The University of Chicago Booth School of Business

Working Paper No. 25-17

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

Brian Jabarian
University of Chicago Booth School of Business

Luca Henkel
Erasmus University Rotterdam

Voice AI in Firms

A Natural Field Experiment on Automated Job Interviews

Brian Jabarian⋆ Luca Henkel

⋆Job Market Paper

This version: August 18, 2025

Click here for the most recent version

Abstract

We study the impact of replacing human recruiters with AI voice agents to conduct job interviews. Partnering with a recruitment firm, we conducted a natural field experiment in which 70,000 applicants were randomly assigned to be interviewed by human recruiters, AI voice agents, or given a choice between the two. In all three conditions, human recruiters evaluated interviews and made hiring decisions based on applicants' performance in the interview and a standardized test. Contrary to the forecasts of professional recruiters, we find that AI-led interviews increase job offers by 12%, job starts by 18%, and 30-day retention by 17% among all applicants. Applicants accept job offers with a similar likelihood and rate interview, as well as recruiter quality, similarly in a customer experience survey. When offered the choice, 78% of applicants choose the AI recruiter, and we find evidence that applicants with lower test scores are more likely to choose AI. Analyzing interview transcripts reveals that AI-led interviews elicit more hiring-relevant information from applicants compared to human-led interviews. Recruiters score the interview performance of AI-interviewed applicants higher, but place greater weight on standardized tests in their hiring decisions. Overall, we provide evidence that AI can match human recruiters in conducting job interviews while preserving applicants' satisfaction and firm operations.

Contact: Brian Jabarian: Booth School of Business, University of Chicago, CDR, CAAI, CESifo, J-PAL, IGL, JILAEE, brian.jabarian@chicagobooth.edu. Luca Henkel: Erasmus University Rotterdam, ERIM, CESifo, JILAEE, IZA, Tinbergen Institute; henkel@ese.eur.nl.

Acknowledgements: We thank Iav Bojinov, Joshua Deutschmann, Drew Fudenberg, Avi Goldfarb, Alex Imas, Jon Kleinberg, Hyunjin Kim, Rem Koning, Andrew Koh, David Laibson, Devin Pope, John A. List, Paolo Piacquadio, Pëllumb Reshidi, Frederik Schwerter, Frank Schilbach, Sahana Subramanyam, Richard Thaler, Catherine Tucker, George Wu, and Luigi Zingales as well as the audiences at the CFX Conference at INSEAD, HBS TOM, Chicago Booth, Stanford CASBS, and Chicago Econ for valuable feedback and suggestions. Bernard Chen, 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 Nicole, Jon-Ray Riveras, Ram Sabaru, James Shen, and Shraddha B. for their indispensable logistical support. At Google Cloud, we are grateful for the cloud 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 University of Chicago Booth Center for Applied Artificial Intelligence, the Becker-Friedman Institute Program in Behavioral Economics Research, Google Cloud Research Program, and the Swiss National Foundation.

Transparency: The study was approved by the ethical committee of the University of Chicago (#IRB24-1894 and #IRB25-1002) and pre-registered on the AEA RCT Registry (#15385). See Appendix K for details. This study was carried out 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 Specialized Services.

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. After the study was completed, on August 1, 2025, Brian Jabarian accepted an unpaid advisory position as Chief AI Economist at PSG Global Solutions (a Teleperformance subsidiary) and a renewed five-year research partnership providing extended access for future data and field experiments.

1 Introduction

Spoken conversations are a complex yet central way to gather information relevant to decision-making. Their interactive nature allows decision-makers to probe other people's motivations, attributes, skills, and needs, often more efficiently than any other form of communication. Accordingly, the use of conversations to aid decisions is widespread: venture capitalists infer entrepreneurs' motivations, bank tellers collect customers' financial details, teachers assess students' knowledge, and doctors identify patients' medical needs. The content and trajectory of conversations, in turn, govern how capital is allocated, financial access granted, educational qualifications awarded, and medical care provided.

Until recently, only humans could initiate and sustain prolonged conversations with each other. Advances in language processing, speech synthesis, and real-time interaction systems have now equipped generative artificial intelligence (AI) with the technical capabilities to engage in vocal conversations with humans as well. These *AI voice agents* process natural language inputs and generate dynamically tailored spoken responses. This enables decision-makers for the first time to use AI to gather information from humans through live and natural conversations. However, evidence on the real-world impact of AI voice agents remains scarce. In particular, evidence on their economic value to firms and how humans respond to them in the workplace is virtually absent.

This paper investigates the impact of AI voice agents in a high-stakes setting: job interviews. Being a decisive stage of the matching of applicants with firms, job interviews allow firms to collect in-depth information from applicants. We conducted a large-scale natural field experiment in which applicants were randomly assigned to be interviewed by either a human recruiter or an AI voice agent, while leaving the hiring decision to human recruiters. We collected comprehensive data at the interview, applicant, recruiter, and firm level to evaluate the consequences of substituting human interviewers with AI voice agents. Specifically, we assess (i) the AI's performance as interviewer, (ii) applicants' reaction to facing an AI interviewer, (iii) recruiters' response to the applicant information collected by the AI, and (iv) consequences for the hiring firm, both in terms of the costs of the recruitment process and the performance of hired applicants.

We conduct our natural field experiment through a partnership with one of the global leaders in recruiting process outsourcing (RPO), PSG Global Solutions, a subsidiary of Teleperformance (TP) Specialized Services (hereafter referred to as "the firm"). Our sample consists of 70,884 applications for 48 different job postings located in the Philippines and 43 different client accounts (23 Fortune 500 and 20 European Leaders), who cover all major industries. Applicants apply for entry-level customer service positions. In the hiring process, applicants who pass an initial screening are randomly assigned to one of the following three interview conditions. In the *Human interviewer* condition, they speak with a human recruiter in their interview. In the *AI interviewer* condition, they speak with an AI voice agent (named Anna AI), which reveals its artificial identity at the beginning of the interview. Lastly, in the *Choice of interviewer* condition, applicants can choose whether a human recruiter or an AI voice agent interviews them. Importantly, in all conditions, a human recruiter reviews the conversation after the interview and makes a hiring decision, regardless of whether a human or an AI conducted the interview (and applicants are informed about this).[1] This feature allows us to separate the interactional component of the hiring process from the evaluative component.

During interviews, recruiters follow a structured guide that lists the required questions and topics, but can tailor follow-ups to each applicant. Interviews begin with questions about the eligibility and suitability of the applicant, proceed through career goals, experience, and education, and conclude with job details and an opportunity for applicants to ask questions. The AI voice agent uses the same guide and sequencing. After the interview, applicants take a standardized test assessing language and analytical skills. Recruiters then assess applicants' performance in the interview and test, and afterwards make a threshold-based hiring decision: an applicant is of sufficient quality or not. 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. To successfully start their job, applicants who receive an offer must still pass job-specific validation checks and complete onboarding and training.

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. Interestingly, a forecast survey we conducted with the firm's recruiters revealed that they expected the AI voice agent to perform worse in terms of interview quality, offer likelihoods, and retention rates. In contrast, 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 and a 17% higher likelihood of having an employment spell lasting at least one month, the effects being significant at the 1% level in all instances. We also find positive effects for applicants from AI-led interviews

[1] Recruiters review both AI-led and human-led interviews. When human-led, they only review interviews conducted by themselves.

when we condition our sample on applicants who have accepted their job offer. Among those, applicants in the *AI Interviewer* condition have a 6% higher likelihood of starting their job ($p = 0.005$) relative to applicants in the *Human Interviewer* condition. They also have a 6% higher likelihood of still being employed after one month ($p = 0.036$).

To understand why offer rates are higher under AI, we apply natural language processing techniques to the interview transcripts. We first confirm that recruitment interview scores and comments, together with text-based features extracted from interview transcripts, strongly predict offer decisions in the *Human Interviewer* condition, even after controlling for standardized test scores.[2] AI-led interviews are more likely to be comprehensive, defined as meeting a threshold of required topics and an organic opening and closing, compared to human-led interviews (42% vs. 39%). In general, they cover more topics on average. We also extract linguistic features of applicants' responses in the interviews. AI-led interviews exhibit higher values on the linguistic features that, in human-led interviews, predict higher offer rates (e.g., conversational exchange) and lower values on features linked to lower rates (e.g., backchannel cues, applicant-posed questions). These patterns indicate that information elicited through AI voice agents is more relevant to recruiters while reducing low-signal behaviors, possibly due to its consistent prompting and structured turn-taking. 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 voice agent had technical difficulties.

Next, we analyze applicants' responses to the introduction of AI voice agents, finding no evidence of backlash against their introduction. 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, where AI-led interviews are perceived as significantly less natural, and in reported gender-based discrimination. Here, switching to AI nearly halves the rate of reported discrimination (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 re-

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

spondents 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 human recruiters evaluate applicants. A total of 131 recruiters evaluated applicants, with a core group of 43 handling most of them. Recruiters give significantly higher interview scores to applicants interviewed by the AI voice agent compared to those they interviewed 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 quantitative 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 facing applicants interviewed by AI, they put less weight on interview scores and more on language test scores compared to the applicants interviewed by themselves.

Related literature. Our paper contributes to four 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 particular, AI-based writing assistance increases output and quality in a professional writing task (Noy and Zhang, 2023), exam scores of law students (Peng et al., 2023), and job postings by employers (Wiles and Horton, 2024). Similarly, generative AI tools help customer support workers resolve issues quicker (Brynjolfsson et al., 2025), software developers code faster (Peng et al., 2023), writers become more creative (Doshi and Hauser, 2023), and students solve math problems better (Kumar et al., 2025). Studying AI-based business advice, Otis et al. (2025) show that it has uneven effects on entrepreneurs' revenues and profits.[3] Importantly, in these studies, humans remain in charge of the core labor task. The AI tools augment workers' tasks by assisting and providing them with information. 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

[3] Chen and Chan (2024) show that, in the context of designing advertisements, different ways in which AI tools assist humans may lead to positive or negative impacts on outcomes. Similarly, Dell'Acqua et al. (2025) show how an AI tool that assists consultants can increase but also decrease productivity depending on the task.

collect and interpret information generated in a natural language conversation to aid a human in making decisions. Thus, we provide causal evidence from a natural setting on the impact of automating a full production stage with AI agents, both in terms of economic consequences and behavioral responses from humans.[4]

Second, we contribute to the literature on human responses to information provided by AI. Previous 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. We add to this literature a setting in which the AI provides quality signals that human recruiters must weigh alongside additional information to make hiring decisions.

Third, with our hiring setting, we contribute to the literature on the use of AI and algorithms more generally in labor markets. A large body of literature has studied the impact of technology that helps recruiters in screening applicant information prior to job interviews or assists in evaluating 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). 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 articles study the stages before or after the interview and keep the interview under human control. In our setting, the AI is used for the interview itself, which is one of the most labor-intensive parts of the hiring process. Because we randomize AI use 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).[5]

[4] Recruiters usually perform two expert tasks: conducting job interviews and evaluating applicants. Our evidence on automating interviews 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; David and Thompson, 2025), in this case by redirecting recruiter expertise toward evaluation and potentially raising standards in low-entry job assessments.

[5] For instance, human recruiters' capabilities of detecting talent are found to be lower in unstructured compared to structured interviews (see, e.g., for a meta-analysis McDaniel et al., 1994), and in the presence of rules-based compared to discretionary hiring (Estrada, 2019). They are also subject to behavioral biases,

Lastly, we contribute to a growing set of behavioral research that investigates how humans and, in particular, workers and managers in labor settings perceive, trust, and interact with AI (Fumagalli et al., 2022; Dargnies et al., 2024). Recent work on AI in persuasion finds that AI agents are less effective than humans in debt collection calls (Choi et al., 2025), while AI agents can change beliefs in conspiracy theories (Costello et al., 2024). We complement this literature by providing evidence on human-AI interactions when the interaction takes place in a field setting with real-world economic consequences. Our behavioral data allow us to assess the human response to AI introduction in terms of revealed preferences, procedural trust, perceived discrimination, and conversational quality, as well as to study heterogeneous sorting into AI versus human interviews.

The remainder of the paper proceeds as follows. In Section 2, we describe the technical background of the AI voice agent, the market setting, and the firm's recruitment workflow. Section 3 describes the experimental design and sample. Section 4 presents the main recruitment outcomes and 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

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

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.

see for example Kausel et al. (2016) and Radbruch and Schiprowski (2025).

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, such as books, websites, or articles, and they learn 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 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, where appropriate.

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 recruitment 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).[6] 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.

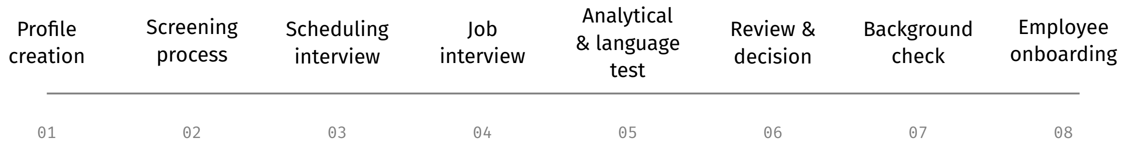
Industry background. The customer service industry in the Philippines is a major sector, estimated to employ more than 1.5 million workers (Hernandez, 2024). The Philippines has become the world's leading provider of call center support, with a large proportion working to assist customers in the US. 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 extremely 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 employees (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

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

Figure 1: Recruitment process



several recruiting firms simultaneously. Interviews are scheduled as soon as possible and hiring decisions are also made quickly. The large pool of applications means 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.

2.3 Recruitment process

To remain competitive, 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 recruitment site of the firm (*Walk-in* mode). Once the firm receives an application, a profile is created in the recruitment software, and the formal recruitment process begins. Figure 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 particular, for applicants in the *Walk-in* mode, all eligible applicants are invited to an interview after expressing their interest in person. 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, with more details yielding a higher score. Applicants whose scores are below a certain threshold are screened out.[7] The rest are invited to an interview. Interview invitations are sent by telephone, text messages, and email. See Appendix Section D.2 for the content of the invitations. Applicants are quasi-randomly assigned to recruiters using a Round-robin scheduling algorithm (Silberschatz et al., 2018).

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

Job interviews. Interviews are conducted in two modes. In the *Walk-in* mode, interviews take place in person, while in the *Remote* mode, recruiters call the applicants. In the latter, 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 E.1 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 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.

Standardized analytical and language 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, including sentence completion, error identification, synonym construction, reading comprehension, and providing contextual vocabulary. Scores are based on the CEFR framework, i.e., range in 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. These parts assess how quickly and accurately applicants can spot errors from strings or images, how well they judge logical statements, and how well they perform in solving math and data-interpretation puzzles, respectively. 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. Once an applicant has completed both interview and standardized test, a recruiter reviews (i) the interview transcript and recording, (ii) the audio recording, and (iii) the performance in the standardized test. Based on the performance of the interview and test, the recruiter then decides whether to extend a job offer to the applicant or not. 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 also provide a short justification in an open-ended text format. For details on the scoring, see Appendix Section E.2. Based on the interview and test scores, a recruiter then decides whether an applicant is suitable to be hired or not. As such, hiring decisions are essentially threshold-based. While there exist some general monthly hiring targets for each application site, the fact that there is a continuous supply of applicants and demand from the client companies means each recruiter decides whether an applicant is above or below a quality bar. If a recruiter judges an applicant to be suitable for the job, they assess whether their current 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. After training, they start their regular work in their new job. A key challenge at this step is that usually a substantial fraction of applicants accept the offer but do not show up for the training period, are absent during the training period, or do not pass the additional validation checks employed by the client company.

3 Experimental design

Our experiment aims to test the causal impacts 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*.

3.1 Treatments

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. During the interview, recruiters follow structured interview guidelines as we described in more detail in the previous section.

AI Interviewer. In the *AI Interviewer* condition, applicants are instead interviewed by an AI voice agent. The agent follows a conversation pathway that mimics the same questions and follow-up strategies used by human interviewers.[8] 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 nearest application center of the firm. That is, in both modes, the 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. If they do not choose within a certain time frame, the AI voice agent calls them. The AI agent explains the process, provides them with insights into how the AI agent works, and then asks them to make the choice.[9]

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 makes a hiring decision. 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. Naturally, 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 recruiters themselves ask the applicants assigned to the *Choice of Interviewer* condition to make a choice, they know if the candidates self-

[8] For a detailed discussion on the technical specifications of the AI voice agent, see Appendix Section I.

[9] 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.

selected for an interview with the AI voice agent or with a human recruiter. Importantly, all recruiters evaluate applications in all conditions and are instructed to apply the same assessment criteria for interviews and hiring, regardless of the treatment or the interview mode.

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

3.2 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 43 different client accounts, which operate in the technology, insurance, telecommunications, retail, finance, healthcare, and transportation. 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.[10] 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* treatment (10,421 in *Walk-in*, 29,682 in *Remote*), 13,557 applications (20.22%) were randomized into the *Human Interviewer* treatment (3,478 in *Walk-in*, 10,079 in *Remote*), and 13,396 (19.98%) into the *Choice of Interviewer* treatment (3,469 in *Walk-in*, 9,927 in *Remote*).[11] In Appendix Table B.1, 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 4,160 applications with employee data.

Applicant characteristics. Of all applications, 64,556 were submitted by unique individuals, which means that 8% of the applicants submitted more than one application during the experiment. The majority of 58% of applicants learn about the job through digital job postings, 19% of the applications happen through referrals, and the rest through other

[10] See Appendix Figure A.1 for information on the regions and the distribution of applications per region.

[11] The weights were imposed by the firm.

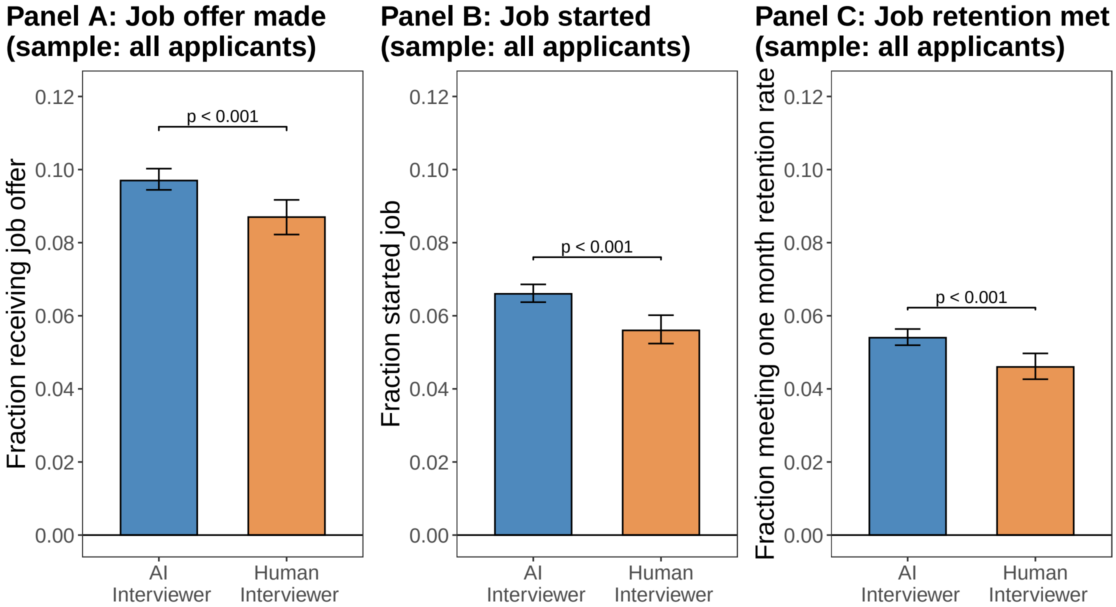
sources such as word of mouth. In total, 60% of applicants were female. Most applicants were between 20 and 30 years old and had some prior experience in customer service jobs.

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, with questions measuring (i) overall interview satisfaction (Net Promoter Score, NPS), (ii) perceived recruiter quality, (iii) perceived interview quality, (iv) perceptions of fairness and discrimination, and (v) opinions, usage, and knowledge of AI. 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). See Appendix G for the instructions for each version. 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 more data on recruiters' beliefs and opinions, the firm conducted a firm-wide survey with recruiters after the experiment concluded. The survey took about five minutes and contained two blocks. One block asked recruiters to predict the results of the experiment and some general attitudes. The other asked them about their experiences evaluating AI-led interviews. The latter was only fielded if they had evaluated AI-led interviews before. See Appendix H for the instructions. The survey was completed by 173 recruiters. Of those, 133 evaluated AI-led interviews, and 98 evaluated applicants who took part in the experiment. These 98 recruiters evaluated 82% of applicants in our experiment.

Figure 2: 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.

4 Results on main outcome variables

In this section, we compare three key outcomes of the recruitment process between the *AI Interviewer* and *Human Interviewer* conditions. As outcomes, we consider the likelihood that an applicant receives a job offer, successfully starts the job, and is employed one month after starting (one-month retention rate). To assess the mechanisms behind our findings, we subsequently analyze the interview transcript data.

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 recruiters expected applicants to receive lower, 49% equal, and 15% higher offer rates. Similarly, 48% expected them to have lower, 39% equal, and 13% higher retention rates, and 61% expected AI-led interviews to be of lower quality.

4.1 Results on hiring decisions and job outcomes

For the analysis, we consider two samples: first, the unconditional sample, where we compare the three outcomes between all applicants that were randomized into either the

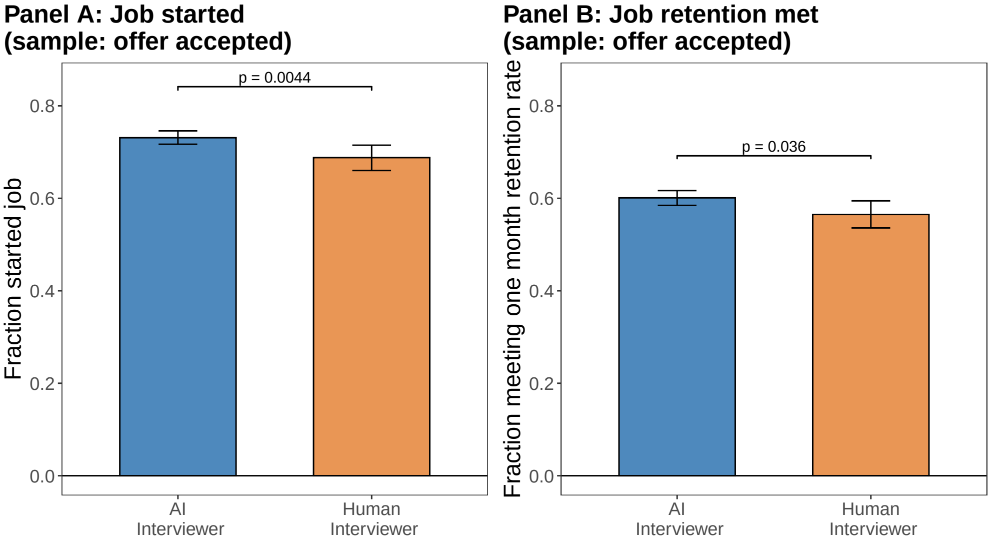
Human Interviewer or *AI Interviewer* condition. Second, the conditional sample, where we condition on applicants who have accepted their job offer.

Sample: all applicants. Figure 2 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 at any conventional level ($p < 0.001$, two-sample proportion test). Moving to the job outcomes, we find that 5.63% (763) and 6.62% (2,653) 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). Lastly, we compare the likelihood that applicants are still working at their job one month after starting. In total, 4.62% of applicants (626) in *Human Interviewer*, and 5.42% of applicants (2,172) in *AI Interviewer* are still working. Hence, applicants who were interviewed by the voice AI have a significantly higher unconditional retention rate ($p < 0.001$, two-sample proportion test)

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 display the results in Figure 3. 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, 68.75% in *Human Interviewer* and 73.14% in the *AI Interviewer* condition successfully start their job and 56.52% and 60.07% are still employed after one month. Accordingly, applicants in *AI Interviewer* who have accepted their job offer have higher fractions of job starters and employees staying for at least one month ($p = 0.004$ and $p = 0.036$, respectively, two-sample proportion test).

Robustness. In Appendix Table B.2, we show that our results are robust to include additional pre-treatment controls and fixed effects. Specifically, we regress our three key recruitment outcomes on treatment status. We then 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 ID of the recruiter who was assigned to the application, the city of the application site that received the application, and the specific job

Figure 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.

posting that the application targeted. Across all three outcomes, whether an application led to an offer, a successful job start, and an employment spell of at least one month, we find very similar treatment effects with and without added controls and fixed effects.

Employment outcomes. Our main job outcome variable is the one-month retention rate, which measures the quality of the employee-employer match. In addition to the retention rate, we also observe, for employees who left their job after starting, whether they left voluntarily (i.e., the leave was 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 job, we find that 58.54% of employees hired from the *Human Interviewer* condition leave voluntarily, while 58.83% of employees hired from the *AI Interviewer* condition do so (thus, 41.55% and 41.17, respectively, leave involuntarily). Hence, there is no difference in leaving reasons between conditions ($p = 1.00$, two-sample proportion test).

4.2 Mechanism: Analyzing interview transcripts

Why do applicants in the *AI Interviewer* condition receive offers with a higher likelihood? In the following, we investigate a key potential explanation: the AI voice agent collects ″better″ information during the interview – in the sense that the collected information reveals more precisely the type of applicants and is therefore more relevant for hiring decisions.

4.2.1 Mechanism empirical analyses: setup

To test the mechanism, we analyze the interview content using data from the interview transcripts.

Interview transcript data. The firm shared with us raw verbatim transcripts. In total, we have transcripts available for 34,109 applications. We used LLM-prompting to convert the raw transcripts into structured data.[12] Afterwards, we used further prompts and text analysis methods to create several transcript-based variables. The first two variables classify and categorize each interview as a whole. We complement these with eight additional variables derived using standard natural language processing methods, which capture the linguistic features of the applicant's responses during the interview.

Number of topics covered. 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. The Appendix Table E.1 displays the full list of possible topics and Appendix Section J.3 the prompt.

Interview type classification. We classify each transcript into one of ten mutually exclusive interview types using a combination of role-based, chain-of-thought, and few-shot promoting (see Appendix Section J.4 for details). *Comprehensive interviews* open and close organically, cover at least eight canonical topics, and are characterized by high applicant engagement. *Disengaged interactions* interviews involve applicants who are unresponsive,

[12] Specifically, we executed two prompts using *gemini-2.0-flash*. First, we use few-shot prompts, i.e., prompts with fully-worked, labeled examples following Brown et al. (2020), to tag each utterance as originating from either the interviewer or the applicant. A second prompt then removed all personally identifiable information from the transcripts with anonymized placeholders (see Appendix Section J.1 and J.2 for the prompts).

distracted, or disinterested, leading to a lower number of topics covered. *Early*, *Midway*, and *Late Screen-outs* are ended by the recruiter in the interview because the applicant does not meet some requirements (e.g., living too far away). *Candidate Unavailability* applies when the interview ends because the applicant is unavailable for the interview. *Telephony Failures* are caused by general telephony problems on either side, while *AI-system Failure*, specific to AI-led interviews, occurs when the AI voice agent crashes. *AI-Aversion* denotes cases where the applicant during the interview explicitly expresses unwillingness to continue speaking with the AI voice agent. See Appendix Table E.3 for a more detailed summary of each type.

Linguistic features. Using standard natural language processing methods, we constructed eight variables capturing key linguistic features of applicants' responses in each interview. Specifically, we measure the (1) vocabulary richness, (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 interviewer and applicant and (7) the number of questions posed by the applicant. Finally, (8) we construct an index of linguistic style matching between interviewer and applicant.

4.2.2 Mechanism empirical analyses: results

Does interview content matter for offer decisions? We start our analysis by assessing the relevance of job interviews for offer decisions. Anecdotally, the firm describes job interviews as a crucial step in the hiring process, with recruiters placing a large 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 to test scores in determining offers, 33% of recruiters say that interview performance is more important, and 63% say they are equally important. Only 4% see test scores as 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 four variables that capture or assess interview content to predict offer decisions. The first two are directly from the recruiters. These are (i) the numeric score (1,2,3) with which recruiters rate applicants' interview performance and (ii) the open-ended text assessment with which recruiters describe the applicants' performance. For open-ended text, we use natural language processing to classify

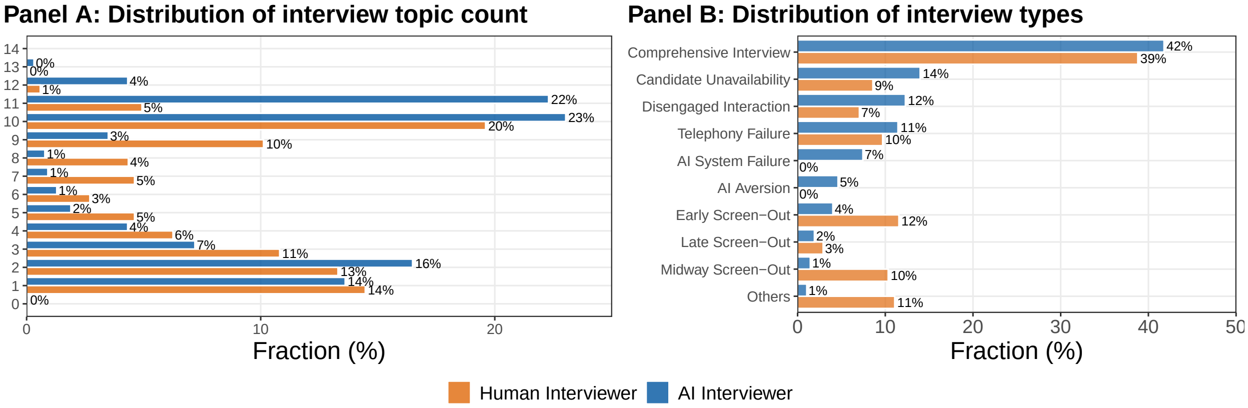
Table 1: Predicting hiring decisions using interview variables

	Dependent variable: Received job offer			
Panel A: Assessment variables				
	(1)	(2)	(3)	(4)
Interview score by recruiter	0.249***	0.147***		
	(0.008)	(0.018)		
Sentiment of interview text assessment by recruiter			0.194***	0.095***
			(0.009)	(0.015)
Controls and fixed effects	No	Yes	No	Yes
Test scores	No	Yes	No	Yes
Observations	4,661	2,477	3,430	1,880
R^2	0.179	0.237	0.139	0.248
Panel B: Transcript variables				
Interview is comprehensive	0.338***	0.095**		
	(0.020)	(0.047)		
Number of topics covered			0.036***	0.007
			(0.002)	(0.007)
Controls and fixed effects	No	Yes	No	Yes
Test scores	No	Yes	No	Yes
Observations	1,891	667	1,891	667
R^2	0.188	0.245	0.125	0.242

Notes: The table shows OLS estimates analyzing the predictive power of interview variables on job offer decisions. 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. "Interview is comprehensive" is an indicator equal to one if the interview opened and closed organically and covered at least eight canonical topics. "Topics covered in interview" is a variable counting the number of topics that were covered in the interview. 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: $^{*}\ p < 0.1$,$^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

text assessments according to their sentiments (coded as follows: negative sentiment = -1, neutral = 0, positive = 1). The latter two variables are inferred variables from the inter-

Figure 4: Transcript variables per treatment



Notes: Panel A displays the distribution of the number of interview topics per interview split by treatment. For the full list of topics and their definitions, see Appendix Table E.1. Panel B displays the fraction of interview transcripts that fall in each of a number of interview types. Appendix Table E.3 provides the details about each interview type's definition.

view transcripts, specifically (iii) an indicator of whether the interview is categorized as an *comprehensive interview*, and (iv) the number of topics covered.

Table 1 provides the results. In Panel A, we regress offer decisions on interview score in column (1) and on text sentiment in column (3). We find that both strongly predict offer decisions. This holds even when we include controls and fixed effects and, most importantly, applicants' test scores (columns (2) and (3)). Accordingly, offer decisions are not only determined by applicants' performance in the analytical and language tests, but are also significantly predicted by interview performance as assessed by the recruiters. In Panel B of Table 1, we focus on the content variables inferred from the interview transcripts. We find that both whether an interview is comprehensive and the number of topics covered predict offer decisions, although the topic variable loses significance once we include controls, fixed effects, and applicants' test scores.

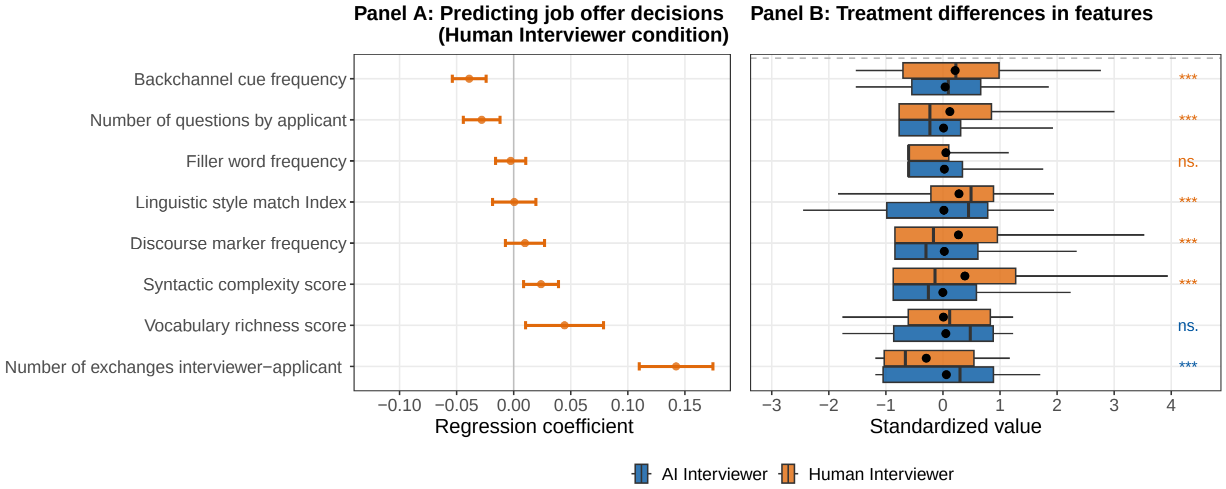
Treatment differences in topic counts. Having established that interview content matters for hiring decisions, we turn to analyzing differences in interview content between treatment conditions. We start with topic counts, with results displayed in Panel A of Figure 4. We find that AI-led interviews are much more likely to cover a very high number of topics (≥ 10) than human-led interviews: 50% versus 25%, respectively. On average, AI-led interviews cover significantly more topics, both in terms of average (6.78 compared to 5.53, $p < 0.001$, two-sample t-test) and median (9 compared to 5, $p < 0.001$, two-sample Wilcoxon signed rank test). The difference in the overall distributions is also highly significant ($p < 0.001$, Kolmogorov–Smirnov test).

Treatment differences in interview type. In Panel B of Figure 4, we plot the distribution of the interview types. In both treatments, the largest proportion of interviews are comprehensive interviews. Importantly, the interviews conducted by the AI voice agent are slightly more likely to be comprehensive (42% compared to 39%). This difference is significant at the 1% level and robust to the inclusion of controls and fixed effects (see Appendix Table B.4). We also see a strong difference in the percentage of interviews that are categorized as *Screen-outs*, which occur much more frequently for human-led interviews. With respect to the two AI-specific interview types *AI Aversion* and *AI System Failure*, we find that the former occurs in 5% of the cases, while the latter case of AI-specific technical failure occurs in 7% of cases.

Treatment differences in linguistic features. Next, we zoom in on the linguistic features. Figure 5 displays the results. As before, we first validate which features matter for job offer decisions. We run a joint regression of job offer decisions on all linguistic variables within the *Human Interviewer* condition. Results are displayed in Panel A. 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. Negative predictors are the frequency with which applicants use backchannel cues and the number of questions they pose. In Panel B of Figure 5, we then compare how the features differ between treatments. Overall, average scores significantly differ for six of the eight features, and we reject equality of distributions in every instance ($p < 0.01$, Kolmogorov–Smirnov test, Bonferroni corrected). Importantly, we find that features are higher in the *AI Interviewer* condition for those features that positively predict job offers, and higher in the *Human Interviewer* condition for those that negatively predict job offers or which have no significant predictive power. These patterns indicate that AI-led interviews elicit more of the linguistic behaviors that recruiters reward (conversational exchange) while reducing those linked to lower hiring odds, such as backchannel cues and applicant-posed questions. This alignment potentially reflects AI's consistent prompting and turn-taking, which may encourage fuller responses and reduce the need for clarifications.

These results indicate that the AI voice agent conducts interviews differently from human recruiters in ways that align with higher offer rates. Recruiters value comprehensive interviews that cover a high number of topics and contain certain linguistic features. AI-led interviews are more likely to meet these criteria, as they produce more comprehensive conversations with broader topic coverage and higher scores on linguistic features that, in human-led interviews, predict higher job offers.

Figure 5: 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.4. Panel B displays the distribution of the variables split by treatment using box plots. Each box represents the interquartile range (25th to 75th percentile), 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 look at the behavior of the applicant in more detail. We investigate their behavior in two stages of the recruitment process. After the interview, we analyze their choices on whether to accept job offers and their responses to the customer experience survey. Before the interview, we examine applicants' choices in the *Choice of interviewer* condition.

5.1 Applicants' response to job offers

Acceptance of job offers. Applicants may react differently to the AI voice agent by not accepting a job offer. This could be caused by them updating negatively or positively about the quality of the job after experiencing the AI voice agent. 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 ($p = 0.086$, two-sample proportion test). In *Chosen Human Interviewer*, the acceptance rate is 92.73%, while in *Chosen AI Interviewer*, it is 92.57%. We find no differences in acceptance rates between the Direct and Choice conditions (Human: $p = 0.55$, AI: $p = 0.67$, two-sample test of proportions) or between applicants choosing different interviewers within the choice condition ($p = 0.92$, two-sample test of proportions). Overall, among those applicants who receive a job offer, there appears to be no negative reaction to the AI voice agent in the form of higher refusals of offers.

5.2 Applicants' interview experience

We complement the behavioral variables of the applicants with our survey. Our survey allows us to more closely measure applicants' attitudes and perceptions of their interview experience.

Net promoter score. We start by comparing how applicants rate their propensity to recommend the recruiting firm to a friend, the central feedback metric that the firm and the industry in general use. We find that the average rating of the applicants in the *AI Interviewer* condition is 8.97 on a scale of 1 to 10, while it is 8.84 in the *Human Interviewer* condition. Accordingly, AI voice agents lead to a small and insignificant increase in score ($p = 0.25$, t-test).

Perceived recruiter quality. Turning to how applicants assess the quality of their recruiter, we find no differences between the *AI Interviewer* and *Human Interviewer* treatments regarding how applicants assess recruiters' knowledge about the firm ($p = 0.58$, t-test) and role ($p = 0.11$, t-test), or the degree to which they found their time to be valued by the recruiter ($p = 0.24$, t-test). We only find a small difference in the perceived relevance of the questions, where the applicants rate the AI voice agent as asking slightly more relevant questions ($p = 0.044$, t-test). Our index of perceived recruiter quality that aggregates these individual items similarly does not differ between treatments ($p = 0.92$, t-test). Hence, applicants rate the quality of the recruiter similarly.

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 induces our index of perceived interview quality to be higher for human-led interviews compared to 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 is capable of delivering an interview quality 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 almost halves with the AI voice agent, although note that these reports form a relatively small sample (62 out of 1818 respondents in *AI Interviewer*, and 22 out of 346 respondents in *Human Interviewer*), as is common with data on discrimination.

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 categorize whether applicants indicate in their text a negative, positive, or neutral interview experience.[13] 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 the responses have a 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.8. 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 or questions generally, or with the AI voice agent specifically. In *Human Interviewer*, a total of 13% mention problems. For the full results, see the Appendix Table A.5. The categorization results are therefore in line with the sentiment analyses and show positive responses to AI-led interviewers. However, since there might be selection into who provides an open-ended response, these results should be interpretative with care.

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

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

AI voice agent is rated as less natural, but also has fewer applicants who report gender discrimination and 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 and recruiters do not have access to the survey, and this was communicated to applicants in the survey, applicants may still believe that they can influence the process with their responses. For example, this may lead them to provide overly positive answers. Although such a misreporting would bias the level of survey responses, it is less likely to influence our relative treatment comparison as it would require an interaction of misreporting and treatment. To empirically assess the extent of the bias, we randomly varied whether the applicants received the survey invitation directly after their interview or after a final decision had been made on their application. This allows us to test whether applicants' responses differ. We find no evidence that applicants change their responses strategically, as we cannot reject the null hypothesis of equal means for our survey variables. See Appendix Table B.9 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 recruiter. 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 AI voice agent. The remaining applicants did not respond to the interview invitation sent by text, nor picked up the follow-up call. Of those making a choice, 7,885 (81.63%) chose the AI voice agent as interviewer. Accordingly, in both modes, a majority of applicants prefer to interview with an AI voice agent instead of a human recruiter. This pattern is relatively stable throughout the experiment, as Figure A.4 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 applicants' preferred time, including right after receiving the interview invitation. Indeed, as we show in Section 7.1, AI-led interviews take place much faster than human-led interviews. A second reason may be applicants' attitudes towards AI. Using our survey evidence reveals that our sample generally perceives AI to have a major impact on the workplace and that the 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.[14] We obtain similar results when looking at how respondents generally assess the impact of AI on workers.

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 the detailed regression results, see Appendix Table B.3.[15]

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 choose to receive the AI interviewer, while lower-quality applicants choose the human interviewer. Second, negative AI sorting means the reverse pattern. Third, applicants' choice and quality might not be correlated.

We start by analyzing the association between applicants' choices in *Choice of Interviewer* and their test scores. Applicants who chose the AI voice agent scored, on average, a 3.14 score (out of 6) in the language test and a 47.54 score (out of 100) in the analytical test. In contrast, applicants who chose the human interviewer scored higher test results of 3.37 and 49.77 on the language and analytical test, respectively. These differences are statistically significant ($p < 0.001$ and $p = 0.001$, respectively, t-tests). The differences are robust to the inclusion of controls and fixed effects; for details, see Appendix Table B.5. 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.2.

As a next step, we analyze whether there exist sorting patterns conditional on receiv-

[14] Perhaps unsurprisingly, recruiters are more pessimistic about the impact of AI on them. We asked the same items in the recruiter survey and found that 68% of recruiters believe that AI will have a major impact on them personally, and only 12% believe the impact to be generally positive.

[15] 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).

ing 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 confound of the previous analyses. It could be that experiencing the AI voice agent relative to the human recruiter affects the performance in the tests. This could be a direct influence, e.g., by applicants increasing or decreasing their effort after experiencing the AI voice agent. More plausibly, the influence could come through differential attrition, as applicants are more or less likely to drop out of the recruitment process before taking the test after completing the interview with the AI voice agent.

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 who get it assigned, while applicants who choose the human interviewer have higher test scores. Still, three out of the four differences are no longer significant after including controls and fixed effects. Accordingly, the sorting effects are slightly weaker when comparing performance in the choice condition to the assigned conditions. For details, see Appendix Table B.6.

Taken together, our results suggest the presence of negative AI sorting.

6 Recruiter response to voice AI interviews

In this section, we analyze the behavior of recruiters.

6.1 Predicting recruiters' offer decisions

Here, we analyze how recruiters score interviews and use the available signals from the interview and the intended test scores. We base our analysis on the *Walk-in* mode, because in the *Remote* mode, it was not mandatory for recruiters to log their interview scores into the system. In Appendix Section F, we show that the results in this section are generally robust when we consider the full sample.

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.[16] For a subset of 10,779 applications, we also observe their justification

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

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.3, we plot the distribution of scores across the two conditions. We find that the higher scores in *AI Interviewer* exclusively come from a higher frequency of a score of 2 instead of a 1 score, while the frequency of 3 scores 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 AI voice agent, consistent with such interviews receiving higher interview scores.

6.1.1 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

Table 2: Predicting job offer decisions of recruiters

	Dependent variable: *Job Offer Made*	
	(1)	(2)
Interview score (std.)	0.091*** (0.010)	0.100*** (0.010)
Language test score (std.)	0.108*** (0.010)	0.115*** (0.010)
Analytical test score (std.)	0.034*** (0.011)	0.028*** (0.010)
Interview score (std.) × *AI Interviewer*	−0.047*** (0.012)	−0.029** (0.012)
Language test score (std.) × *AI Interviewer*	0.028** (0.012)	0.022* (0.011)
Analytical test score (std.) × *AI Interviewer*	−0.003 (0.012)	0.001 (0.012)
Mean DV in Human Interviewer	0.38	0.38
Controls and fixed effects	No	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: * $p < 0.1$,** $p < 0.05$, *** $p < 0.01$.

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 applicant 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.

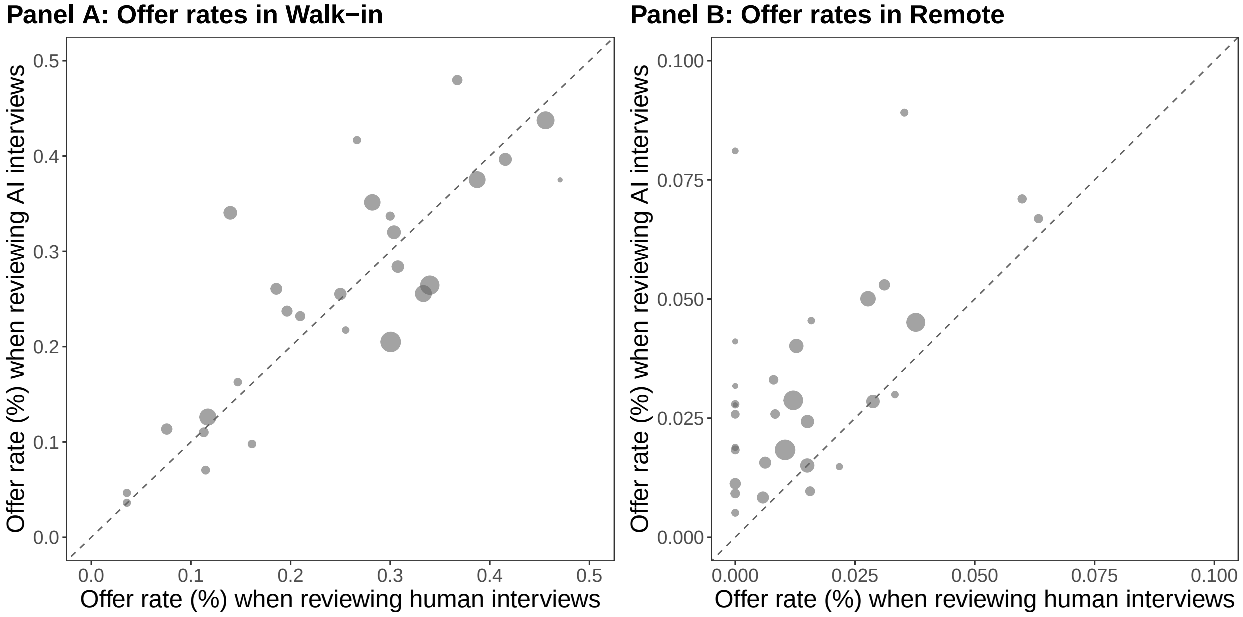
Importantly, we find a significant interaction when comparing the influence of each signal in *Human Interviewer* relative to *AI Interviewer*. The language score is significantly more predictive of offers in *AI Interviewer*, while the interview score is significantly less predictive. This suggests 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 much higher (and only statistically significant) for recruiters who consider the interview as more important than the standardized scores. For details, see Appendix Table B.7. This suggests that recruiters indeed weigh the signals from AI-led interviews differently.

6.2 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 6: 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 6, 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.10 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.10). 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 report how the adoption of the AI voice agent impacts time-to-hire by shifting the queue problem from scheduling and interview completion to the evaluation stage. Second, we explore how the costs of implementing AI compare to the costs of human recruiters.

7.1 The impact of AI introduction on recruitment process length

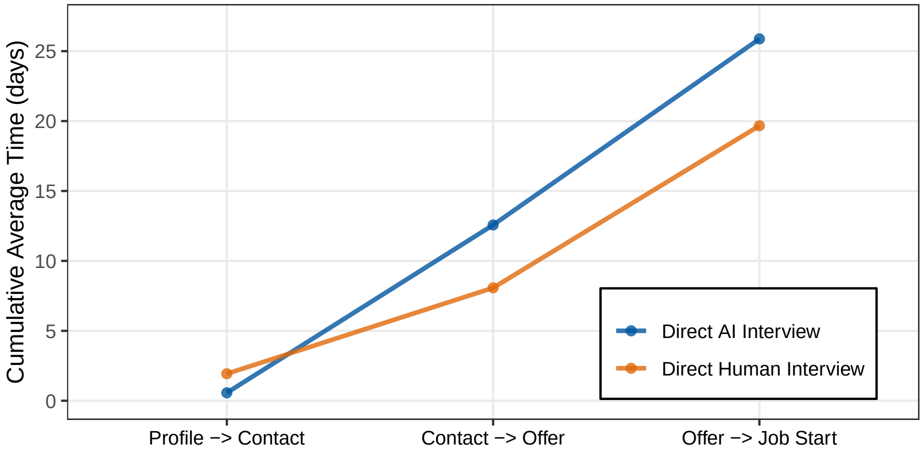
Setup. The firm tries to reach out to applicants as fast as possible after they submit an expression of interest to minimize the chance that applicants go to competing companies. Accordingly, the time from expression of interest to interview is a key factor influencing organizational efficiency in this setting. The second key factor is the time between the interview and the offer decision. Lastly, less important for the recruiting firm but relevant to the client companies is the time from the offer decision to when applicants start their job. Introducing the AI voice agent plausibly influences the first two variables: as the agent is available 24/7, making scheduling much easier for applicants. However, after the interviews, the human recruiters may take longer to review an applicant, given that the recruiters have not had the interview experience with them. The third variable is plausibly less likely to be influenced by the introduction of the AI voice agent, as the process is standardized and conducted independently of interviews. Any differences here can only reflect differences in applicant characteristics, e.g., how quickly they hand in information.

We focus on applicants in the *Remote* mode, and start by looking at the time of successful hires, that is, applicants who have started their job. For them, we analyze time differences at three points during the recruitment process. First, the time from when an applicant submits an expression of interest to when an attempt to conduct an interview is made. Second, the time from the first interview contact to the offer decision. Third, from the offer decision to when the applicants start their job.

Results. Figure 7 displays the results. We find that the median time (average time) from profile creation to interview is 0.32 days (0.56) in *AI Interviewer* and 0.51 (1.93) in *Human Interviewer*. Both the difference in medians ($p < 0.001$, two-sample Wilcoxon signed-rank test) and averages are significant ($p = 0.093$, two-sample t-test), albeit the latter one only at the 10% level. From the first to the second stage, applications in *AI Interviewer* take a median (average) time of 6.69 days (11.46), while in *Human Interviewer*, the time is 2.58 (6.14). Accordingly, the time is much smaller in *Human Interviewer* compared to *AI Interviewer* ($p = 0.011$, two-sample Wilcoxon signed-rank test; $p = 0.004$, two-sample t-test). Lastly, regarding the time from the offer stage to date that applicants start their job, this takes AI-led interview applicants a median (average) time of 12.00 (12.69) days, while for Human-led interview applicants it takes 12.00 (11.59) days. Hence, we find no difference in speed across both conditions ($p = 0.70$, two-sample Wilcoxon signed-rank test; $p = 0.50$, two-sample t-test). This result is expected, as the process after receiving a job offer does not depend on who conducted the interview.

When looking at the overall time from profile creation to job starting, we find that *AI*

Figure 7: 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. "Profile → Contact" denotes the time it takes from an applicant's expression of interest to an interview taking place. "Contact → Offer" denotes the time it takes from an interview taking place to a recruiter making an offer decision. "Offer → Job Start" denotes the time it takes from a recruiter making an offer decision to the applicant starting their job.

Interviewer applicants take a median (average) time of 22.00 days (24.71), while in *Human Interviewer*, the time is 19.00 (19.67). These differences are marginally significant ($p = 0.062$, two-sample Wilcoxon signed-rank test; $p = 0.052$, two-sample t-test). Accordingly, the time saved by employing AI voice agents in interviews is offset by human recruiters taking longer to review them, leading to a slightly longer overall recruitment process.[17]

7.2 When do voice AI interviews pay off? Stylized estimates

We now use approximate estimates of our partner firm's cost structure to compare the cost efficiency of AI-led and human-led interviews in two stylized environments: (i) *static*, in which the accuracy and costs per interview are fixed; and (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).

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

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

[17] This result reflects a classic queue optimization problem (Hassin and Haviv, 2003), now shifted to AI-led interview, where time saved upfront must be balanced against longer downstream evaluations by humans.

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 communicated estimate per interview, ensuring that our cost figures do not overstate the expense of human-led interviews.[18] The mid- and high-cost levels ($3.50 and $6.37, respectively) scale the baseline to reflect the middle- and high-wage market environments observed in the firm's global operations. As our experiment occurred in the Philippines, the human costs in our setting correspond to the c_H^L calibration.

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

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

These calibrations are derived from the publicly available API pricing of ElevenLabs, an AI vendor comparable to the one that our firm partner relies on, which we cannot disclose. In August 2025, AI list rates as low as $0.08 per minute (annual billing).[19] 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.[20] In all scenarios, to be even more con-

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

[19] 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.

[20] 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

servative, we also include a one-time fixed deployment cost of $F = \$10,000$ paid to the AI vendor.[21]

Static 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 3, 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 instance, 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.

Dynamic 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 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.[22] The actual launch month rate was far below 25%, hence this choice inflates early periods, making our break-even estimates conservative.

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.

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

[22] This value comes from our classification of the proportion of interviews labeled as "AI System Failure" in Figure 4, based on transcript-level completion classifications covering our preregistered experimental period between March 7 and June 7, 2025.

Table 3: Static and one-year dynamic break-even counts under three AI prices and wage environments

Panel A: Static cost gaps and break-even interviews

	Low AI $1.30		Mid AI $2.06		High AI $3.03	
Market	Gap ($)	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)

	Low AI $1.30		Mid AI $2.06		High AI $3.03	
Market	Gap ($)	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).

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

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

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

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

which treats the year as if all interviews occurred in the month 12 gap.[23]

As shown in panel B of Table 3, 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.

8 Conclusion

In this paper, we examine the consequences of substituting human interviewers with AI voice agents. Our large-scale natural field experiment provides empirical evidence on the capabilities but also challenges of deploying AI voice agents. We find that AI-led interviews increase offer rates and yield comparable or even improved employee-employer match quality relative to human-led interviews. At the same time, we document nuanced behavioral responses: recruiters adjust their workflow and how they interpret applicant information from AI-led interviews. Besides, while most applicants are comfortable interacting with AI, even preferring it when given a choice, there is evidence of negative quality sorting among those who choose the AI interviewer. Overall, our results show that AI voice agents can match human recruiters in operating a complex but key task in hiring – conducting job interviews – with early signs of better outcomes in different dimensions and new behavioral and organizational challenges that would need to be tackled by firms that aim to fully maximize the economic promises of generative AI.

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

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*, [6]

Agarwal, Nikhil, Alex Moehring, and Alexander Wolitzky (2025). “Designing Human-AI Collaboration: A Sufficient-Statistic Approach.” *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. [70]

Angelova, Victoria, Will Dobbie, and Crystal Yang (2023). “Algorithmic Recommendations and Human Discretion.” *Working Paper*, [6]

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]

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). “Finding Talent in the Age of AI.” *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*, [19]

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

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]

Choi, James J, Dong Huang, Zhishu Yang, and Qi Zhang (2025). "How Good Is AI at Twisting Arms? Experiments in Debt Collection." *Working Paper*, [7]

Costello, Thomas H., Gordon Pennycook, and David G. Rand (2024). "Durably Reducing Conspiracy Beliefs through Dialogues with AI." *Science* 385 (6714): eadq1814. [7]

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 (2024). "Aversion to Hiring Algorithms: Transparency, Gender Profiling, and Self-Confidence." *Management Science*, mnsc.2022.02774. [7]

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

David, H, and Neil Thompson (2025). "Expertise." *Working Paper*, (w33941): [6]

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]

Estrada, Ricardo (2019). "Rules versus Discretion in Public Service: Teacher Hiring in Mexico." *Journal of Labor Economics* 37 (2): 545–79. [6]

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

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. National Bureau of Economic Research. [6]

Hassin, Refael, and Moshe Haviv (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*. vol. 59, Springer Science & Business Media. [36]

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]

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]

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. [7]

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]

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. [6]

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., 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. [7]

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]

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. [6]

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*, [83]

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*, [83]

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: Distribution of applications by recruitment center location

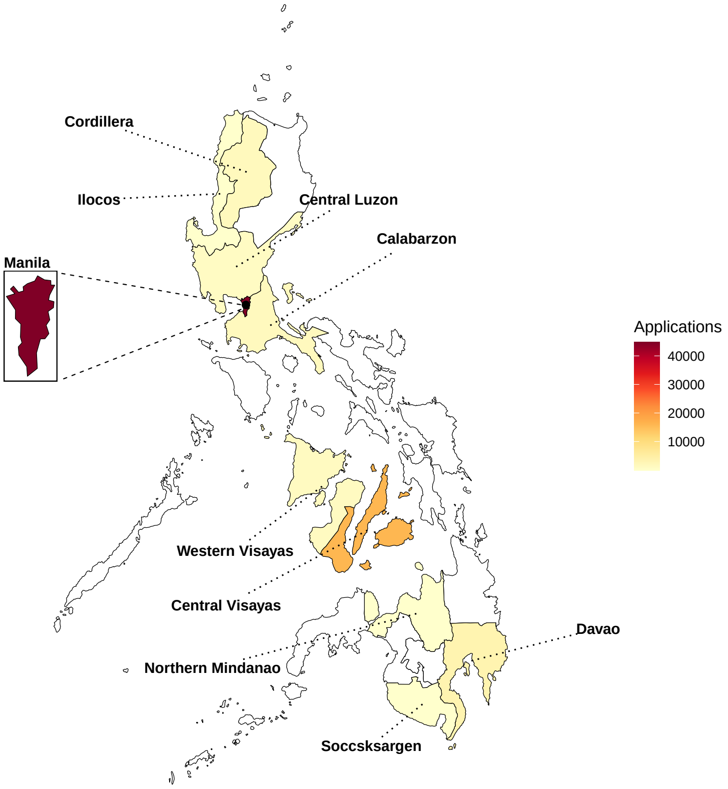
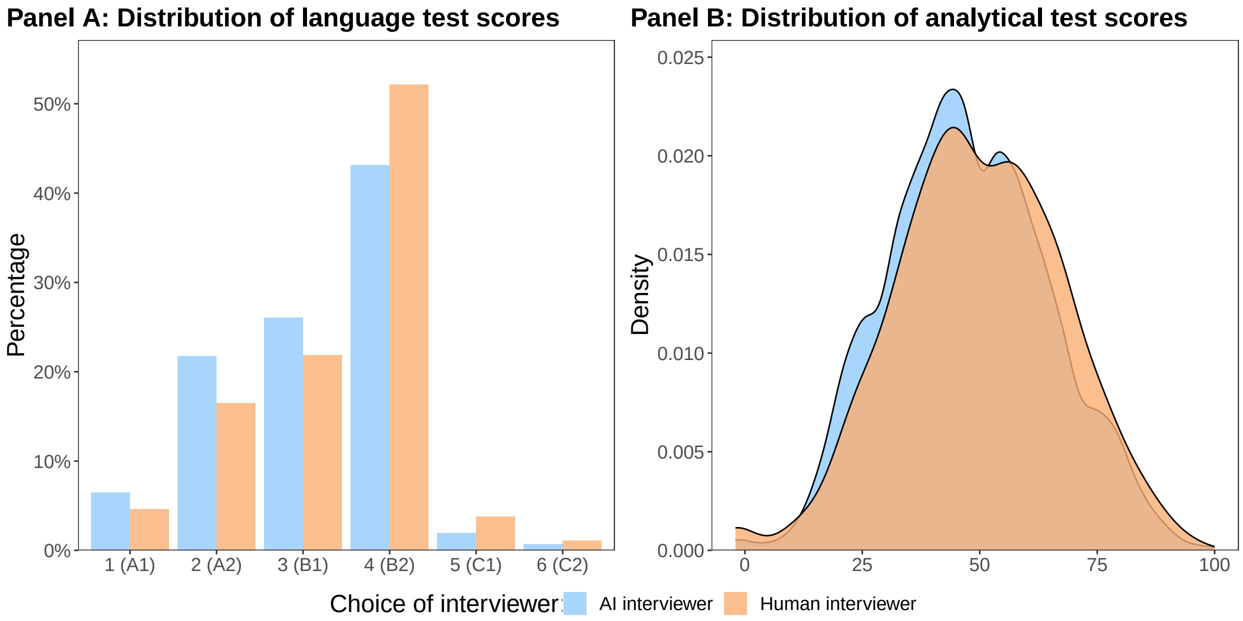
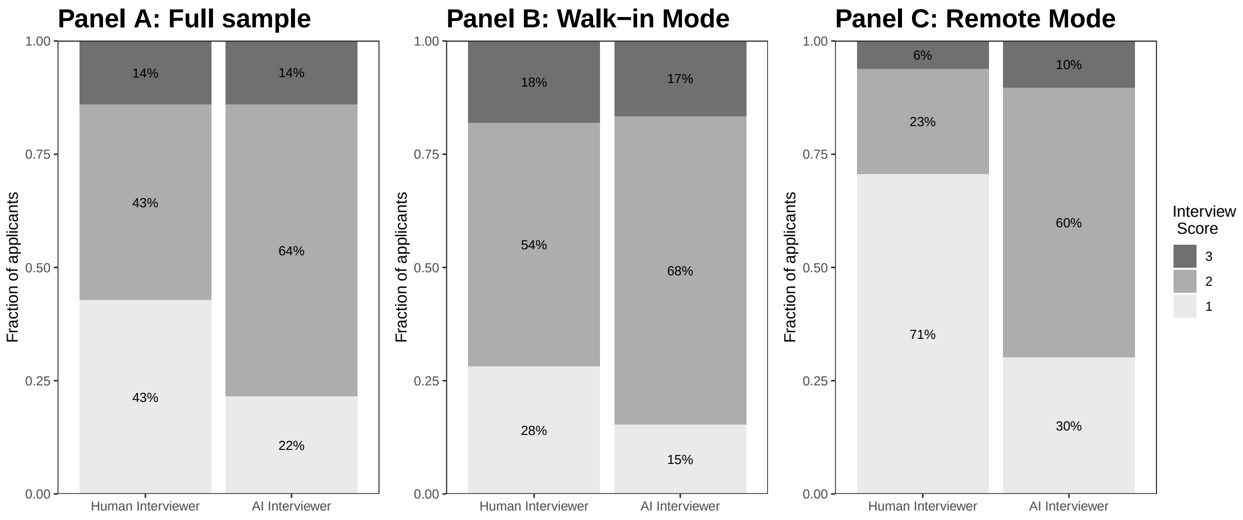


Figure A.2: Distribution of standardized test scores depending on applicants' choice of interviewer



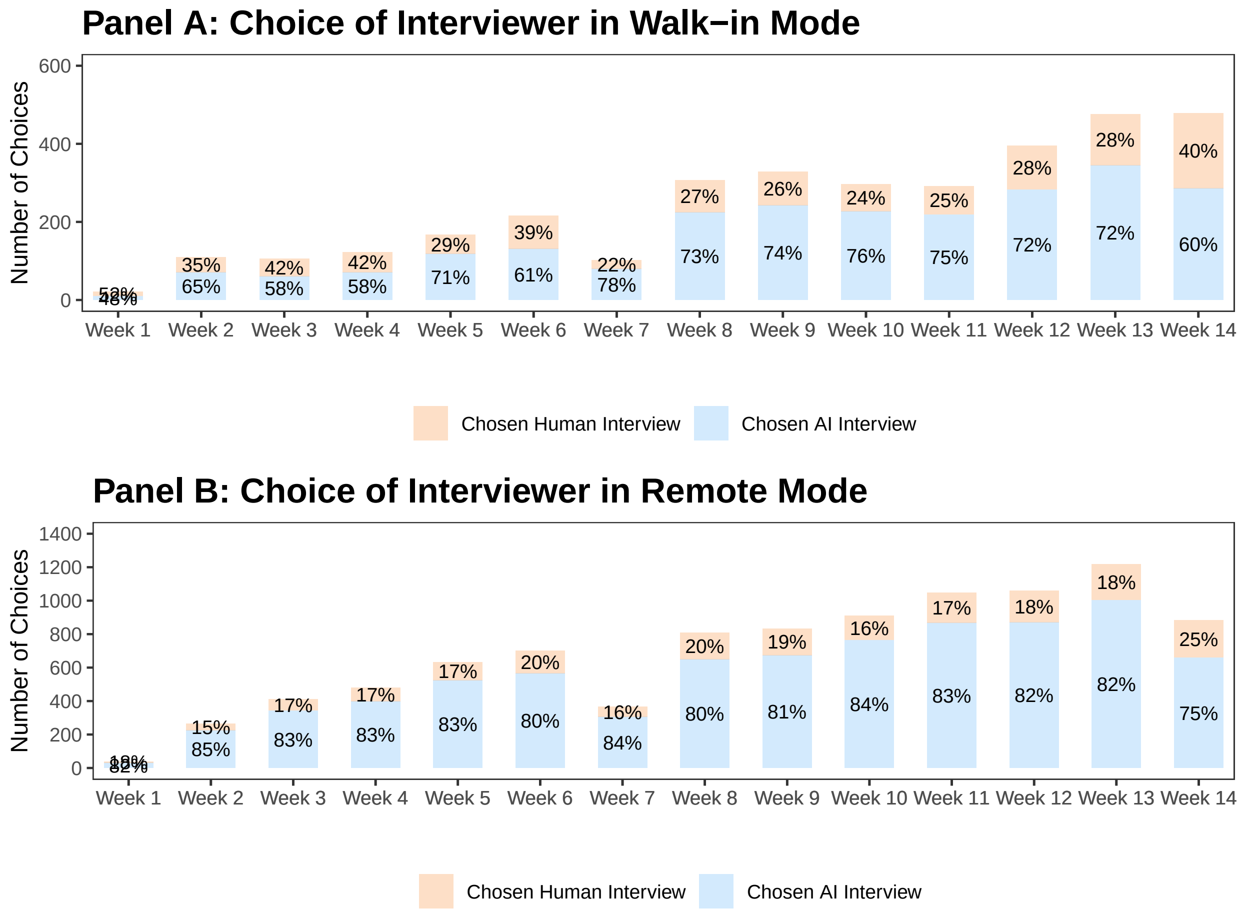
Notes: The figure displays the distribution of test score results of applicants, split by their interviewer choice in the *Choice of interviewer* condition. Panel A displays the results from the language test, which assesses applicants' writing and reading capabilities in English. Scores are based on 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.3: Distribution of interview scores across treatments



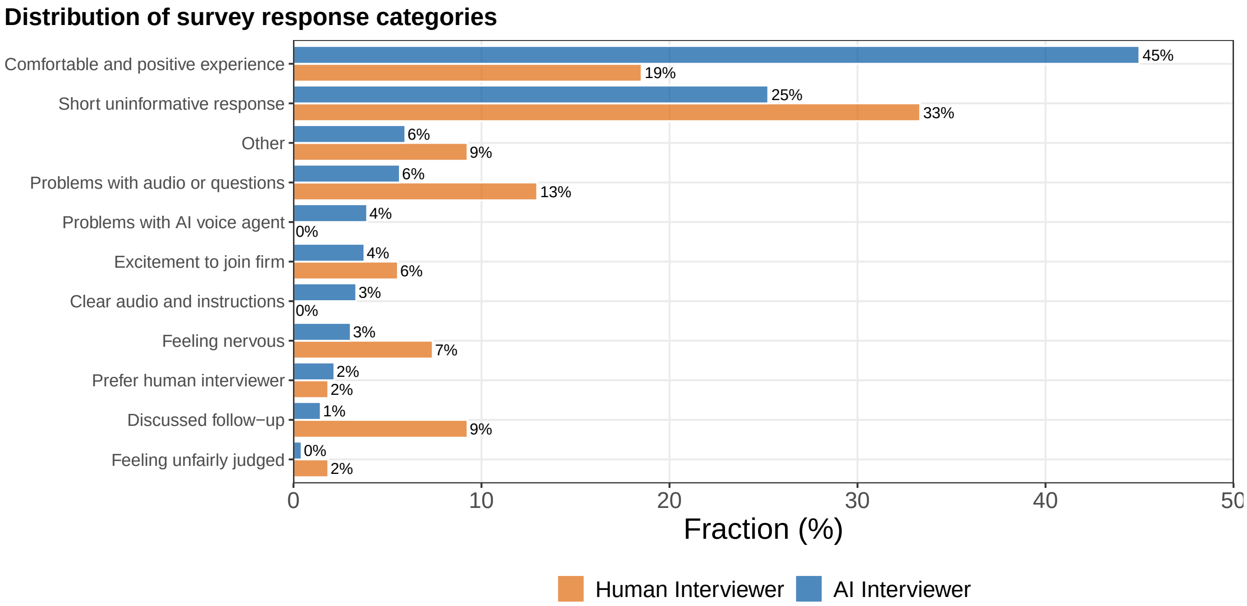
Notes: The figure displays the distribution of the interview score with which recruiters assess each interview. The score is 1-poor, 2-medium, 3-good. For details on the scoring, see Appendix Table E.2. Panel A displays the full sample results. Panel B displays the *Walk-in mode* subsample, in which 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.4: 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.5: 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.8.

B Additional tables

Table B.1: 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	0.18	0.52	0.03
Source is referral	0.19	0.19	0.19	0.90	0.40	0.24
Source is digital ad	0.59	0.59	0.59	0.44	0.84	0.30
Mode is Walk-in	0.26	0.26	0.26	0.43	0.68	0.78
Initial engagement score	37.05	37.10	37.21	0.85	0.60	0.65
Observations	13,561	40,181	13,417			
Panel B: *Walk-in mode*						
Gender (Women=1)	0.56	0.56	0.56	0.97	0.88	0.94
Source is referral	0.12	0.11	0.11	0.55	0.64	0.38
Source is digital ad	0.05	0.06	0.06	0.66	0.79	0.55
Observations	3,484	10,463	3,477			
Panel C: *Remote mode*						
Gender (Women=1)	0.62	0.61	0.62	0.13	0.48	0.02
Source is referral	0.22	0.22	0.21	0.60	0.64	0.27
Source is digital ad	0.77	0.77	0.78	0.60	0.67	0.29
Initial engagement score	50.49	50.42	50.49	0.70	0.99	0.68
Observations	10,077	29,718	9,940			

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.2: Treatment effect on key recruiting outcomes

Panel A: sample all applicants									
	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***	0.0099***	0.0095***	0.0095***	0.0080***	0.0077***	0.0077**
	(0.0028)	(0.0026)	(0.0037)	(0.0023)	(0.0022)	(0.0032)	(0.0021)	(0.0021)	(0.0031)
Mean DV *Human Interviewer*	0.0870	0.0870	0.0870	0.0563	0.0569	0.0569	0.0462	0.0466	0.0466
Controls and fixed effects	No	Yes	Yes	No	Yes	Yes	No	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	0.0003	0.1095	0.1095	0.0002	0.0940	0.0940
Panel B: sample offer accepted									
AI Interviewer				0.0439***	0.0353**	0.0353	0.0355**	0.0380**	0.0380*
				(0.0158)	(0.0156)	(0.0219)	(0.0170)	(0.0169)	(0.0213)
Mean DV *Human Interviewer*				0.6875	0.6953	0.6953	0.5652	0.5708	0.5708
Controls and fixed effects				No	Yes	Yes	No	Yes	Yes
Clustering				App.	App.	Rec.	App.	App.	Rec.
Observations				4,708	4,575	4,575	4,708	4,575	4,575
R^2				0.0017	0.0701	0.0701	0.0009	0.0660	0.0660

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.3: 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	No	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.4: 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	No	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.5: 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	No	Yes	No	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.6: 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	No	Yes	No	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	No	Yes	No	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.7: 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.) × *AI Interviewer*	-0.021	-0.024	-0.086^{***}	-0.059^{*}
	(0.018)	(0.017)	(0.029)	(0.031)
Language test score (std.) × *AI Interviewer*	0.042^{**}	0.036^{**}	0.045	0.038
	(0.017)	(0.016)	(0.031)	(0.032)
Analytical test score (std.) × *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	No	Yes	No	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.8: 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.9: 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.10: 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 × *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)

Engagement-score algorithm. Each applicant receives a base score that depends on how their profile entered the system (self-application vs. recruiter-added).

Signal weighting. 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.

Threshold rule. 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.

Randomization. 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 Interviews details

E.1 Interview structure

Table E.1: Voice AI or 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

E.2 Interview scoring grid

Recruiters - whether they are evaluating their own human-led interview or reviewing an AI-led conversation - score interviewers on a three-point scoring system, displayed in Table E.2.

Table E.2: 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

E.3 Transcript type classification

Table E.3: 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

Table E.4: Overview of interview transcript feature variables

Variable	**Description**
Vocabulary richness score	Number of unique words divided by the square root of the total number of words in applicant's responses. Higher scores indicate greater vocabulary diversity and linguistic sophistication.
Syntactic complexity score	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 more nuance in responses.
Discourse markers frequency	Average number of sequential, causal, and clarifying discourse markers per minute (e.g., "first", "because", "specifically") in applicant's responses. Higher values indicate more frequent use of discourse markers.
Filler words frequency	Average number of basic and conversational fillers per minute (e.g., "uh", "uhm", "like", "you know") in applicant's responses. Higher values indicate more frequent use of filler words.
Backchannel cue frequency	Average number of verbal backchannel cues per minute, i.e., 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	Total number of conversational exchanges, where an exchange is first the interviewer speaks and then the applicant.
Number of questions by applicant	Total number of questions asked by the applicant.
Linguistic style match index	Similarity of linguistic style between applicant and interviewer. The index 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 $1-\frac{\lvert \text{rate}_{\text{interviewer}}-\text{rate}_{\text{applicant}} \rvert}{\text{rate}_{\text{interviewer}}+\text{rate}_{\text{applicant}}+0.0001}$, with rates representing percentage usage in each speaker's text. Higher index values indicate higher linguistic style similarity between interviewer and applicant.

F Robustness analyses recruiter perspective

Table F.1: Predicting job offer decisions of recruiters full sample

	Dependent variable:	
	Job Offer Made	
	(1)	(2)
Interview score (std.)	0.077*** (0.010)	0.088*** (0.010)
Language test score (std.)	0.109*** (0.010)	0.111*** (0.009)
Analytical test score (std.)	0.033*** (0.010)	0.029*** (0.010)
Interview score (std.) × *AI Interviewer*	−0.038*** (0.011)	−0.026** (0.011)
Language test score (std.) × *AI Interviewer*	0.015 (0.011)	0.015 (0.010)
Analytical test score (std.) × *AI Interviewer*	−0.003 (0.011)	0.001 (0.011)
Mean DV in *Human Interviewer*	0.38	0.37
Controls and fixed effects	No	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: * $p < 0.1$,** $p < 0.05$, *** $p < 0.01$.

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 Technical background of the AI voice agent

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. The partner's AI voice agent, facing the candidate, is supplied by a specialized AI-SaaS vendor. It stacks three generative technologies in real time: (i) multilingual automatic speech recognition (ASR); (ii) a job-specific large language model (LLM) and (iii) streaming text-to-speech (TTS). The ASR front end ('Babel') converts accented Filipino-English audio into text; the transcript is routed through an LLM fine-tuned in thousands of proprietary interviews, which selects the next question or follow-up to prompt a human response; the response is rendered in natural speech by a neural vocoder and sent back to the candidate. End-to-end latency averages 0.4–1.8s, indistinguishable from human turn-taking, because the entire stack runs on the vendor's dedicated GPU cluster. 'Contextual pathway prompts' inject the relevant job description at every turn, while hard guardrails veto off-topic or non-compliant content. Every utterance, text, and timestamp is archived, resulting in millisecond-level measures of pauses, interruptions, and sentiment that conventional interviews cannot capture.

Operational challenges. Live hiring interviews magnify the risks of the classic language model. A delay above a second breaks conversational flow; ASR slips on noisy mobile lines derail question logic; hallucinations expose the firm to legal risk; and applicants may try to game the agent by parroting scripted buzzwords. Meeting these constraints required custom acoustic models, controller-layer guardrails, and low-latency GPU inference, deployment investments far beyond an off-the-shelf API call, often only required for using AI as a signal provider.

J Text analysis: prompt content

J.1 Speaker labeling prompt

Speaker labeling prompt

```
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:

Catherine: Hi, this is Catherine 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.

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

Catherine: Okay

Output:

```

```
Interviewer: Hi, this is Catherine 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 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.4 Interview classification prompt

Overall, 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, in-context examples help the model recognize category cues, e.g., the statement “I’m not comfortable talking with AI” signals the AI Aversion category. This approach allows Gemini to leverage in-context learning without altering its underlying parameters. 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

Priority 3 - Interview Analysis
- If none of above categories apply, check for these following categories

8. Disengaged Interaction: The candidate demonstrates disinterest, unresponsive, distracted, and poor continuity. This category applies when the candidate initially engaged in the conversation.
    - Number of topics is strictly less than 8 topics (topic_count < 8)
9. Comprehensive Interview: Natural opening and closure; topic_count at least 8 (>= 8) expected topics or more; high-quality engagement from both parties.
    - The candidate answers questions fully and asks relevant questions.
    - There is a concluding remark. e.g. "We’ll see you on the next hiring process. Good luck!"
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.

Priority 4 - Others
- If none of above categories apply, check for the following category
11. Others: An interview is not in English or does not fit to above categories.

Expected output:
1. Category: [Chosen Category Name]
2. Explanation: [Your concise explanation, citing evidence from the provided data and transcript.]

"""
```

J.5 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 preregistered the experiment at the AEA RCT Registry (trial number #15385, link: https://www.socialscienceregistry.org/trials/15385). The preregistration 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 preregistration.

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 preregistered period in which the experiment took place.

Treatment conditions. Our three experimental conditions were implemented as preregistered. 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, we also focus on applicants who made a choice in our analyses involving the *Choice of interviewer* condition.

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 4.2.2 because we received transcript data.
- Whether the applicant receives a job offer
  - This variable is reported in Section 4.1.
- 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 4.1.

- Retention rate: whether the applicant still works at the firm after starting their job.
  - This variable is reported in Section 4.1.

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 transcript, but not audio files. Interview transcript variables are analyzed in Section 4.2. 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 & 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.

Further variables. Analyses that are based on applicants' standardized test scores were not part of the preregistration. The reason is that at the time the preregistration was written, it was unclear whether the firm would provide data access to the scores. We pre-registered the applicant survey and report its results. The recruiter survey was not pre-registered.