving an AI voice agent instead of a human recruiter in an interview on several recruitment outcome variables. Controls include an applicant's gender, source of application, pre-treatment engagement score, and whether they have applied before to any of the firm's job postings. Fixed effects include week, recruiter, application side, and job posting fixed effects. An observation is an application. Standard errors in parentheses are either clustered at the applicant level ("App." in row "Clustering") or recruiter level ("Rec"). Significance levels: $^{*}\ p < 0.1$,$^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table B.6: Treatment effect on key recruiting outcomes among applicants who accepted an offer

Panel A:

	Dependent variable:					
	Started job			Employed after one month		
	(4)	(5)	(6)	(7)	(8)	(9)
AI Interviewer	0.0452***	0.0364**	0.0364*	0.0372**	0.0351**	0.0351*
	(0.0158)	(0.0156)	(0.0219)	(0.0167)	(0.0166)	(0.0194)
Mean DV *Human Interviewer*	0.6884	0.6962	0.6962	0.6060	0.6132	0.6132
Controls and fixed effects	–	Yes	Yes	–	Yes	Yes
Clustering	App.	App.	Rec.	App.	App.	Rec.
Observations	4,708	4,575	4,575	4,708	4,575	4,575
R^2	0.0018	0.0700	0.0700	0.0011	0.0611	0.0611

Panel B:

	Dependent variable:								
	Employed after two months			Employed after three months			Employed after four months		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
AI Interviewer	0.0351**	0.0330*	0.0330	0.0239	0.0265	0.0265	0.0177	0.0222	0.0222
	(0.0171)	(0.0170)	(0.0210)	(0.0172)	(0.0171)	(0.0202)	(0.0152)	(0.0137)	(0.0190)
Mean DV *Human Interviewer*	0.5562	0.5632	0.5632	0.4792	0.4849	0.4849	0.2609	0.2623	0.2623
Controls and fixed effects	–	Yes	Yes	–	Yes	Yes	–	Yes	Yes
Clustering	App.	App.	Rec.	App.	App.	Rec.	App.	App.	Rec.
Observations	4,708	4,575	4,575	4,708	4,575	4,575	4,708	4,575	4,575
R^2	0.0009	0.0523	0.0523	0.0004	0.0595	0.0595	0.0003	0.2559	0.2559

Notes: The table shows OLS estimates analyzing the treatment effect of receiving an AI voice agent instead of a human recruiter in an interview on several recruitment outcome variables. The sample is restricted to those applicants who accepted a job offer. Controls include an applicant's gender, source of application, pre-treatment engagement score, and whether they have applied before to any of the firm's job postings. Fixed effects include week, recruiter, application side, and job posting fixed effects. An observation is an application. Standard errors in parentheses are either clustered at the applicant level ("App." in row "Clustering") or recruiter level ("Rec"). Significance levels: $^{*}\ p < 0.1$,$^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table B.7: Predicting job offer decisions with interview variables

	Dependent variable: Received job offer (within *Human Interviewer* condition)					
	(1)	(2)	(3)	(4)	(5)	(6)
Interview score by recruiter	0.249***	0.193***	0.147***			
	(0.008)	(0.016)	(0.018)			
Sentiment of interview justification by recruiter				0.194***	0.157***	0.095***
				(0.009)	(0.014)	(0.015)
Test completed	–	Yes	Yes	–	Yes	Yes
Test scores included	–	–	Yes	–	–	Yes
Controls and fixed effects	–	–	Yes	–	–	Yes
Observations	4,661	2,477	2,477	3,430	1,880	1,880
R^2	0.179	0.049	0.237	0.139	0.059	0.248

Notes: The table shows OLS estimates analyzing the predictive power of interview variables on job offer decisions in the *Human Interviewer* condition. The dependent variable is an indicator equal to one if an application leads to a job offer, and zero otherwise. "Interview score by recruiter" is the 1,2,3 score that recruiters assign to applicants' interview performance, with higher values indicating higher performance. "Sentiment of interview text assessment by recruiter" is a 1, 0, -1 coded variable indicating whether the sentiment of the text with which recruiters describe applicants' interview performance is positive, neutral or negative, respectively. In columns (2) and (4), we additionally include control variables, fixed effects, and applicants' test scores in the standardized language and analytical test. Controls include an applicant's gender, source of application, pre-treatment engagement score, and whether they have applied before to any of the firm's job postings. Fixed effects include week, recruiter, application side, and job posting fixed effects. An observation is an application. Standard errors in parentheses are clustered at the applicant level. Significance levels: $^{*}\ p < 0.1$,$^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table B.8: Treatment differences in the distance between average and actual recruiter question

Topic	% of interviews covered (1)	*AI Interviewer* distance score (2)	*Human Interviewer* distance score (3)	H_0: (2) = (3) p-value (4)
Introduction	99.5%	0.103	0.217	$p < 0.001$
Source verification	2.1%	0.486	0.364	$p < 0.001$
Location and commute	65.7%	0.166	0.288	$p < 0.001$
Motivation & attrition risks – Motivation and goals	58.4%	0.091	0.332	$p < 0.001$
Motivation & attrition risks – Shifting schedule	56.9%	0.046	0.189	$p < 0.001$
Motivation & attrition risks – Age verification	55.8%	0.041	0.330	$p < 0.001$
Motivation & attrition risks – Training	51.0%	0.027	0.248	$p < 0.001$
Motivation & attrition risks – Planned vacation	52.2%	0.031	0.101	$p < 0.001$
Motivation & attrition risks – Other	55.4%	0.118	0.341	$p < 0.001$
Education verification	53.8%	0.089	0.187	$p < 0.001$
Compensation expectations	53.2%	0.035	0.208	$p < 0.001$
Employment history	54.3%	0.141	0.272	$p < 0.001$
Re-hire eligibility check	64.2%	0.041	0.270	$p < 0.001$
Availability	52.2%	0.099	0.276	$p < 0.001$
Data verification	51.7%	0.095	0.254	$p < 0.001$
Needs assessment	0.50%	0.462	0.360	$p < 0.05$
Profiling	8.3%	0.283	0.445	$p < 0.001$
Further procedure	49.4%	0.064	0.326	$p < 0.001$
Wrap-up & referrals – Referral	0.3%	0.412	0.341	$p > 0.10$
Wrap-up & referrals – Closing	65.6%	0.191	0.301	$p < 0.001$

Notes: P-values are obtained from pairwise t-tests and adjusted for multiple hypothesis testing using the Bonferroni correction.

Table B.9: Predicting applicants' interviewer choices

	Dependent variable:	
	Choice of Interviewer (AI = 1)	
	(1)	(2)
Perceived impact of AI on applicant (direction)	0.058	0.108**
	(0.040)	(0.043)
Controls and fixed effects	–	Yes
Observations	186	177
R^2	0.011	0.241

Notes: The table shows OLS estimates predicting applicants' interviewer choices in the *Choice of interviewer* treatment using their survey responses. The outcome variable is an indicator variable equal to one if an applicant chose the AI interviewer and zero otherwise. Higher values of "Perceived impact of AI on applicant" indicate a more positive impact of AI on applicants themselves. Controls include an applicant's gender, source of application, and pre-treatment engagement score. Fixed effects include week, recruiter, application side, and job posting fixed effects. An observation is an applicant. Standard errors in parentheses are clustered at the applicant level. Significance levels: * $p < 0.1$,** $p < 0.05$, *** $p < 0.01$.

Table B.10: Treatment differences on transcript data

	Dependent variable:	
	Interview is comprehensive	
	(1)	(2)
Direct AI Interview	0.0299***	0.0482***
	(0.0116)	(0.0099)
Mean DV in Human Interviewer	0.3879	0.3880
Controls and fixed effects	–	Yes
Observations	29,221	28,785
R^2	0.0002	0.2086

Notes: The table shows OLS estimates analyzing treatment differences between receiving an AI voice agent instead of a human recruiter in an interview on the type of interview. The dependent variable is an indicator variable equal to one if the interview is classified as *Comprehensive interview* and zero otherwise. *Comprehensive interview* means it opens and closes organically and covers at least eight canonical topics. Controls include an applicant's gender, source of application, pre-treatment engagement score, and whether they have applied before to any of the firm's job postings. Fixed effects include week, recruiter, application side, and job posting fixed effects. An observation is an application. Standard errors in parentheses are clustered at the applicant level. Significance levels: * $p < 0.1$,** $p < 0.05$, *** $p < 0.01$.

Table B.11: Differences in test scores depending on applicants' choice of interviewer

	Dependent variable:			
	Language test score (1-6)		Analytical test score (0-100)	
	(1)	(2)	(3)	(4)
Has chosen AI	-0.227^{***}	-0.170^{***}	-2.228^{***}	-1.665^{**}
	(0.040)	(0.044)	(0.689)	(0.775)
Mean DV when human is chosen	3.37	3.37	49.77	49.82
Controls and fixed effects	–	Yes	–	Yes
Observations	3,377	3,318	3,435	3,374
R^2	0.009	0.046	0.003	0.036

Notes: The table shows OLS estimates predicting applicants' test scores using their interviewer choice. "Has chosen AI " is an indicator variable equal to one if an applicant in the *Choice of Interviewer* chose the AI voice agent and zero if they chose the human interviewer. Controls include an applicant's gender, source of application, pre-treatment engagement score, and whether they have applied before to any of the firm's job postings. Fixed effects include week, recruiter, application side, and job posting fixed effects. An observation is an application. Standard errors in parentheses are clustered at the applicant level. Significance levels: $^{*}\ p < 0.1$,$^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table B.12: Differences in test scores when applicants choose instead of being assigned either the human or AI interviewer

Panel A: Choice of AI versus assigned AI interviewer

	Dependent variable:			
	Language test score (1-6)		Analytical test score (0-100)	
	(1)	(2)	(3)	(4)
Has chosen AI	-0.040^{*}	-0.005	-1.179^{***}	-0.686^{*}
	(0.022)	(0.022)	(0.374)	(0.376)
Mean DV in assigned AI	3.18	3.18	48.72	48.62
Controls and fixed effects	–	Yes	–	Yes
Observations	13,857	13,601	14,120	13,861
R^2	0.000	0.033	0.001	0.025

Panel B: Choice of human versus assigned human interviewer

	Dependent variable:			
	Language test score (1-6)		Analytical test score (0-100)	
	(1)	(2)	(3)	(4)
Has chosen human	0.099^{**}	0.036	0.598	-0.091
	(0.039)	(0.040)	(0.683)	(0.713)
Mean DV in assigned human	3.27	3.26	49.17	49.11
Controls and fixed effects	–	Yes	–	Yes
Observations	3,672	3,583	3,740	3,649
R^2	0.002	0.059	0.000	0.044

Notes: The table shows OLS estimates predicting applicants' test scores using their interviewer choice. In Panel A, "Has chosen AI " is an indicator variable equal to one if an application was in the *Choice of Interviewer* condition and the applicant chose the AI voice agent and zero if the application was in the *AI interviewer* condition instead, where applicants were interviewed by the AI voice agent without a choice. In Panel B, "Has chosen human " is an indicator variable equal to one if an application was in the *Choice of Interviewer* condition and the applicant chose the human interviewer and zero if the application was in the *Human interviewer* condition instead, where applicants were interviewed by a human recruiters without a choice. Controls include an applicant's gender, source of application, pre-treatment engagement score, and whether they have applied before to any of the firm's job postings. Fixed effects include week, recruiter, application side, and job posting fixed effects. An observation is an application. Standard errors in parentheses are clustered at the applicant level. Significance levels: * $p < 0.1$,** $p < 0.05$, *** $p < 0.01$.

Table B.13: Heterogeneity in job offer decisions by gender

	Dependent variable:		
	Received job offer		
	(1)	(2)	(3)
Female applicant	-0.0192^{***}	-0.0074	0.0037
	(0.0051)	(0.0047)	(0.0169)
Female applicant $\times$ *AI Interviewer*	0.0003	-0.0022	-0.0086
	(0.0059)	(0.0054)	(0.0187)
Mean DV among male applicants	0.0986	0.0987	0.4169
Controls and fixed effects	–	Yes	Yes
Test scores	–	–	Yes
Observations	53,515	52,367	13,866
R^2	0.0013	0.1683	0.2015

Notes: The table shows OLS estimates analyzing gender differences in job offer decisions across treatments. Controls include source of application, pre-treatment engagement score, and whether they have applied before to any of the firm's job postings. Fixed effects include week, recruiter, application side, and job posting fixed effects. An observation is an application. Standard errors in parentheses are either clustered at the applicant level. Significance levels: $^{*}p < 0.1$,$^{**}p < 0.05$, $^{***}p < 0.01$.

Table B.14: Heterogeneity in predicting job offer decisions of recruiters

	Dependent variable: Job Offer Made			
	Recruiters who consider interview $\leq$ test (survey)		Recruiters who consider interview $>$ test (survey)	
	(1)	(2)	(3)	(4)
Interview score (std.)	0.087***	0.117***	0.142***	0.120***
	(0.016)	(0.015)	(0.026)	(0.028)
Language test score (std.)	0.067***	0.077***	0.107***	0.114***
	(0.015)	(0.015)	(0.029)	(0.030)
Analytical test score (std.)	0.027*	0.012	0.026	0.016
	(0.016)	(0.015)	(0.029)	(0.029)
Interview score (std.) $\times$ *AI Interviewer*	−0.021	−0.024	−0.086***	−0.059*
	(0.018)	(0.017)	(0.029)	(0.031)
Language test score (std.) $\times$ *AI Interviewer*	0.042**	0.036**	0.045	0.038
	(0.017)	(0.016)	(0.031)	(0.032)
Analytical test score (std.) $\times$ *AI Interviewer*	−0.001	0.006	0.026	0.031
	(0.018)	(0.017)	(0.032)	(0.031)
Mean DV in Human Interviewer	0.32	0.32	0.42	0.41
Controls and fixed effects	–	Yes	–	Yes
Observations	4,574	4,538	2,147	2,091
R^2	0.101	0.243	0.152	0.242

Notes: The table shows OLS estimates predicting job offer decisions of recruiters using standardized test scores and interview scores. The dependent variable is an indicator variable equal to one if an application led to a job offer. Test scores are standardized. "*AI Interviewer*" is an indicator variable equal to one if the application was in the *AI Interviewer* condition, and zero if the application was in the *Human Interviewer* condition. We split the sample by recruiters who state in the recruiter survey that interview score is equally or less important than the test scores for offer decisions(columns (1)-(2)) and recruiters who state that interview scores are more important (columns (3)-(4)). Controls include an applicant's gender, source of application, pre-treatment engagement score, and whether they have applied before to any of the firm's job postings. Fixed effects include week, recruiter, application side, and job posting fixed effects. An observation is an application. Standard errors in parentheses are clustered at the applicant level. Significance levels: * $p < 0.1$,** $p < 0.05$, *** $p < 0.01$.

Table B.15: Predicting job offer decisions of recruiters full sample

	Dependent variable: Job Offer Made	
	(1)	(2)
Interview score (std.)	0.077^{***}	0.088^{***}
	(0.010)	(0.010)
Language test score (std.)	0.109^{***}	0.111^{***}
	(0.010)	(0.009)
Analytical test score (std.)	0.033^{***}	0.029^{***}
	(0.010)	(0.010)
Interview score (std.) $\times$ *AI Interviewer*	-0.038^{***}	-0.026^{**}
	(0.011)	(0.011)
Language test score (std.) $\times$ *AI Interviewer*	0.015	0.015
	(0.011)	(0.010)
Analytical test score (std.) $\times$ *AI Interviewer*	-0.003	0.001
	(0.011)	(0.011)
Mean DV in *Human Interviewer*	0.38	0.37
Controls and fixed effects	–	Yes
Observations	12,934	12,732
R^2	0.102	0.211

Notes: The table shows OLS estimates predicting job offer decisions of recruiters using standardized test scores and interview scores. The dependent variable is a dummy equal to one if an application led to a job offer. Test scores are standardized. "*AI Interviewer*" is a dummy equal to one if the application was in the *AI Interviewer* condition, and zero if the application was in the *Human Interviewer* condition. Controls include an applicant's gender, source of application, pre-treatment engagement score, and whether they have applied before to any of the firm's job postings. Fixed effects include week, recruiter, application side, and job posting fixed effects. An observation is an application. Standard errors in parentheses are clustered at the applicant level. Significance levels: $^{*}\ p < 0.1$,$^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table B.16: Categories of open-ended survey responses

Category	Definition	Example response
Comfortable and positive experience	Applicant mentions that they had a comfortable and positive experience.	"Yes, thank you, I just want to say I really appreciated how welcoming and professional the interview felt. The questions were thoughtful, and it gave me a clear picture of the role and firm culture. It made me even more interested in being part of the team. Thanks again for the opportunity!"
Short uninformative response	Applicant gave a short answer, expressing that they have nothing to add.	"Nothing else."
Problems with audio or questions	Applicant mentions a problematic interview experience due to misunderstanding of answers, frequent interruptions, or imperfect audio quality.	"The line is getting cut a lot and I haven't answered the question yet but the interviewer is already asking another question."
Clear audio and instructions	Applicant mentions that instructions and speaker audio were clear.	"I love how AI speaks clearly and how he/she ask about specific questions that need to be clear."
Excitement to join firm	Applicant mentions their interest and excitement in joining the firm.	"It's great and i hope I can join the team"
Problems with AI voice agent	Applicant mentions problems with the AI voice agent, such as a lack of understanding of applicant input, clarity or emotional cues.	"AI is a useful tool for interviews and makes the process easier. However, it's different from talking to a real person. Sometimes, when I reply, the AI doesn't fully understand me, which can lead to misunderstandings. That makes communication a bit challenging at times."
Feeling nervous	Applicant mentions that they were nervous for their first job interview.	"It's good, I didn't really give my best answers to the interview since I was nervous because its my first time interview but for first timer it was good."
Feeling unfairly judged	Applicant feels or expresses concern about being unfairly judged by the interviewer	"Aside of not having a bpo experiences, there were questions that I felt and think may impact my application. I just hope my application be considered nor not discriminate of how I speak in English"
Prefer human interviewer	Applicant prefers to be interviewed by a human interviewer.	"Sometimes, you would need to have human interaction on these types of interviews."
Discussed follow-up	Applicant discusses follow-up procedures.	"From the day i passed my final interview until now i didnt received any notification from you."
Other	Applicant responses that could not be classified into the other categories.	"What can you advice of the people got hired in BPO"

Table B.17: Timing of survey invitation balance tests

Variable	*Survey fielded*		H_0: (1) = (2)
	Post interview	Post recruitment	p-value
	(1)	(2)	(3)
Net promoter score	8.98	8.91	0.37
Perceived impact of AI index (direction)	0.28	0.26	0.77
Perceived impact of AI index (magnitude)	2.37	2.35	0.63
Perceived recruiter quality index	3.79	3.76	0.23
Perceived interview quality index	3.77	3.75	0.73
Knowledge of AI	2.48	2.35	0.10
Usage of AI	6.74	6.45	0.19
Observations	1,844	920	

Notes: Columns (1) and (2) display mean values of the survey variables depending on the time in the recruitment process when the survey was sent to applicants. Column (3) displays p-values obtained using a t-test.

Table B.18: Association of review experience with job offer rate among recruiters

	Dependent variable:	
	Average job offer rate (%)	
	(1)	(2)
Number of applications assigned to recruiter	-0.00006^{***}	-0.00013^{***}
	(0.00001)	(0.00004)
AI Interviewer		0.01862
		(0.02150)
Number of applications assigned to recruiter $\times$ *AI Interviewer*		0.00007^{**}
		(0.00003)
Recruiters	112	112
Observations	224	224
R^2	0.03181	0.04271

Notes: The table shows OLS estimates. The dependent variable is a recruiter's job offer rate (0-1), i.e., the sum of offers a recruiter gave to applicants during the experiment divided by the total sum of applications evaluated by the recruiter. "Number of applications assigned to recruiter" is the total number of applications a recruiter received per treatment as part of the experiment. In columns (1), we pool both *AI interviewer* and *Human interviewer* conditions. In column (2), we add an indicator variable indicating treatment status. An observation is a recruiter per treatment unit. Standard errors in parentheses are clustered at the recruiter level. Significance levels: * $p < 0.1$,** $p < 0.05$, *** $p < 0.01$.

C Job description wording

Title: Customer Expert

Our *Customer Service Representatives* and *Technical Support Representatives* are vital members of our company. You will field customer inquiries and find innovative ways to respond. You will have the chance to work in a highly collaborative and engaging environment that provides dynamic work experience with different cultures, as well as unlimited opportunities to grow your potential and develop your career.

As a *Customer Service Representative / Technical Support Representative*, your responsibilities will include:

- Handling and carefully responding to all customer inquiries via inbound calls and email
- Providing excellent customer service through active listening
- Working with confidential customer information in a secure manner
- Aiming to resolve issues on the first call by being proactive
- Appropriately and adequately communicating with customers

Working hours:
Monday to Friday – 8:00 PM to 5:00 AM PH Time / 10:00 PM to 7:00 AM PH Time

Background Requirements:

- NBI clearance
- Birth certificate
- Fit-to-work clearance

Compensation: Depends on job, between Php 16,000 and Php 25,000

Minimal Requirements:

- SHS Grad / HS Grad
- Average communication skills

D Detailed recruitment process

D.1 Engagement-score algorithm (non-proprietary summary)

Each applicant receives a base score that depends on how their profile entered the system (self-application vs. recruiter-added). The algorithm adds points for (i) valid phone and e-mail information, (ii) successful delivery of SMS/e-mail messages, (iii) message openings, and (iv) positive "Yes" responses; it subtracts points for explicit "No" responses. Missing or invalid contact details do not affect the score. Once the cumulative score exceeds a low, account-specific threshold, the applicant is queued for interview scheduling. Applicants who do not reach the threshold are not contacted. Applicants who are queued for interview scheduling are randomized into one of the three experimental treatments *Human interviewer*, *AI interviewer*, or *Choice of interviewer*.

D.2 Invitation text

D.2.1 Treatments *Human interviewer* and *AI interviewer*

Email: If a recruiter decides to interview an applicant, the following invitation text is sent via email to them. The text is identical across the *Human interviewer* and *AI interviewer* conditions.

[Subject line:] Interview Invitation: Schedule Your interview for the [Name Position]

Hi [First name] [Last name],

I hope you're doing well! My name is [Name recruiter], from [Name recruiting partner], the recruiting partner of [Name recruiting firm]. We've had a chance to review your application for the [Name Position]. We currently have an immediate need to fill this position. Not all roles require [Position-specific] experience, providing opportunities for various backgrounds and skill sets. If you are interested, please click on the button below.

[Button "Get Interview Call"]

[Clicking the button will redirect applicants to the firm's website, where they can schedule their interview.]

Phone: At the same time as the email is sent, the following invitation text is sent via text message to applicants' phones. Again, the text is identical across the *Human interviewer* and *AI interviewer* conditions.

Subject: Interview Invitation: Schedule Your Interview for the [Name position]

Hi [First name] [Last name],

We are reaching out to you regarding your application for the [Name position]. We'd like you to schedule your interview. We've sent the interview invitation to your email.

D.2.2 Treatment *Choice interviewer*

Applicants in the *Choice interviewer* treatment receive the same email and phone message as in the other two treatments. The only difference is that clicking the button "Get Interview Call" in the email will redirect them to an interview scheduling preference page. The text on the page is as follows:

Interview Scheduling Preference

Congratulations! You have been shortlisted for an interview. Please select your preferred interviewer:

- *AI Interviewer: The call can be scheduled at your convenience.*
- *Human Interviewer: You'll need to schedule the interview based on the human recruiter's availability.*

[After selecting the interviewer, applicants can schedule their interview exactly as in the other two treatments.]

D.3 Details on the interview process

D.3.1 How human recruiters conduct interviews

Human recruiters are provided with a structured interview script that ensures consistency while allowing flexibility in addressing individual candidate profiles. The interview begins with a standardized introduction in which the recruiter confirms the candidate's identity and explains the purpose of the call. The recruiter then proceeds with scenario-based questions tailored to the candidate's background. For example, if the candidate has gaps in their employment history, the recruiter asks about the reasons for those gaps and how the candidate maintained their motivation during that period. Similarly, if a candidate has frequently changed jobs, the recruiter asks about the reasons for these transitions.

In addition to these customized questions, the script includes general questions applicable to all candidates. These questions explore the candidate's recent employment, their strategies for handling stress, and their expectations regarding salary. For candidates with specific backgrounds, such as financial accounting experience or technical support, the

recruiter conducts mock calls or role-playing exercises to assess their practical skills in handling customer inquiries. For example, a candidate with a background in financial accounts might be asked to role-play a scenario where a customer inquires about a declined payment or a loan application process.

Although the script provides a comprehensive structure, recruiters are allowed to deviate from it as long as the essential questions are addressed. This flexibility allows recruiters to adapt their questioning to better suit the flow of the conversation and to probe deeper into areas of interest or concern. The interview process also includes a secondary round of questions, known as the "Validation Interview," where recruiters further assess the candidate's problem-solving abilities, teamwork experiences, and adaptability.

D.3.2 How AI voice agents conduct interviews

The AI voice agent is instructed to follow the same structure as human recruiters. At the beginning of the interview, the AI voice agent uses the following standardized text as the introduction:

AI voice agent*: Hi [Applicant name]. This is Anna, [firm name]'s AI recruiter and I am calling about the [job] role you applied to recently. Do you have a couple of minutes to chat about your application?*

Applicant: *[Example response: Hi, yes I have time.]*

AI voice agent: *Great! Since I am an AI recruiter, as I ask you questions, if you are not clear on my question, please feel free to ask me for clarification. Does that work for you?*

Applicant: *[Example response: Yes, that works for me.]*

AI voice agent: *Ok, I also want to let you know a human recruiter will review the recording from our discussion today and will make the final decision on your application for employment. The questions I will be asking you are the same questions my human counterpart would ask. Does that sound OK?*

E Interview transcript variables

Data preparation. The firm shared with us raw verbatim transcripts. In total, we have transcripts available for 34,109 applications, which is a subset of all interviews conducted. We used LLMs to transform them into structured and anonymized transcripts to be able to use the text as data. Specifically, we preprocessed the raw transcripts using prompt methods based on few shots, roles, and instructions with the *gemini-2.0-flash* model, following

the techniques described in Brown et al. (2020). First, we create speaker tags for each candidate and recruiter and then standardized this classification to each turn within a job interview. Second, we remove personal identifiable information (PII), such as name, address, phone number, and email address, from the speakers or the companies mentioned during the job interview. See Appendix Sections J.1 and J.2 for the prompts.

E.1 Interview type transcript variable

We performed an LLM-driven natural language processing analysis on anonymized structured transcripts using few-shot, role-based, and instruction-based prompt engineering techniques to categorize interviews into mutually exclusive types. We combine multiple prompting strategies to achieve a reliable and accurate transcript classification. Specifically, we combine role-based prompting, chain-of-thought (CoT) prompting, and in-context examples (zero- or few-shot). First, role-based prompting appears at the start of the prompt ("You are an interview expert"), which focuses the model on domain knowledge of recruitment and interviewing. Second, CoT prompting provides a structured decision hierarchy, reducing misclassifications when labels overlap (Wei et al., 2022; Vatsal and Dubey, 2024). Finally, we included in-context examples to help the model recognize category cues. The prompt can be found in Appendix Section J.3. In Table E.1, we display the classification of interview types, which was developed in collaboration with the company to achieve a comprehensive coverage of the different types of interviews.

Table E.1: Transcript type classification

Category	Definition	Duration	Topic Coverage
Comprehensive Interview	Interview has a natural opening and closure, a high-quality engagement, and contains $\geq$ eight expected topics.	Average	High (≥ 8)
AI Aversion	The candidate explicitly expresses unwillingness to continue speaking with an AI recruiter.	Short	Low
Early Screen-Out	Interview ends early because the candidate is immediately disqualified based on a non-negotiable requirement related to the job (e.g., location).	Short	Very Low (0–2)
Midway Screen-Out	Interview ends after some initial engagement due to a mismatch discovered during the conversation (e.g., conflicting school plans).	Medium	Moderate (3–7)
Late Screen-Out	Interview proceeds nearly to completion but the candidate fails a final, critical criterion (e.g., rehire status).	Long	High (8+)
Telephony Failure	Interview ends due to issues with cellular network, signal loss, or VOIP instability.	Varies	Varies
AI System Failure	LLM/voice agent stalls, crashes, or fails to respond in the interview.	Varies	Varies
Disengaged Interaction	Applicant is disinterested, unresponsive, and/or distracted during the interview, and the interview has poor continuity.	Varies	Low-Moderate (< 8)
Candidate Unavailability	Applicant states they are unable to talk, and the interview ends for this reason.	Short	Very Low (0–2)
Others	Interview type that does not fit in any other category.	Varies	Varies

E.2 Recruiter transcript variables

We use the following five variables to capture features of recruiters' language and behavior in the interviews. To construct the variables, we used LLM-based promoting with *gemini-2.5-flash*, a transformer model, and regex NLP methods.

Topic coverage. We use a controlled-vocabulary prompt to categorize how many of the maximum number of 14 topics that recruiters can cover in a given transcript are substantively covered. A topic is counted as 'covered' only when (i) the interviewer explicitly probes the theme, and (ii) the candidate offers a nontrivial reply (at least three content words) that contains at least one keyword from a topic-specific lexicon supplied to the model. To reduce semantic drift and ensure consistency across transcripts, we use the firm's predefined topic labels. We also add a final, "other" topic to capture off-topic conversations, hence there are 15 in total. Appendix Table B.2 displays the full list of possible topics, and Appendix Section J.4 the prompt. Topic coverage is then a percentage, defined as the ratio of topics covered divided by the maximum number of 15 topics.

Following topic order. This variable measures the consistency of the order of topics in a given interview transcript compared to the firm's guideline topic order. See Table B.2 for the template order and description of topics. For each transcript, we extract the sequence in which topics appear and remove repeated occurrences so that each topic is represented once. Next, we compute the Kendall rank correlation coefficient (τ) between the transcript's topic sequence and the guideline's topic sequence. Kendall's τ measures the degree of similarity in the ordering between the two sequences, ranging from -1 to 1. Higher τ values thus indicate stronger similarity between actual and guideline topic order.

Following guideline questions. To measure the extent to which recruiters follow the guideline questions, we calculate the semantic similarity between recruiter questions asked in each interview and the template questions supplied by the firm as guidelines. To do so, we first match each question to one of the 14 topics (Table B.2) and convert the text into a 384-dimension embedding using the all-MiniLM-L6-v2 sentence transformers model.[23] In the same way, we convert the template questions to word embeddings, averaging the embeddings in case of multiple template questions. We then computed the cosine similarity score between the recruiter's word embedding of the question and corresponding

[23]The model all-MiniLM-L6-v2 is a sentence embedding model that is fine-tuned to perform well on sentence similarity tasks. The model is obtained from https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

embedding of the template questions for that particular topic. For each interview, we then average the similarity scores across all topics. Higher values indicate higher overall similarity of recruiter questions with the guideline questions.

Vocabulary richness score. This variable is defined as the number of unique words divided by the square root of the total number of words spoken by the recruiter. Higher scores indicate greater vocabulary diversity and linguistic sophistication.

Question distance score. This variable measures the semantic variance of recruiter questions per topic. We again first match each question to one of the 14 topics (Table B.2) and convert the text into a 384-dimension embedding using the all-MiniLM-L6-v2 sentence transformers model. Then, separately for each topic and condition (AI voice agent and human recruiter), we measured the "topic centroid embedding" by calculating the average of word embeddings of all recruiter questions within that topic per condition. We then calculate the topic distance score as 1 - cosine similarity score between the word embedding of each question and the corresponding topic centroid embedding. For each interview, we then calculate the average distance score across all questions asked for the respective topic. A higher question distance score indicates greater semantic variation in the phrasing of the questions relative to the average recruiter question.

E.3 Applicant transcript variables

We use the following eight variables to capture linguistic features of applicants. The relevant phrases and markers were captured using spaCy and regex NLP methods.

Vocabulary richness score. This variable is defined as the number of unique words divided by the square root of the total number of words spoken by the applicants. Higher scores indicate greater vocabulary diversity and linguistic sophistication.

Syntactic complexity score. This variable captures the average number of subordinate clauses, specific details, and explanatory phrases per response used by the applicant, normalized by response length. Higher values capture greater thoroughness and nuance in responses.

Discourse markers frequency. This variable captures the average number of sequential, causal, and clarifying discourse markers per minute (e.g., "first", "because", "specifically") in the applicant's responses. Higher values indicate more frequent use of discourse markers.

Filler words frequency. This variable represents the average number of basic and conversational fillers per minute (e.g., "uh", "uhm", "like", "you know") in the applicant's responses. Higher values indicate more frequent use of filler words.

Backchannel cue frequency. This variable measures the average number of verbal backchannel cues per minute — short verbal cues supplied by the applicant to indicate attention or agreement (e.g., "sure", "got it", "mhm", "okay", "yeah", "yes"). Higher values indicate more frequent use of backchannel cues.

Number of exchanges interviewer–applicant. This variable counts the total number of conversational exchanges, where an exchange is defined as a sequence in which the interviewer speaks, followed by the applicant.

Number of questions by applicant. This is the total number of questions asked by the applicant during the conversation.

Linguistic style match index. This index captures the similarity of linguistic style between the applicant and interviewer. It is constructed as the average similarity score across nine function word categories: (1) personal pronouns (e.g., "I", "you"), (2) impersonal pronouns (e.g., "this", "it"), (3) articles (e.g., "a", "the"), (4) auxiliary verbs (e.g., "am", "have"), (5) high-frequency adverbs (e.g., "very", "well"), (6) prepositions (e.g., "in", "around"), (7) conjunctions (e.g., "but", "while"), (8) negations (e.g., "not", "no"), and (9) quantifiers (e.g., "many", "few"). Each category score is defined as

$$1 - \frac{|\text{rate}_{\text{interviewer}} - \text{rate}_{\text{applicant}}|}{\text{rate}_{\text{interviewer}} + \text{rate}_{\text{applicant}} + 0.0001},$$

where rates represent the percentage usage in each speaker's text. Higher index values indicate greater linguistic style similarity between the interviewer and applicant.

F Examples of full interviews

The following transcripts present two sample interviews, one from the *Human Interviewer* and one from the *AI Interviewer* condition. As noted in the main text, the content of interviews varies across application sites and job profiles. These examples are therefore provided for illustrative purposes and are not necessarily representative of the interviews conducted in the experiment. To protect applicant confidentiality, the transcripts have been edited to preserve anonymity and prevent any possibility of direct or indirect identification.

F.1 Human Interviewer example

Interviewer: Hello.
Candidate: Hello, good morning.
Interviewer: Hi, this is [Anonymized] from [Anonymized] Recruitment. I'll be conducting the interview and I'll be asking you personal details and some common interview questions. Then I'll be giving you feedback prior to your position, okay? [Topic 1: Introduction]
Candidate: Yes, of course.
Interviewer: Now, I just want to remind you that this phone call is recorded for quality assurance purposes.
Candidate: Yes.
Interviewer: Okay. So can you state your complete name, your middle name, your age, and your birthplace?
Candidate: My name is [Anonymized] and I'm [Anonymized].
Interviewer: I'm sorry, how old are you?
Candidate: [Anonymized].
Interviewer: [Anonymized]?
Candidate: Yeah, correct.
Interviewer: Can you turn down on your phone maybe?
Candidate: Yes.
Interviewer: Have you ever worked for [Anonymized] in the past?
Candidate: No. This is my second job in [Anonymized].
Interviewer: And where are you currently residing? [Topic 2: Location Verification]
Candidate: In [Anonymized].
Interviewer: If you are ever hired, will there be any issues commuting or traveling from your place that you work?

Candidate: No, there's no issues at all because right now we do have a direct train from [Anonymized] going to your place.
Interviewer: May I know what will be your mode of transportation or how long will it take you to travel? [Topic 2: Mode of Transportation]
Candidate: Yeah, just a few. I think it will be around an hour.
Interviewer: Can you state your highest educational attainment? [Topic 3: Educational Background]
Candidate: I did [Anonymized].
Interviewer: Hello, I'm sorry. I'm sorry I can't hear you clearly. Can you turn the muffled sound if you can? Can you transfer for a better reception or maybe using your headset?
Candidate: Yeah, I'm currently using my headset right now.
Interviewer: If you can, can we just mute your phone so that you can hear me clearly?
Candidate: Okay, ill just remove my headset. Sorry. Hello?
Interviewer: Hello, can you hear me clearly?
Candidate: Okay, that's better. Thank you so much. First, I graduated [Anonymized]. Then I took up my masters degree in [Anonymized] but unfortunately I just finished [Anonymized] because there were some circumstances that I have to face here in the Philippines. Thats why I go back there.
Interviewer: Okay. May I know what's your masters degree?
Candidate: Masters degree in [Anonymized], majoring in [Anonymized].
Interviewer: May I ask you if you have plans to pursue your studies? [Topic 4: Candidate's Commitment and Attrition Risks - Future Study Plan]
Candidate: Right now, no because I need to focus first with [Anonymized]. Thats the reason why I came back here in the Philippines.
Interviewer: Thank you. Are you currently employed?
Candidate: Sorry?
Interviewer: Are you currently employed?
Candidate: No. I'm not employed right now.
Interviewer: Do you have any pending applications from the other companies? [Topic 4: Candidate's Commitment and Attrition Risks - Other Applications]
Candidate: To be honest, I just submitted my application and fortunately this is the first company who called me for an interview. Thats why I'm here right now.
Interviewer: Thank you. How about any previous work experience? All in all? Any foreign? [Topic 5: Past Employment History]
Candidate: First, my first three years I worked in [Anonymized] as [Anonymized]. Then after that, I went abroad to [Anonymized]. I worked as a [Anonymized] for [Anonymized]

years. Then I go back here in the Philippines to work as [Anonymized] for [Anonymized]. Then I went to [Anonymized] and worked as an [Anonymized].
Interviewer: So the most recent job that you had is in [Anonymized], right?
Candidate: Yeah, correct.
Interviewer: Can you please tell me more about it?
Candidate: As a [Anonymized], I'm the one responsible for [Anonymized]. All the problems related with [Anonymized], they put in [Anonymized]. Then every morning I have to check [Anonymized] and check all the priorities which need to handle first. Also, I have to give feedback to [Anonymized]. And I also receive some [Anonymized] from the customers regarding the problems or issues with their equipment or any computer related problems. And especially, I need to resolve all those issues to reduce the downtime on their side. Basically, that's all my job.
Interviewer: And may I know what's the reason why you left that job?
Candidate: I left my job. Thats what I'm saying because I have to face some problems here in the Philippines because [Anonymized], they have some issues. I have to weigh in which is better for the future. But I will not sacrifice my [Anonymized]. So I'd rather choose [Anonymized] rather than [Anonymized].
Interviewer: May I ask if you have any certificates of employment from the previous job?
Candidate: Yeah, I do.
Interviewer: So what made you decide to apply?
Candidate: Yeah, because I submitted applications and fortunately this is the first company who called me for an interview. Also, I think since I'm challenging, like trying something like [Anonymized] which is very in demand in this country. And I also want to improve my comm skills as well because working in [Anonymized] will improve your comm skills.
Interviewer: May I know what's your salary expectation? [Topic 6: Compensation]
Candidate: To be honest, I would like to. I dont have a specific since I dont have any background experience in [Anonymized]. So I would like to assume that from the range between [Anonymized] to [Anonymized]
Interviewer: [Anonymized] to [Anonymized], right?
Candidate: Yes, yes.
Interviewer: The range that we can offer you if ever you're going to get hired will be from [Anonymized] to [Anonymized] depending on what account are you going to be with. Are you okay with it?
Candidate: Yes.
Interviewer: Okay. So I'm sure you have [Anonymized] experience.

Candidate: Yeah, [Anonymized]. The customers will call [Anonymized] to help to resolve their problem or some issues. For example, for instance, they have an internet issue and the client or the customer will call [Anonymized] and the job of [Anonymized] is to resolve, to find the root cause and to resolve the issues of [Anonymized].
Interviewer: And how would you define quality [Anonymized]?
Candidate: Oh yeah. First, you must be polite to the service and you must also handle the average customer and you should be able to resolve all the [Anonymized] issues and concerns. Especially with less supervision from [Anonymized]. You must be able to handle that on your own.
Interviewer: So if ever youre going to encounter irate or difficult [Anonymized], how do you think you can handle them effectively?
Candidate: Yeah, I always stay calm. I already encountered with irate [Anonymized] in [Anonymized]. I stay calm and explain to them everything in [Anonymized]: I explain everything clearly and I try to focus on the problem and solution to resolve the issues of [Anonymized].
Interviewer: What are skills do you think you have that can be a great asset for the job position?
Candidate: First, I'm a hardworking and dynamic self-disciplined person and I can also work as an individual or a team player and I'm good at handling [Anonymized].
Interviewer: If ever you're going to encounter conflict between your peers or your colleagues, how are you going to address them?
Candidate: I will have to help them personally and privately, not in front of the audience, and help them and try to resolve the issues.
Interviewer: Do you have any [Anonymized] or at least most of the time I mentioned?
Candidate: No. [Anonymized]
Interviewer: Do you have any other certifications aside from your work?
Candidate: No, not at all.
Interviewer: Do you have any travel plans within the next six months, international or local? [Topic 7: Candidate's Commitment and Attrition Risks - Planned Vacation]
Candidate: No, I don't have one.
Interviewer: Have you ever had the willingness to work on-site?
Candidate: Yeah, of course.
Interviewer: You can work in shifting schedules, in extended hours, holidays, weekends, weekdays? [Topic 7: Candidate's Commitment and Attrition Risks - Working during Holidays]
Candidate: Yeah, no problem at all. Because when I worked in [Anonymized] for [Anonymized],

I worked in various shifts for at least [Anonymized].
Candidate: So working in various shifts is not a problem at all.
Interviewer: Can you start as soon as possible if ever youre going to get hired?
Candidate: Yeah, absolutely. And I'm willing to start.
Interviewer: And may I know how long are you planning to work here with us? [Topic 7: Candidate's Commitment and Attrition Risks - Long-term Intention]
Candidate: Sorry?
Interviewer: How long are you planning to work here with us?
Candidate: As long as I'm enjoying my job in our company, Ill pursue my goal. Because if I'm going to start here as a regular [Anonymized], thats not my goal.
Candidate: My goal is to become a higher position, which I can use my masters degree in handling people, something like that.
Interviewer: Many of you already have your government documents like [Anonymized] to help you by any chance? [Topic 8: Data Verification]
Candidate: Right now I dont have [Anonymized] yet, but I do have [Anonymized].
Interviewer: And may I know what valid IDs do you have with you?
Candidate: I do have my [Anonymized].
Interviewer: And are you vaccinated?
Candidate: [Anonymized]
Interviewer: Do you have any medical issues that we need to know?
Candidate: [Anonymized].
Interviewer: Okay, I think that concludes the end of our interview. The next step is the assessment. Please wait for the contact officer to call you, okay? [Topic 9: Conclusion]
Candidate: Okay, thank you so much.

F.2 AI Interviewer example

Interviewer: Hi [Anonymized]. This is Anna, [Anonymized] AI recruiter and I am calling about the [Anonymized] role you applied to recently. Do you have fifteen minutes to talk about your application? [Topic 1: Introduction + Candidate Availability]
Candidate: Yes.
Interviewer: Great! Since I am an AI recruiter, as I ask you questions, if you are not clear on my question, please feel free to ask me for clarification. Does that work for you?
Candidate: Yes.
Interviewer: Ok, I also want to let you know a human recruiter will review the recording from our discussion today and will make the final decision on your application for

employment. The questions I will be asking you are the same questions my human counterpart would ask. Does that sound OK?
Candidate: Yes.
Interviewer: Let's get started. Have you ever worked for [Anonymized] in the past?
Candidate: No. This is my first time working in, um, [Anonymized].
Interviewer: That's okay, we get a lot of new applicants. Can you tell me a little bit about yourself and why you're interested in this role?
Candidate: Firstly, good afternoon. I'm [Anonymized] from [Anonymized] and I am an [Anonymized] graduate and have also completed [Anonymized] semester of [Anonymized] at [Anonymized] in [Anonymized]. My year of experience in [Anonymized] has equipped me with strong customer service and skills, and I'm confident in my ability to make a significant contribution to your team.
Interviewer: That's great to hear about your experience and education. This position is an on-site position and is located at [Anonymized]. Can you share with me where you reside? [Topic 2: Location Verification]
Candidate: I'm residing in [Anonymized]. Is there any available position in [Anonymized]?
Interviewer: Our office is actually located in [Anonymized]. How far a drive would this be for you?
Candidate: I think, um, it's one and a half hour from [Anonymized]. But it would be better if there's any open position in [Anonymized] .
Interviewer: Ok, how about transportation to and from work. Can you share with me the mode of transportation you would use to commute to our office daily? [Topic 2: Mode of Transportation]
Candidate: Oh, from [Anonymized], I'd estimate my commute to [Anonymized] would be around [Anonymized] minutes to [Anonymized] depending on traffic. But to ensure punctuality, uh, also plan to commute at least [Anonymized] before my shift, starts since it's a one ride trip only, and I'm confident about reaching the time reaching the office on time.
Interviewer: The minimum age requirement for this position is 18 years of age, are you at least 18 years of age? [Topic 3: Age Requirement]
Candidate: I'm [Anonymized] years old.
Interviewer: Most of the opportunities we have are on a shifting schedule where you could work during holidays and/or weekends. Have you had previous employment that has required you to work a similar schedule? [Topic 4: Candidate's Commitment and Attrition Risks - Working during Holidays]
Candidate: Yes.

Interviewer: In addition to a shifting schedule, we sometimes require overtime. Would you be able to accommodate extended hours when needed? [Topic 4: Candidate's Commitment and Attrition Risks - Working Overtime]
Candidate: Yes. I will.
Interviewer: I'd like to ask you some questions about your career goals and motivations. When you read our job advertisement, what attracted you to the position at [Anonymized] [Topic 4: Candidate's Commitment and Attrition Risks - Motivations]
Candidate: Actually, my previous coworker referred me to because as per him, there's open roles for a [Anonymized] account and as I have [Anonymized] experience and I think I will be fit for this role. That's why I applied for this job.
Interviewer: That's interesting. As you consider your career goals, how long could you envision working for [Anonymized] and performing the duties of this job? [Topic 4: Candidate's Commitment and Attrition Risks - Long-term Intentions]
Candidate: I think I see myself um, still working to your company [Anonymized] to [Anonymized] years. If there's a good position or good environment, I think I I'm able to work as long as I work on my previous job.
Interviewer: As you consider making a job change, what is important to you as you evaluate job opportunities?
Candidate: Right now, my goal is to work at your company since I as I check on my multimedia reviews on [Anonymized] that your company's team good. That that's why I want to focus on working with company as well.
Interviewer: Thank you for your answers, I appreciate the insight.
Candidate: That's why I think I didn't need to find another job. I'm just interested in the job position that you have open.
Interviewer: I'd like to shift back to a few more questions about the position we are considering you for. Our training requires mandatory attendance. If selected for the position, would you be able to attend all training?
Candidate: Yes. I will. Mm-mm.
Interviewer: Do you have any planned vacations over the next 6 months? [Topic 4: Candidate's Commitment and Attrition Risks - Planned Vacations]
Candidate: No. I don't have any plans. My, um, go for this year or this month is work. Focus on work.
Interviewer: Are there any activities that may affect your attendance or any prior commitments that you have to attend to?
Candidate: For now, there aren't any.
Interviewer: Let's take a few minutes to review your education background. What is your

highest level of education? [Topic 5: Educational Background]
Candidate: I am a [Anonymized] graduate and have also completed two semester of [Anonymized] at [Anonymized].
Interviewer: That's interesting. Are there any activities that may affect your attendance or any prior commitments that you have to attend to?
Candidate: For now, I think there are no activities in the future that affect my attendance because I will now talk as to work, and I have no any vacation pending.
Interviewer: Let's take a few minutes to review your education background. You've already told me about your highest level of education completed, which is [Anonymized] and having completed [Anonymized] at [Anonymized]. I just want to confirm, is that correct?
Candidate: Yes.
Candidate: I'm an [Anonymized] and a [Anonymized], um, completer, uh, for [Anonymized]. I think it's correct.
Interviewer: Since you have completed [Anonymized], do you plan to return to [Anonymized]?
Candidate: For now, I don't think so. I just my goal for now is to get a job. Permanent job, to sustain my bill needs.
Interviewer: If we can shift our discussion to compensation expectations, that would be helpful. What is your expected salary? [Topic 6: Compensation]
Candidate: I consider my ex considering my experience in customer facing enroll, I was hoping for a salary closer to [Anonymized] to [Anonymized].
Interviewer: The roles that we have open right now are between the salary package range of [Anonymized] to [Anonymized] per Month, depending on your qualifications, you could be offered a salary package between these amounts. Would you be okay with this range?
Candidate: Is there any closer for [Anonymized] to [Anonymized]? May I know the specific or may I know the account first?
Interviewer: The roles that we have open right now are between the salary package range of [Anonymized] to [Anonymized], depending on your qualifications, you could be offered a salary package between these amounts. Would you be okay with this range?
Candidate: Yes. I would be okay.
Interviewer: Thank you again for the information you've shared. I have a few more questions to go over, so let's continue. I'd like to talk about your [Anonymized] experience. can you share with me the volume of [Anonymized] on a daily basis? [Topic 7: Past Employment History]
Candidate: My previous job, I have been in [Anonymized] near [Anonymized] and I re-

ceived [Anonymized] well, we received [Anonymized] about [Anonymized]. That's our goal for the day, and I get [Anonymized] that's passing the [Anonymized] for a day.
Interviewer: That sounds like a significant vo-
Candidate: yes. Continue.
Interviewer: Were these [Anonymized] solving technical issues or were you responding to [Anonymized]?
Candidate: We conduct troubleshooting and also we create [Anonymized], and we talk to a customer and explain their needs, what their needs, and also like that.
Interviewer: Now, I'd like to ask about your previous work experience. Have you worked with other companies before this role?
Candidate: My previous job, I ended my journey there [Anonymized].I think. Yeah. That's my last job.
Interviewer: So, it sounds like you have some work experience, specifically from your last job.
Candidate: Yes.
Interviewer: Now, could you please provide the name of the most recent company you worked for?
Candidate: I've been for [Anonymized] for a year, and my first job was [Anonymized]. Then the second one, is I've I have [Anonymized].
Interviewer: Now, can you tell me what are your employment dates for those companies you mentioned.
Candidate: My employment date like, probably this job I've been there for [Anonymized] and ended [Anonymized]. It's almost [Anonymized]
Interviewer: So, your employment dates are from [Anonymized] to [Anonymized]. That's helpful to know.
Candidate: Mhmm. That's correct.
Interviewer: Now. Can you tell methe location of the company we are discussing. Where is this company located?
Candidate: This company must my previous job was located at [Anonymized].
Interviewer: So your previous job was located at [Anonymized]. Now, I'd like to ask, do you have any subsequent work experience with other companies?
Candidate: I'm also I also work as health care account at [Anonymized]. Before I apply for at [Anonymized], then I applied for this.
Interviewer: That's interesting. Now, could you please provide the name of the second most recent company you worked for?
Candidate: Second reason, company I work for. I work for [Anonymized]. I'm [Anonymized]

there. I've been there almost [Anonymized].
Interviewer: I'd like to know more about your work history. What are your employment dates for these previous roles, if you have them available?
Candidate: For a while, let me double check here because I don't I don't remember. I think it's to be sure, let's me let me double check here on my resume for a while.
Interviewer: Where is this company located?
Candidate: Okay. Before I applying for a [Anonymized] as a [Anonymized] for [Anonymized] account. Before that, I've been in a [Anonymized] as a [Anonymized]. Duration was [Anonymized]. It's located at [Anonymized].
Interviewer: So, it seems like you have some experience in customer service. Do you have any subsequent work experience with other companies besides [Anonymized] and [Anonymized]?
Candidate: I also work at [Anonymized]. I've been there for duration is [Anonymized] to [Anonymized] as a [Anonymized]. All my job all my previous job experience is [Anonymized] roles.
Interviewer: Could you please provide the name of the third most recent company you worked for?
Candidate: My third is the [Anonymized]. It's a [Anonymized], and my position there a [Anonymized]. The duration is [Anonymized] until [Anonymized]. This company was located at [Anonymized].
Interviewer: You mentioned you worked at [Anonymized] as a [Anonymized], can you tell me why you left that company?
Candidate: I left at [Anonymized] because I'm sick.
Interviewer: If you are offered the job, when would you be able to start? [Topic 8: Availability]
Candidate: I can for now because since I don't have any work, I can work
Interviewer: Do you have any pending applications this this week in [Anonymized]
Candidate: This is my first application before I ended my contract on my previous job.
Interviewer: Do you have any pending applications with other companies?
Candidate: I don't have any pending applications.
Interviewer: What is your contact number? [Topic 9: Data Verification]
Candidate: Okay. My contact number and my cell phone number [Anonymized]
Interviewer: Could y
ou please share your [Anonymized] number with me?
Candidate: Okay. Same as I said earlier, my phone number through [Anonymized] is [Anonymized] **Interviewer:** May I know your Marital/Civil Status, please?

Candidate: [Anonymized].
Interviewer: [Anonymized] do you have any questions you would like for me to document and have reviewed by the recruiting team?
Candidate: Yes. I want to ask if there's any available position for your local account in tour site at [Anonymized] because I was prepared my by my previous by my previous coworker and I really want to join their team because I I believe, like, I have a role because or I fitted the job this job because I have a [Anonymized]. As per my referral, your current employee, that's health care account for your local account or bill care account is now open for hiring.
Interviewer: I have noted your question, and I will make sure to document it, after this call, someone from our team will get back to you regarding the available position for [Anonymized], and we will also verify the information about [Anonymized] being open for hiring.
Candidate: Okay. Thank you. That's all.
Interviewer: Thank you [Anonymized]. This concludes the questions I have for you today. Following this interview, our recruiting team will be reviewing the summary of our conversation to determine next steps regarding your application for employment. Due to the high demand and urgency to fill our openings, we encourage you to complete an online assessment. By taking this assessment right away, you'll demonstrate your commitment to this opportunity. [Topic 10: Conclusion + Assessment Test]
Candidate: Mhmm.
Interviewer: Great! We'll see you on the next hiring process. Good luck!

G Applicant survey instructions

Invitation text short survey. We want to hear about your recent candidate experience. Please take a few minutes to share your feedback. Your feedback is confidential and does not impact any employment decisions.

Invitation text long survey. We want to hear about your recent candidate experience. Please take a few minutes to share your feedback. As an appreciation for your time, we will send you a gift card in the amount of 4USD when you complete this survey. Your feedback is confidential and does not impact any employment decisions.

G.1 Long survey wording

G.1.1 Procedural trust

1. Based on your experience, how likely is it that you would recommend our company to a friend or colleague as a place to apply for work? [NPS question]

2. Was the recruiter knowledgeable about the company?

    - Very knowledgeable
    - Somewhat knowledgeable
    - Slightly knowledgeable
    - Not knowledgeable at all

3. Was the recruiter knowledgeable about the role you were applying for?

    - Very knowledgeable
    - Somewhat knowledgeable
    - Slightly knowledgeable
    - Not knowledgeable at all

4. Were the questions asked during your phone interview relevant to the job you applied for?

    - Very relevant
    - Somewhat relevant
    - Slightly relevant
    - Not relevant at all

5. Do you feel your time was valued during the recruitment process?

    - Very much valued
    - Somewhat valued
    - Slightly valued
    - Not valued at all

6. Did you feel that the recruiter was able to follow up appropriately based on your answers?

- Always
- Often
- Sometimes
- Rarely
- Never

G.1.2 Social experience

1. How natural did the interaction with the recruiter feel?
    - Very natural
    - Somewhat natural
    - Neutral
    - Somewhat unnatural
    - Very unnatural
2. How comfortable did you feel during the interview with the recruiter?
    - Very comfortable
    - Somewhat comfortable
    - Neutral
    - Somewhat uncomfortable
    - Very uncomfortable
3. Did you find talking to the recruiter stressful?
    - Not at all stressful
    - Somewhat stressful
    - Moderately stressful
    - Very stressful
    - Extremely stressful
4. How frequently did you receive live feedback from the recruiter during your interview?
    - Very frequently

- Somewhat frequently
- Occasionally
- Rarely
- Never

G.1.3 Perceived Discrimination

1. Did you feel discriminated by the recruiter because of your gender identity?
    - Yes
    - No
    - Not sure
2. Do you believe the interview process was fair compared to your past interview experiences?
    - Much more fair
    - More fair
    - About the same
    - Less fair
    - Much less fair
    - This was my first ever job interview. (N/A).

G.1.4 General AI awareness, knowledge and usage

1. Select the correct definition of Artificial Intelligence (AI)
    - AI is the process of enhancing industrial machinery efficiency using automated control systems for mechanical and electrical improvements. [Incorrect]
    - AI involves developing computer systems to perform tasks that usually require human intelligence, such as language understanding and pattern recognition. [Correct]
    - AI is a method to develop software applications specifically designed to manage financial transactions and banking operations efficiently. [Incorrect]
    - AI refers to the creation of complex spreadsheets for data analysis and business forecasting, emphasizing numerical computations. [Incorrect]

- I don't know. [Incorrect]

2. Select the correct definition of Generative Artificial Intelligence (GenAI)

    - GenAI automates genetic analysis to modify DNA sequences for medical purposes. [Incorrect]
    - GenAI is AI that creates new content, such as text and images, by learning from existing data. [Correct]
    - GenAI involves the creation of algorithms for solving complex mathematical problems and optimizing industrial processes. [Incorrect]
    - GenAI is a system that focuses on generating engineering techniques to enhance crop yield and resistance to pests in agriculture. [Incorrect]
    - I don't know. [Incorrect]

3. How often do you use the following products?

    - ChatGPT [Daily, Weekly, Monthly, Never, I have never heard of it, Not rated N/A]
    - Character.AI [Daily, Weekly, Monthly, Never, I have never heard of it, Not rated N/A]
    - QuillBot [Daily, Weekly, Monthly, Never, I have never heard of it, Not rated N/A]
    - Midjourney [Daily, Weekly, Monthly, Never, I have never heard of it, Not rated N/A]

4. Thinking about customer service, which of the following uses artificial intelligence (AI)?

    - A detailed Frequently Asked Questions webpage [Incorrect]
    - An online survey sent to customers that allows them to provide feedback [Incorrect]
    - A contact page with a form available to customers to provide feedback[Incorrect]
    - A chatbot that immediately answers customer questions [Correct]
    - Not sure [Incorrect]

5. When using email, which of the following uses artificial intelligence (AI)?

- The email service marking an email as read after the user opens it [Incorrect]
- The email service allowing the user to schedule an email to send at a specific time in the future [Incorrect]
- The email service categorizing an email as spam [Correct]
- The email service sorting emails by time and date [Incorrect]
- Not sure [Incorrect]

6. Thinking about online shopping, which of the following uses artificial intelligence (AI)?

    - Storage of account information, such as shipping addresses [Incorrect]
    - Records of previous purchases [Incorrect]
    - Product recommendations based on previous purchases [Correct]
    - Product reviews from other customers [Incorrect]
    - Not sure [Incorrect]

G.1.5 AI perception on the labor market

[The following items were taken from *American Trends Panel Wave 119* (2022).]

Over the next 20 years, how much impact do you think the use of artificial intelligence (AI) in the workplace will have on...

1. Workers generally

    - A major impact
    - A minor impact
    - No impact
    - Not sure
    - No answer

2. You, personally

    - A major impact
    - A minor impact
    - No impact

- Not sure
- No answer

Thinking about the use of artificial intelligence (AI) in the workplace over the next 20 years, what do you think the outcome will be for...

3. Workers generally
    - AI will help more than it hurts
    - AI will equally help and hurt
    - AI will hurt more than it helps
    - Not sure
    - No answer
4. You, personally
    - AI will help more than it hurts
    - AI will equally help and hurt
    - AI will hurt more than it helps
    - Not sure
    - No answer

G.1.6 Open-ended feedback

1. Do you have any additional feedback you'd like to share about your interview experience?

G.2 Short survey wording

G.2.1 Procedural trust

1. Based on your experience, how likely is it that you would recommend our company to a friend or colleague as a place to apply for work? [NPS question]
2. Was the recruiter knowledgeable about the company?
    - Very knowledgeable
    - Somewhat knowledgeable

- Slightly knowledgeable
- Not knowledgeable at all

3. Were the questions asked during your phone interview relevant to the job you applied for?

    - Very relevant
    - Somewhat relevant
    - Slightly relevant
    - Not relevant at all

G.2.2 Social experience

1. How natural did the interaction with the recruiter feel?

    - Very natural
    - Somewhat natural
    - Neutral
    - Somewhat unnatural
    - Very unnatural

2. How frequently did you receive live feedback from the recruiter during your interview?

    - Very frequently
    - Somewhat frequently
    - Occasionally
    - Rarely
    - Never

G.2.3 Perceived discrimination

1. Did you feel discriminated against by the recruiter because of your gender identity?

    - Yes
    - No
    - Not sure

G.2.4 General AI awareness, knowledge and usage

1. Select the correct definition of Generative artificial intelligence (GenAI)

    - GenAI automates genetic analysis to modify DNA sequences for medical purposes. [incorrect]
    - GenAI is AI that creates new content, such as text and images, by learning from existing data. [Correct]
    - GenAI involves the creation of algorithms for solving complex mathematical problems and optimizing industrial processes. [incorrect]
    - GenAI is a system that focuses on generating engineering techniques to enhance crop yield and resistance to pests in agriculture. [incorrect]
    - I don't know [incorrect]

2. How often do you use the following products?

    - ChatGPT [Daily, Weekly, Monthly, Never, I have never heard of it, Not rated N/A]
    - Character.AI [Daily, Weekly, Monthly, Never, I have never heard of it, Not rated N/A]
    - QuillBot [Daily, Weekly, Monthly, Never, I have never heard of it, Not rated N/A]
    - Midjourney [Daily, Weekly, Monthly, Never, I have never heard of it, Not rated N/A]

G.2.5 AI perception on the labor market

Over the next 20 years, how much impact do you think the use of artificial intelligence (AI) in the workplace will have on...

1. Workers generally

    - A major impact
    - A minor impact
    - No impact
    - Not sure
    - No answer

2. You, personally
    - A major impact
    - A minor impact
    - No impact
    - Not sure
    - No answer

G.2.6 Open-ended feedback

1. Do you have any additional feedback you'd like to share about your interview experience?

H Recruiter survey instructions

Welcome! Thank you for taking part in this short survey. The survey asks about your experiences and views on hiring. We are interested in hearing your thoughts and perspectives. There are no right or wrong answers - we're simply interested in your honest opinion. The survey will only take a few minutes to complete, and your responses will remain confidential.

1. Did you evaluate interviews conducted by Anna, our AI voice agent?
    - Yes
    - No

H.1 Predicting differences between human and AI-led interviews

For the next questions, please consider all interviews that there were conducted at PSG in the last three months, i.e., from March to June 2025.

1. Across all interviews, do you expect AI-led interviews to be of **higher, lower, or equal quality** compared to human-led interviews?
    - Much higher quality
    - Slightly higher quality
    - About the same quality

- Slightly lower quality
- Much lower quality

2. Across all interviews, do you expect AI-interviewed candidates to **receive job offers at a higher, lower, or equal rate** compared to human-interviewed candidates?

    - Higher
    - Equal
    - Lower

3. *[Shown only if "Higher" was selected in the previous question 2.]*
   On the previous page, you indicated that you think AI-interviewed candidates receive job offers at a higher rate. **If out of 1,000 candidates who were interviewed by human recruiters, 85 got a job offer, how many of the 1,000 do you think would have gotten a job offer if they were interviewed by the AI instead?**

4. *[Shown only if "Lower" was selected in the previous question 2.]*
   On the previous page, you indicated that you think human-interviewed candidates receive job offers at a higher rate. **If out of 1,000 candidates who were interviewed by human recruiters, 85 got a job offer, how many of the 1,000 do you think would have gotten a job offer if they were interviewed by the AI instead?**

5. Across all candidates who eventually received job offers, do you expect AI-interviewed candidates to **stay longer, shorter, or for a similar length employed** compared to human-interviewed candidates?

    - AI-interviewed are **much longer** employed
    - AI-interviewed are **slightly longer** employed
    - AI-interviewed are **equally long** employed
    - AI-interviewed are **slightly shorter** employed
    - AI-interviewed are **much shorter** employed

6. Across all candidates who eventually started their job, do you expect AI-interviewed candidates to **have higher, the same, or lower on-the-job productivity** compared to human-interviewed candidates?

    - AI-interviewed are **much more** productive

- AI-interviewed are **slightly more** productive
- AI-interviewed are **similarly** productive
- AI-interviewed are **slightly less** productive
- AI-interviewed are **much less** productive

H.2 Experience evaluating AI voice agent interviews

[Shown only if "Yes" selected for the question of whether the recruiter evaluated AI voice agent interviews]

1. Compared to evaluating the interviews you conducted yourself, how **easy/difficult was it for you to evaluate** AI-led interviews in terms of time and effort?

    - Much more difficult to evaluate AI
    - Somewhat more difficult to evaluate AI
    - About the same
    - Somewhat easier to evaluate AI
    - Much easier to evaluate AI

2. When you decide whether to make an offer to an applicant, how **important** are an applicant's [Name of standardized test] test scores compared to their performance in the interview itself?

    - Interview performance is **much more important** than [Name of test] scores
    - Interview performance is **somewhat more important** than [Name of test] scores
    - Interview performance and [Name of test] scores are **equally important**
    - Interview performance is **somewhat less important** than [Name of test] scores
    - Interview performance is **much less important** than [Name of test] scores

3. Compared to the interviews you conducted yourself, how do you rate the **quality of the information** you received from AI-led interviews?

    - Much lower quality from AI
    - Somewhat lower quality from AI
    - About the same quality

- Somewhat better quality from AI
- Much better quality from AI

4. Overall, how would you rate the introduction of AI-interviews in the recruiting process?

   - Very negative
   - Negative
   - Neutral
   - Positive
   - Very positive

5. Compared to the interviews you conducted yourself, **how high are your standards** for applicants from AI-led interviews?

   - Much lower standards for applicants from AI
   - Lower standards for applicants from AI
   - About the same standards
   - Higher standards for applicants from AI
   - Much higher standards for applicants from AI

6. Please share any additional thoughts or suggestions regarding your experience with AI-led interviews, especially improvements or changes you would like to see.

H.3 AI perception on the labor market

Artificial intelligence (AI) can be used by employers to collect and analyze data, make decisions, and complete tasks. Some employers are using AI in hiring, for worker evaluations, or even to do jobs humans used to do.

Over the next 20 years, how much impact do you think the use of artificial intelligence (AI) in the workplace will have on...

7. Workers generally

   - A major impact

- A minor impact
- No impact
- Not sure
- No answer

8. You, personally

    - A major impact
    - A minor impact
    - No impact
    - Not sure
    - No answer

Thinking about the use of artificial intelligence (AI) in the workplace over the next 20 years, what do you think the outcome will be for...

9. Workers generally

    - AI will help more than it hurts
    - AI will equally help and hurt
    - AI will hurt more than it helps
    - Not sure
    - No answer

10. You, personally

    - AI will help more than it hurts
    - AI will equally help and hurt
    - AI will hurt more than it helps
    - Not sure
    - No answer

I Comparing the costs of AI and human interviews

This appendix presents a stylized cost model to compare the cost efficiency of AI-led and human-led interviews. We use approximate estimates of our partner firm's cost structure to compare the cost efficiency of AI-led and human-led interviews in three stylized environments: (i) *static*, in which the accuracy and costs per interview are fixed; (ii) *dynamic*, in which we assume the existence of an AI error rate that decreases with calendar time (foundation model updates and crash rate controlled by the AI vendor of our partner firm) and with the volume of cumulative interviews (improved conversational paths controlled by our partner firm), and (iii) *stochastic*, in which we allow human and AI interview costs to vary stochastically.

Human recruiter costs. We represent the cost of human interviews, c_H, as follows:

$$c_H(w) \;=\; t_H\, w \;+\; b_H,$$

where w is the hourly wage, t_H is interviewer time (minutes per interview) and b_H is a fixed cost per interview covering, for example, bundle supervision, training, turnover and back office overhead. We report results for three representative adjusted calibrations. The cost figures are adjusted estimates and do not represent and should not be interpreted as the actual accounting records of the firm partner.

$$\left(c_H^{\mathrm{L}}, c_H^{\mathrm{M}}, c_H^{\mathrm{H}}\right) = (\$2.48, \$3.50, \$6.37)$$

Here, c_H^L corresponds to wages in low-wage market environments, c_H^M to mid-wage, and c_H^H to high-wage. These calibrations result from the application of a conservative downside adjustment coefficient of 0.67 that we applied to the firm's original communication estimate per interview, ensuring that our cost figures do not overstate the expense of human-led interviews.[24] The mid- and high-cost levels ($3.50 and $6.37, respectively) scale the baseline to reflect the market environments of middle and high wages observed in the firm's global operations. As our experiment occurred in the Philippines, the human costs in our setting correspond to the calibration c_H^L.

AI voice agent costs. For the cost of AI interviews, $c_A I$, we assume three price tiers of the vendor for the marginal cost per interview:

$$\left(c_{\mathrm{AI}}^{\mathrm{L}}, c_{\mathrm{AI}}^{\mathrm{M}}, c_{\mathrm{AI}}^{\mathrm{H}}\right) = (\$\,1.30, \$\,2.06, \$\,3.03)$$

[24] One could as well conduct this analysis with a distribution of values for such a coefficient.

These calibrations are derived from the publicly available API pricing of ElevenLabs, an AI vendor comparable to the one on which our firm partner relies, which we cannot disclose. In August 2025, AI list rates as low as \$0.08 per minute (annual billing).[25] Given a median duration of AI-led interviews of 9.60 minutes in our sample, this implies a baseline marginal cost of \$0.768 per interview. To remain conservative, by accounting for potentially higher API call costs between March 7 and June 7, 2025, as well as other AI-related expenses such as compliance, we apply a coefficient of 1.69 to this baseline, leading to \$1.30 by AI-interview for the low-cost tier. The mid- and high-cost tiers (\$2.06 and \$3.03, respectively) correspond to x2.68 and x3.95 the baseline, reflecting plausible variation in vendor pricing and feature bundling.[26] In all scenarios, to be even more conservative, we also include a one-time fixed deployment cost of $F = \$10,000$ paid to the AI vendor.[27]

Next, we compare these costs across three environments - static, dynamic, and stochastic.

I.1 Static cost environment

In the static environment, both c_{AI} and c_H are constants. We first compute the gaps between them, $c_H - c_{\text{AI}}$, for each level, yielding nine cases. As shown in Panel A of Table I.1, AI remains more costly in only one case ($c_H^{\text{L}} - c_{AI}^{\text{H}}$). For all other cases, the break-even number of interviews follows directly from

$$n^* = \frac{F}{c_H - c_{\text{AI}}}, \qquad F = \$10{,}000. \tag{1}$$

For example, when the firm faces c_{AI}^{L} and c_H^{H}, AI becomes cost effective after only 1,972 interviews. In contrast, with c_{AI}^{M} and c_H^{L}, AI requires 23,810 interviews to break even. Finally, if the human cost remains low (c_H^{L}) while the AI cost is high (c_{AI}^{H}), AI never catches up to the human benchmark. Thus, AI adoption in static settings is highly sensitive to the relative positioning of cost tiers.

[25] ElevenLabs API pricing for "Conversational AI" (annual subscription, highest offered level). At this tier, start-ups receive 22,000 minutes per month of AI voice agent usage, which would cover our firm partner's monthly needs during the simulated experiment period (March 7–June 7, 2025). Accessed on August 17, 2025.

[26] Examples of these features that may incur additional cost: a larger menu of voice accents or diverse gender tones, multilingual support, more advanced generative AI models powering the voice agent, and different audio quality levels – ranging from standard 128kbps MP3 to 192kbps MP3 at 44.1kHz, or even higher 'ultra" and ultra lossless" settings – as well as access to additional minutes beyond the tier limit, or the use of premium speech-to-text models.

[27] This fixed cost could alternatively be structured as a recurring monthly subscription, in addition to per-API-call charges.

I.2 Dynamic cost environment

In the dynamic environment, we assume an AI error rate, $\varphi(t)$, i.e., the percentage of AI-led interviews that terminate prematurely due to *AI system crash*. For example, the voice agent or its back-end API ends the call unexpectedly and produces an unusable or incomplete transcript, a phenomenon that we also observe in our data. Costs are thus:

$$c_{\text{AI}}^{\text{eff}}(t) = c_{\text{AI}} + \varphi(t)c_H, \tag{2}$$

We calibrate two conservative anchors for $\varphi(t)$. First, a launch-month crash rate of $\varphi(0) = 0.25$, well above any observed monthly rate, to avoid understating the early missing costs that are potentially absent from this reduced-form approach. Second, a one-year crash rate of $\varphi(12) = 0.05$, only assuming modest stability gains beyond the 7% average observed to date, just after three months of the AI agent's launch.[28] The actual launch month rate was far below 25%, hence this choice inflates early periods, making our break-even estimates conservative.

The following functional form combines calendar-time improvement from upstream model updates and throughput-driven learning-by-doing:

$$\varphi(t) = \varphi_0 \ \exp(-\kappa t) \left(\tfrac{n_0+\lambda t}{n_0}\right)^{-\gamma}, \qquad n(t) = n_0 + \lambda t. \tag{3}$$

Here, κ captures *calendar* improvements, λ is throughput (interviews/month), γ is the learning elasticity *volume*, and n_0 scales the onset of experience. The instantaneous log-improvement is

$$\frac{d\log\varphi(t)}{dt} = -\kappa - \gamma\frac{\lambda}{n_0+\lambda t},$$

so both channels operate; the volume term dominates early.

These two calibration points $(\varphi(0), \varphi(12))$ identify a one-parameter family of (κ, γ) consistent with throughput λ and the onset scale n_0. From

$$\frac{\varphi(12)}{\varphi(0)} = \exp(-12\kappa)\left(\tfrac{n_0+12\lambda}{n_0}\right)^{-\gamma},$$

we obtain

$$\kappa \ = \ \frac{-\ln\big[\varphi(12)/\varphi(0)\big] \ - \ \gamma\,\ln\big[(n_0+12\lambda)/n_0\big]}{12}.$$

[28] This value comes from our classification of the proportion of interviews labeled as "AI System Failure" in Figure A.6, based on transcript-level completion classifications covering our pre-registered experimental period between March 7 and June 7, 2025.

Table I.1: Static and one-year dynamic break-even counts under three AI prices and wage environments

Panel A: Static cost gaps and break-even interviews

Market	Low AI $1.30		Mid AI $2.06		High AI $3.03	
	Gap ($)	n^* (int.)	Gap ($)	n^* (int.)	Gap ($)	n^* (int.)
Low-income	1.18	8,475	0.42	23,810	−0.55	—
Mid-income	2.20	4,545	1.44	6,944	0.47	21,277
High-income	5.07	1,972	4.31	2,320	3.34	2,994

Panel B: One-year dynamics ($\varphi_{12} = 5\%$, $\lambda = 5{,}000$/mo)

Market	Low AI $1.30		Mid AI $2.06		High AI $3.03	
	Gap ($)	n^*_{12} (int.)	Gap ($)	n^*_{12} (int.)	Gap ($)	n^*_{12} (int.)
Low-income	1.06	9,434	0.30	33,333	−0.67	—
Mid-income	2.02	4,950	1.27	7,874	0.30	33,333
High-income	4.75	2,105	3.99	2,506	3.02	3,311

Notes: "—" denotes a non-positive gap (AI not cost-competitive).

Then, the approximation of one year (end-of-year gap) with $\varphi(12) = 0.05$ is:

$$n^*_{12} = \frac{F}{c_H - c^{\text{eff}}_{\text{AI}}(12)} = \frac{F}{c_H - \left[c_{\text{AI}} + 0.05\, c_H\right]},$$

which treats the year as if all interviews occurred in the month 12 gap.[29]

As shown in panel B of Table I.1, incorporating a one-year crash rate of $\varphi(12) = 5\%$ slightly reduces the cost gaps and increases the break-even thresholds compared to the static environment. First, in the low-income market, AI at the low-price tier remains cost-competitive but requires 9,434 interviews to break even, while at the mid-tier the threshold raises to 33,333 interviews. At the high price of AI, the gap is negative, so AI is never cost-competitive. Second, in the mid-income market, the break-even occurs after 4,950 interviews at the low AI price and 7,874 interviews at the mid-tier; at the high price, the threshold becomes cost-competitive at 33,333 interviews. Third, in the high-income market, AI breaks very quickly across all tiers, with 2,105 interviews at the low price, 2,506 at the mid-tier, and 3,311 at the high price. In other words, introducing dynamics shifts the break-even counts upward while maintaining the same qualitative ranking as in the static environment: AI pays off fastest with higher human wages and lower AI prices.

[29] Since early-month gaps are smaller, this is a *lower-bound (optimistic)* count for the true cumulative break-even: $n_{\text{BE,true}} \geq n^*_{12}$.

Stochastic environment. Lastly, we allow human and AI interview costs to vary stochastically. Specifically, we introduce random variation by allowing human and AI interview costs to follow correlated lognormal processes. The results show that greater variability in human costs increases the AI-automation advantage in costs, while higher correlations in cost shocks decrease it. For details, see Appendix Section I.3.

I.3 Stochastic cost environment

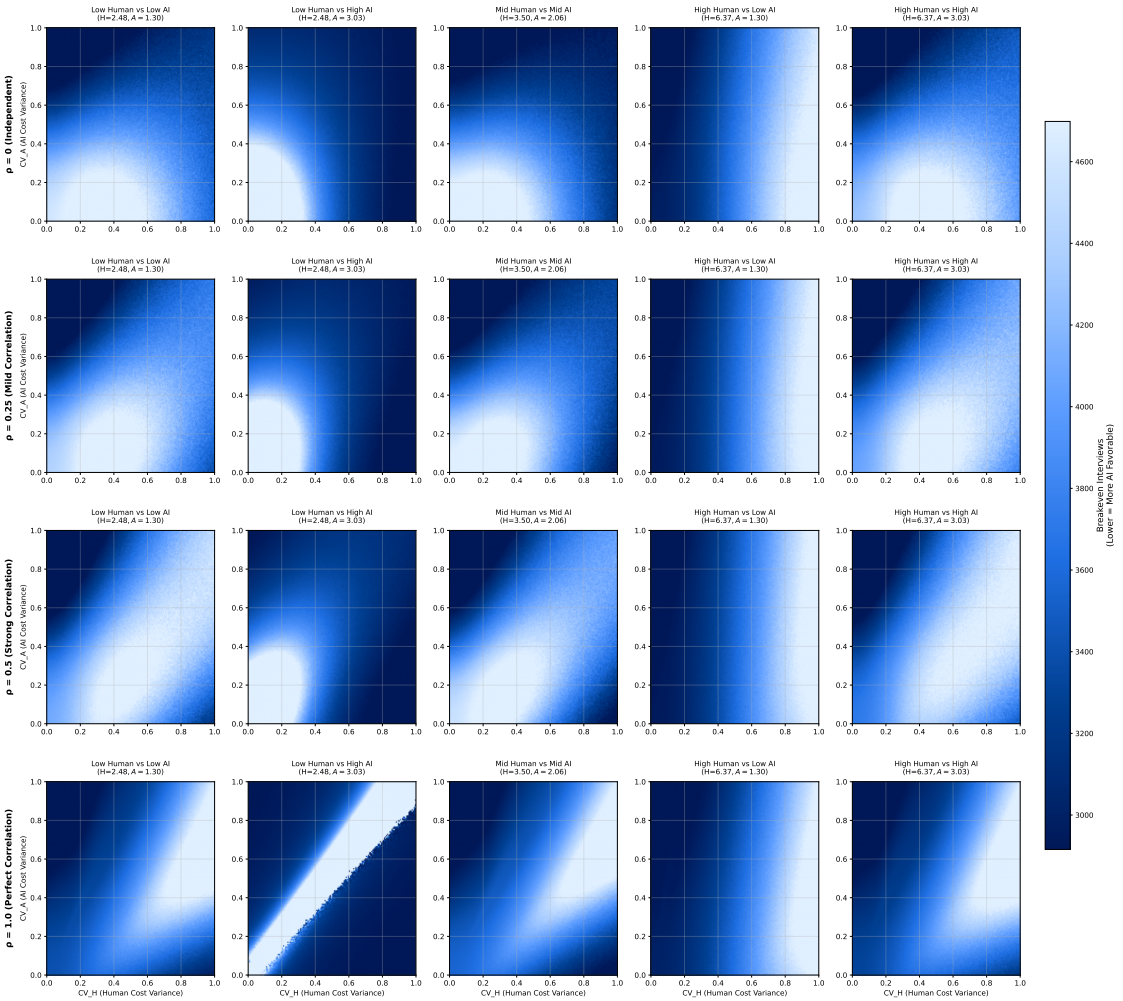
Lastly, we allow human and AI interview costs to vary stochastically. Specifically, we assume that per-interview costs follow correlated lognormal processes, capturing volatility in labor costs and fluctuations in AI usage or vendor pricing. For each combination of the coefficients of variation of human and AI costs, (CV_H, CV_A), and the correlation of shocks $\rho \in \{0, 0.25, 0.5, 1\}$, we simulate 50,000 draws and compute the median break-even number of interviews:

$$n^* = \frac{F}{c_H - c_{\text{AI}}^{\text{eff}}}, \qquad c_{\text{AI}}^{\text{eff}} = c_{\text{AI}} + \phi\, c_H,$$

where F is the fixed cost of AI adoption and ϕ is the percentage of interviews conducted by human recruiters within the AI regime.

Figure I.1 plots the resulting frontier across cost volatility and correlation parameters, with regions in darker blue indicating faster AI cost efficiency (fewer interviews required to break even). Our simulations in Figure I.1 provide several insights on the dynamics of AI adoption and how it varies with the stability of human and AI costs. First, when human interview costs fluctuate more than those of the AI system, the fixed deployment cost of automating interviews with AI is recovered after fewer interviews. Second, as the correlation between human and AI cost shocks increases, this automation advantage vanishes, and the frontier collapses toward a diagonal where volatility symmetry leaves only mean costs to matter. Third, across cost regimes, higher wage environments shift the entire frontier outward, indicating that AI becomes cost-effective more quickly when human labor is both expensive and unstable.

Figure I.1: Dynamic break-even frontier across cost volatility and correlation



Notes: Each panel reports the median number of interviews required for AI-led interviewing to become cumulatively cheaper than human interviewing, simulated over 50,000 lognormal cost draws. Columns vary baseline cost levels; rows vary the correlation between human and AI cost shocks. Darker blue regions indicate faster AI cost efficiency (fewer interviews to break even).

J Text analysis: prompt content

J.1 Speaker labeling prompt

SPEAKER LABELING PROMPT

```
1 You are an expert conversation analyst tasked with meticulously labeling
    an interview transcript. Your goal is to accurately identify each
    speaker and label their lines with either "Interviewer:" or "Candidate
    :". Do not repeat any portions of the original text. Only output the
```

```
    labeled transcript.

If it’s unclear who is speaking, make your best judgment based on the
    overall context of the conversation.

Return the transcript with speaker labels in the format "Speaker:
    Dialogue". *Place each speaker’s turn on a new line, but do not
    include \textbackslash n within the line itself.

\#\#\# Example

Input:

[Anonymized]: Hi, this is [Anonymized] from Teleperformance Recruitment.
    Ill be conducting an interview and Ill be asking you personal details
    and some common interview questions in order to give you feedback on
    how I should proceed your application, okay?

[Name applicant]: Okay.

[Anonymized]: And I just want to remind you that this phone call is
    recorded for quality assurance purposes, okay?
[Name applicant]: Alright, got it.

[Anonymized]: Okay

Output:

Interviewer: Hi, this is [Anonymized] from Teleperformance Recruitment.
    Ill be conducting an interview and Ill be asking you personal details
    and some common interview questions in order to give you feedback on
    how I should proceed your application, okay?

Candidate: Okay.

Interviewer: And I just want to remind you that this phone call is
    recorded for quality assurance purposes, okay?
```

```
Candidate: Alright, got it.

Interviewer: Okay

Here is the transcript to label: [omitted for privacy reasons].
```

J.2 Anonymization prompt

ANONYMIZATION PROMPT

```
You are an expert transcript anonymizer.
Your job is to remove any personally identifiable information (PII) from
    the following text. That includes:
- Person names (replace with **AnonymizedNAME**)
- Email addresses (replace with **AnonymizedEMAIL**)
- Phone numbers (replace with **AnonymizedPHONE**)
- Organization names (replace with **AnonymizedORG**)
- Street address or city or location (replace with **AnonymizedADDRESS**)
- Date of Birth (replace with **AnonymizedDOB**)
Return only the cleaned text - do not add any explanation.
Text to anonymize:
"""{text}"""
```

J.3 Interview classification prompt

The following shows the prompt used to classify interviews into types. Note that the prompt contains the additional category "Expectation Mismatch". For the paper, we put this category under "Other" given that less than 1% of interviews have this category.

INTERVIEW CLASSIFICATION PROMPT

```
You are an interview expert. Your task is to classify the provided
    interview data into one of the following categories.

Interview Transcript:
1. Interview Transcript: {transcript}

```

```
Interview Meta Data:
1. Call Duration: {duration} minutes
2. Number of Topics Covered: {topic_count}
3. Treatment: {treatment}

Instructions:
1. Evaluate interview transcript and metadata based on order of priority, starting from Priority 1 (Interview-Stopping Events), Priority 2 (Screening Out Interviews), Priority 3 (Interview Analysis), to Priority 4 (Others).
2. Provide your reasoning. In your explanation, cite specific phrases from the transcript and data points (like topic_count or call duration) to justify your choice.

Category Definitions Based on Order of Priority:
Priority 1 - Interview-Stopping Events
- Check for these following categories first

1. Candidate Unavailability: The candidate explicitly states they are currently unable to talk (e.g., "I'm driving," "I'm in a meeting," "Can I call you back later?"). The call ends quickly due to this reason.

2. AI Aversion: The candidate explicitly expresses unwillingness to continue speaking with an AI recruiter. e.g., "I'm not comfortable talking to AI," "I'd prefer to speak with a human recruiter.")
   - Note: Do not classify human interviews (Treatment that contains "Human Interview") into "AI Aversion"
3. Telephony Failure: Issues with cellular network, signal loss, or VOIP instability. Conversation is not finished, no conclusion remark
   - Note: when there are some repeated questions, topic_count >= 8, conversation ends with concluding remark, it cannot be classified as "Telephone Failure"
4. AI System Failure: The LLM/voice agent malfunctions (e.g., stalls, crashes, fails to respond, repeats itself endlessly). This is a failure of the AI itself, not the connection.
   - Important Exclusion: This category does not apply if an interview recovers and ends with concluding remark. A few repeated question
```

```
        does not automatically mean "AI System Failure"
    - Note: Do not classify human interviews (Treatment that contains "
        Human Interview") into "AI System Failure"


Priority 2 - Screening Out Interviews
- If interview-stopping events do not apply, check for these following
    categories

5. Early Screen-Out: The interview ends early because the candidate is
    immediately disqualified based on a non-negotiable requirement related
     to the job (e.g., salary expectations, location, visa status).
    Important rules include topic_count is low (0-2) and call duration is
    short. Recruiter states the reason for ending the call due to
    disqualification.
    - Important Rule: A short duration alone does not automatically mean "
        Early Screen-Out." "Early Screen-Out" requires an explicit
        disqualification based on a non-negotiable requirement stated by
        the recruiter.
    - Important Rule: if a call ends without concluding remark from the
        recruiter, "Early Screen-Out" does not apply
6. Midway Screen-Out: The interview ends after some initial engagement
    due to a mismatch discovered during the conversation (e.g.,
    availability issues that weren't immediately apparent, conflicting
    school plans, a skill gap). Important rules include topic_count is
    moderate (3-7) and call duration is in the middle. Recruiter states
    the reason for ending the call due to disqualification.
    - Important Rule: if a call ends without concluding remark from the
        recruiter, "Midway Screen-Out" does not apply
7. Late Screen-Out: The interview proceeds nearly to completion but the
    candidate fails a final, critical criterion (e.g., rehire status, a
    serious attitudinal concern revealed late in the interview). Important
     rules include topic_count is high (>= 8) and call duration is
    typically long. Recruiter states the reason for ending the call due to
     disqualification.
    - Important Rule: if a call ends without concluding remark from the
        recruiter, "Late Screen-Out" does not apply
```

```
39
40 Priority 3 - Interview Analysis
41 - If none of above categories apply, check for these following categories
42
43 8. Disengaged Interaction: The candidate demonstrates disinterest,
     unresponsive, distracted, and poor continuity. This category applies
     when the candidate initially engaged in the conversation.
44    - Number of topics is strictly less than 8 topics (topic_count < 8)
45 9. Comprehensive Interview: Natural opening and closure; topic_count at
     least 8 (>= 8) expected topics or more; high-quality engagement from
     both parties.
46    - The candidate answers questions fully and asks relevant questions.
47    - There is a concluding remark. e.g. "We'll see you on the next hiring
          process. Good luck!"
48 10. Expectation Mismatch: A full interview is conducted, topic_count is
     high (>= 8), but the candidate has a fundamental misunderstanding of
     the role, the company, or the requirements.
49
50 Priority 4 - Others
51 - If none of above categories apply, check for the following category
52 11. Others: An interview is not in English or does not fit to above
     categories.
53
54 Expected output:
55 1. Category: [Chosen Category Name]
56 2. Explanation: [Your concise explanation, citing evidence from the
     provided data and transcript.]
57
58 """
```

J.4 Topic coverage prompt

TOPIC COVERAGE PROMPT

```
You are a highly skilled conversation analyst reviewing interview transcripts. Your task is to determine whether the following interview conversation covers a specific list of topics.

**Topics to Identify:**
['INTRODUCTION', 'SOURCE VERIFICATION', 'LOCATION/COMMUTE/TRANSPORTATION VERIFICATION', 'CHECKING FOR RED FLAGS/COMMITMENT/ATTRITION RISKS', 'EDUCATION VERIFICATION', 'COMPENSATION', 'SCREENING FOR EMPLOYMENT HISTORY', 'REHIRE ELIGIBILITY CHECK', 'AVAILABILITY', 'DATA VERIFICATION', 'NEEDS ASSESSMENT (in preparation for AMCAT)', 'PROFILING', 'ICIMS \& AMCAT', 'REFERRAL']

**Instructions:**

1. Carefully analyze the provided interview transcript to determine which of the topics listed above are substantively discussed.
2. SOURCE VERIFICATION is to understand where candidates find the opening
3. PROFILING is to understand whether candidates have worked in the company before, their past work experiences, or skills related to the job. If no, then it will explain briefly about the company
4. For each topic, consider whether the conversation includes sufficient information or questions to indicate that the topic was genuinely addressed. Brief mentions or passing references should *not* be considered as "covered."
5. Organize your output into three distinct sections: "Topics Covered," "Topics Not Covered," and "Explanations."
6. For the "Explanations" section, provide a concise justification (1-2 sentences) for why each topic is classified as "covered" or "not covered," referencing specific parts of the transcript if possible.

**Output Format:**

*  **Topics Covered:** [List of covered topics, e.g., ['INTRODUCTION', 'SOURCE VERIFICATION']]
```

```
* **Topics Not Covered:** [List of topics not covered, e.g., ['AVAILABILITY', 'COMPENSATION']]
* **Explanations:**
    * INTRODUCTION: [Explanation of why INTRODUCTION is covered, e.g., "The interviewer and candidate exchanged greetings and discussed the purpose of the interview."]
    * SOURCE VERIFICATION: [Explanation of why SOURCE VERIFICATION is covered, e.g., "The interviewer asked the candidate how they found the job posting."]
    * AVAILABILITY: [Explanation of why AVAILABILITY is not covered, e.g., "The interviewer did not ask about the candidate's start date or work schedule."]
    * ... (Continue for all topics, both covered and not covered)
**Interview Transcript:**
"""{text}"""
```

J.5 Topic order prompt

TOPIC COVERAGE PROMPT

```
You are a highly skilled conversation analyst reviewing interview transcripts.

INSTRUCTIONS:
1. Your task is to assign a unique topic id (integer, starting from 1)
2. Output the sequence of topic id as they appear in the transcript
3. If a topic reappears in later conversations, use the same topic id again and do not create a new topic id

TOPICS TO IDENTIFY:
1. INTRODUCTION
- Hi (Candidates name), my name is ______ with the Teleperformance recruiting team. I am calling to thank you for applying to the _________ job opening and was hoping you had time to speak with me about your prior work experience. Is this a good time to talk?

2. SOURCE VERIFICATION
```

- Before we begin, can you share with me where you found this job opportunity listed? We like to track where candidates are finding our job openings.

3. LOCATION/COMMUTE/TRANSPORTATION VERIFICATION
- Ok, let’s jump into some more specific questions. Where do you reside?
- This position is an onsite position and is located at _____, how far a drive would this be for you?
- What mode of transportation would you use?

4. CHECKING FOR RED FLAGS/COMMITMENT/ATTRITION RISKS
- The minimum age requirement for this position is 18 years of age, are you at least 18 years of age or older?
- Most of the opportunities we have are on a shifting schedule where you could even work during holidays and/or weekends. Would you be okay with this type of schedule?
- Also, would you be willing to render over time if necessary?
- Can you share with me what your main motivations are to consider this job opportunity?
- What motivated you to apply to this Teleperformance job opening?
- As you consider your career goals, how long could you envision working for Teleperformance and performing the duties of this job?
- How do your short and long term goals align to the opportunities this job presents?
- The training process we offer is mandatory to attend.
- Do you have any planned vacation in the next 6mos?
- What are the activities that may affect your attendance or any prior commitments that you have to attend to?

5. EDUCATION VERIFICATION
- What is your highest level of education completed?
- If you are a college grad, do you plan to return to school to complete a post undergraduate degree?
- Why did you not continue with (HS/College)

6. COMPENSATION

- If we can shift our discussion to compensation expecations, that would be helpful. What is your expected salary?
- The roles that we have open right now are between the salary package range of (Salary Range of available campaigns); depending on your qualifications, you could be offered a salary package between these amounts. Would you be okay with this range?

7. SCREENING FOR EMPLOYMENT HISTORY
- I’d like to talk more about your call center experience that we spoke about earlier. Can you share with me the volume of calls you managed on a daily basis? Were these calls solving technical issues or were you responding to general inquiries? Were you ever put on a performance plan during your prior employment?
- Tell me about your last 3 companies, starting with the most recent. - (Get employment dates, location)
- What was your role?
- Why did you leave the company?

8. REHIRE ELIGIBILITY CHECK
- A few more questions and we are almost done. Have you ever worked for Teleperformance in the past?
- If yes, where did you work and do you recall your prior manager’s name?

9. AVAILABILITY
- If you are offered the job, when would you be able to start?
- Do you have any pending applications with other companies? If yes, what stage in the interview process are you in?

10. DATA VERIFICATION
- What are your Contact Number & Alternate Contact Number (if there’s any)?
- What is your Viber Account Number?
- When is your Birthdate?
- What is your Gender:
- What is your Marital/Civil Status:
- Who is your Emergency Contact Person (Name, Relationship, Number):
- When is your date of birth?

```
11. NEEDS ASSESSMENT (in preparation for AMCAT)
- Do you have a reliable internet connection? Do you have a laptop/PC available at home?

12. COMPANY OVERVIEW
- Thanks for answering my questions. Based on your background, I can see that you’re a good fit for the CSR/TSR position I’m working on for Teleperformance.
- TP is one of the largest BPO Companies. Our client has over 20 sites in the PH alone. They have different accounts that offer paid training, HMO, and leadership opportunities. Are you ready to hear more about the job?

13. ICIMS & AMCAT
- As soon as you’re finished with the assessment, please let me know and send me a screenshot of your completion. Take note of the time stamp after the assessment. Then I will keep in touch with you with the hour. Do you have any questions so far?

14. REFERRAL
- Btw, we have multiple openings here in TP, can you recommend anyone that has similar skills and experience such as yourself? Could you provide me their names and phone numbers?

15. GENERAL CONVERSATION / TRANSITION
- Okay, thank you for that.
- How your day has been?

OUTPUT:
{{
  "topics": [{{"id": 1, "name": "INTRODUCTION"}}, ...],
  "sequence": [1,2,1,3,...]
}}

OUTPUT INSTRUCTIONS:
1. Do not wrap the json in markdown code fences, like ‘‘‘json ... ‘‘‘
```

```
2. Entire output should be a valid, single, well-formed json object
3. JSON object have to be pretty-printed with an indentation of 2 spaces
4. Output should directly begin with {{ and }}

INTERVIEW TRANSCRIPT:
\"\"\"{text}\"\"\"
```

J.6 Interview review classification prompt

Interview Review Classification Prompt

```
def make_prompt(text:str) -> str:
   return f"""
You are a sentiment analysis assistant for recruiter feedback.
Classify the following recruiter comment as:
- Positive
- Neutral
- Negative
Do not classify multiple sentiments in a comment
**Output Format**
Positive, Neutral, Negative
**Recruiter Comment:**
\\\{text}\\\
"""
```

K Research transparency

We pre-registered the experiment at the AEA RCT Registry (trial number #15385, link: https://www.socialscienceregistry.org/trials/15385). The pre-registration includes details on the experimental design, the planned sample size, variables that were expected to be collected, and an outline of the hypotheses and analysis plan. In the following, we describe in more detail the mapping between the paper and pre-registration.

Sample size. As pre-registered, we included in the experiment all applicants who applied between March 7 and June 7, 2025. For this time period, based on pre-experimental

data, we calculated an expected sample size for the experiment of around 27,000 applications and stated this number in the pre-registration. However, we also noted that "The actual sample size depends on several external factors such as demand from the firms commissioning the client companies. Therefore, the sample size is not directly in the researchers' control, meaning the actual sample size may be (substantially) higher or lower than expected." Indeed, our realized sample size was substantially higher than the expected one during the pre-registered period in which the experiment took place.

Treatment conditions. Our three experimental conditions were implemented as pre-registered. The randomization weights put on each treatment condition were at the discretion of the firm. We stated in the pre-registration that we expect all three conditions to include more than 10% of applicants in a given month. During the experiment, the firm implemented fixed weights of 60% *AI interviewer*, 20% *Human interviewer*, and 20% *Choice of interviewer*. Realized fractions closely match these numbers.

Excluding observations. We pre-registered that we would exclude any applications that were not invited for an interview (and thus not randomized into one of the three conditions). Following this, we excluded 3,828 applications. As pre-registered, for analyses involving the *Choice of interviewer* condition, we focus only on applicants who made a choice.

Key outcome variables. As key outcome variables, we pre-registered the following variables (in the order as they appear in the pre-registration):

- Whether the interview was successfully completed
  - We initially pre-registered a binary variable, but then conducted a more detailed interview-type analysis in Section 4.3 because we received transcript data.
- Whether the applicant receives a job offer
  - This variable is reported in Section 3.
- Interviewer score
  - This variable is reported in Section 6.1.
- Interviewer comment (open-ended text)
  - This variable is reported in Section 6.1.

- Whether the applicant accepts an offer conditional on receiving one
  - This variable is reported in Section 5.1.
- Time from initial application until the interview takes place
  - This variable is reported in Section 7.1.
- Time from initial application until a final decision has been made
  - This variable is reported in Section 7.1.
- Whether the applicant started their job at the respective firm
  - This variable is reported in Section 3.
- Retention rate: whether the applicant still works at the firm after starting their job.
  - This variable is reported in Section 3.

Further pre-registered variables. We further pre-registered to focus on several interview transcript and interview audio variables under the condition that we would receive them, which was not clear at the time of writing the pre-registration. The firm indeed shared the transcripts, but not audio files. Interview transcript variables are analyzed in Section 4. We pre-registered analyzing eight transcript variables. We implemented the following six (in brackets we denote the variable names as reported in the paper): vocabulary richness (vocabulary richness score), filler words and hedging (filler word frequency), turn-taking behavior (Number of exchanges interviewer-applicant), response length (Number of questions by applicant), question-answer alignment (linguistic style match index), conversation frictions (discourse marker frequency). Instead of the two pre-registered variables sentiment polarity and redundancy of information shared, we implemented the variables backchannel cue frequency and syntactic complexity score. Our conclusions are similar if we instead use the pre-registered variables. We pre-registered the applicant survey and report its results.

Variables not pre-registered. Analyses that are based on applicants' standardized test scores, workers' separation reasons, and their performance measures were not part of the pre-registration. The reason is that at the time the pre-registration was written, it was unclear whether the firm would provide data access to these variables. The recruiter survey was not pre-registered. Importantly, the variables and thus the analyses of recruiter behavior in interviews (Section 4.2) were also not pre-registered.