AI and AI-human based Salesforce Hiring using Conversational Interview Videos

Ishita Chakraborty
University of Wisconsin-Madison, ishita.chakraborty@wisc.edu

Khai Chiong
University of Texas-Dallas, khai.chiong@utdallas.edu

Howard Dover
University of Texas-Dallas, howard.dover@utdallas.edu

K.Sudhir
Yale School of Management, k.sudhir@yale.edu*

We develop an AI and AI-human-based model for salesforce hiring using recordings of conversational video interviews that involve two-sided, back-and-forth interactions with messages conveyed through multiple modalities (text, voice, and body language). We extract theory-relevant objective measures of interviewees' sales performance from the text, voice and video modalities as explanatory features in the AI model. Our key contribution to the broader research on persuasion and influence is that we show how to use conversational videos to capture features related to (i) two-way conversational interactivity; (ii) real time adaptation and (iii) human body language, with minimal measurement error relative to extant survey-based approaches that suffer from recall biases. We use rubric-based scores by panels of sales professionals (correlated with hiring decisions) to isolate a candidate's "latent sales ability;" and use these as outcome variables to be predicted by the AI model. The AI model achieves reasonable predictive accuracy, but integrating human input into an AI-Human hybrid model further enhances performance – it improves workforce quality relative to a random benchmark by 67%. While the content of what is spoken is most important in prediction, conversational interactivity, sellers' real-time adaptation to the buyer, and body language also have good explanatory power. Finally, in terms of performance-cost trade-offs, the addition of just one human professional evaluation in the hiring loop in combination with AI is optimal. Further, using human input based on only the two early stages of the interview in a task-based hybrid model is the most cost-effective in improving performance.

Key words: AI-human, Video Analytics, B2B, Salesforce, Machine Learning

---

*February 8, 2024

* The authors are grateful for discussions and comments from Norris Bruce, Jiwoong Shin, Vineet Kumar, Alex Burnap, Soheil Ghili, Aniko Öry, Neeraj Arora, Kevin Chung, Remi Daviet, Emaad Manzoor, Qinglai He and participants at the Yale SOM Internal Marketing Workshop, UConn Marketing PhD seminar, UT Dallas Bass Forms Conference 2022, Wisconsin School of Business Rays of Research Workshop, Conference on AI, ML, and Business Analytics, Temple University 2021, AIM Conference 2020, USC and Marketing Science Conference 2020.

1. Introduction

Entry-level recruitment is a major expense for companies in areas such as sales, professional services, IT, and hospitality. Due to the large number of positions and the need for a consistent influx of entry-level employees, companies interview vast numbers of potential recruits annually for efficient onboarding and training. University campuses are a primary source of this talent, making campus recruitment an annual event for many firms (Wotruba et al. 1989, Rivera 2012). Resumes, while common for initial screening, are less effective for recent graduates due to their similar educational backgrounds and limited experience. Hence, firms conduct on-campus interviews for initial screening, followed by multiple in-house interviews for final selection (Camp et al. 2004). This is especially true for roles like sales and consulting where soft skills are vital (Maree et al. 2019). However, this method is costly, with average college recruiting expenses ranging from \$3,000–\$6,000 per hire.[1] As a result, firms are keen on tech-driven solutions to cut entry-level hiring costs (Black and van Esch 2020, Smith 2019).

Advancements in machine learning now allow the use of interview recordings and prior ratings to train AI models for assessing interviewees' competency and suitability. The COVID-19 pandemic has normalized remote video interviews, a trend likely to continue[2] as it eliminates travel costs, thus enabling companies to broaden their talent pool beyond certain top campuses.[3]

This paper is a first step towards using AI and Human-AI hybrid to analyze video interviews that are conversational in nature.[4] Conversations involve two-sided, multi-instance (back-and-forth utterances), and multi-modal (encompassing language, voice, and body movement) interactions. We derive objective features that influence persuasion from multiple modalities of text, audio,

[1]NACE's 2018 estimate for the average cost of college recruits was \$6,110, a 50% increase from 2014. Campus hires account for about half of entry-level recruits, and this number is growing (NACE Outlook 2018, NACE Outlook 2022)

[2]In 2022, 55% of firms reported that they use video interviewing, while that number is 53% in 2021 (Criteriacorp 2022). In addition, according to the Forbes' Remote Work Statistics (2023), 93% of companies that transitioned to video interviews for remote hiring continue to use it.

[3]Relying solely on school reputation is not always indicative of job success, as they lead firms to overlook potentially top candidates from outside the top tier of schools (Rivera 2015)

[4]Current AI hiring systems focus on resume matching or hard skill testing, and lack soft skill assessment.

and video channels, leveraging state-of-the-art methods such as posture modeling, conversational analysis, and Large Language Models (LLMs) in an important marketing context. In contrast to traditional persuasion studies, which rely on recall, our real-time metrics provide a more precise depiction of interactive dynamics.

We apply our model to entry-level sales intern hiring at a U.S. university. We train our AI model using video recordings of 15-minute situational interviews (sales role-play) structured around the National Collegiate Sales Competition (NCSC). The primary outcome, indicating the interviewee's latent sales ability, is derived from scores given by a panel of industry salesforce hiring managers. Specifically, we use the comprehensive score of interview performance based on the NCSC rubric.[5]

We highlight three key considerations in our AI model development. The first is the legal landscape. Unlike "low risk" AI applications in marketing, hiring is deemed "high-risk" in the EU and several U.S. jurisdictions.[6] This mandates AI hiring systems to offer decision explainability; which requires construction of theory-motivated features from the video interviews as inputs into the AI model.[7] Further, the literature has shown that increasing standardization by (i) using a structured interview format in asking questions and (ii) using a rubric for scoring, improves the psychometric properties of the evaluations and produces more consistent results across evaluators, ensuring that all relevant performance factors are considered in hiring (Campion et al. 1997, Levashina et al. 2014). This structure can also reduce potential biases based on gender, race, and disability (Levashina et al. 2014). Many companies, therefore, use structured interviews and rubric-based assessments as predictors of interviewee job performance to reduce noise and biases in hiring recommendations (McDaniel et al. 1994, Conway et al. 1995, Posthuma et al. 2002, Ashnai et al. 2020).[8]

[5]The NCSC, established in 1999, is the U.S.'s longest-standing university sales competition. The NCSC scoring rubric is well-established in B2B sales curricula among North American business schools (Loe and Chonko (2000), Widmier et al. (2007)). In addition, the skills evaluated in the NCSC rubric are aligned with entry-level sales job requirements as per the US Department of Labor guidelines (US DOL website).

[6]Several U.S. states have enacted laws addressing AI use in hiring, emphasizing interviewee consent, explainability, and bias audits.

[7]Deep learning models can provide post-hoc explainability, but given the scarcity of public video interview data and the vast theoretical literature on persuasion, feature engineering is more pragmatic and legally defensible.

[8]Case interviews of consulting firms, behavior interviews at Amazon, and sales position hiring are typically based on structured interviews and rubrics. `https://aws.amazon.com/careers/how-we-hire/`.

Hence, we view the use of the structured interview format and standardized rubric as strengths of our application.

The second is the lack of objective performance metrics for the interviewees at the time of hiring. Companies therefore have to rely on managerial predictions of latent skills. Note that even in retrospect, it is not feasible to train the model to discriminate between those hired and not hired using performance data, as such data will be available only for those hired. Moreover, across-evaluator heterogeneity in scoring styles, within-evaluator variability, and evaluator biases in scoring can lead to noise in the estimates of latent skills. By using the evaluations by diverse evaluators in training the AI model, we can account for heterogeneity and obtain a true estimate of the candidate's latent sales ability. Once trained, the AI models saves costs by reducing the need for a large panel of human evaluations during large-scale hiring.

The third is the role of human intuition. There is the possibility that the AI might overlook nuanced performance elements that humans intuitively recognize. For example, humans may be able to differentiate whether *hand movements* reflect innate enthusiasm or impatience/nervousness in the context of other cues, that may not have been captured in the AI model. While the validity of such concerns is empirical, integrating human input can enhance accuracy and foster AI acceptance.[9] However, human input is expensive, presenting an accuracy-cost tradeoff. We evaluate the extent to which human input enhances AI predictions and determine the optimal balance in a human-AI hybrid model. By addressing these considerations, we aim to create a robust, legally compliant, and effective AI-based hiring system.

We now discuss some challenges in the model development. For the AI model, a primary challenge is to obtain theory-driven, explainable features from the video interviews. We rely on the extant literature on persuasion and salesforce which is mainly survey-based to identify salespeople's content, interactivity, verbal and non-verbal styles. Our opportunity and challenge here is

[9]Studies (Tambe et al. 2019, Shrestha et al. 2019) indicate that both recruits and recruiters prefer a blend of human and AI input, ensuring legal compliance and comfort.

to directly measure these constructs from the videos encompassing text, voice, and human movement (e.g., posture, and facial expressions) without the potential biases from human recall and evaluation. We use OpenPose (Cao et al. 2019), to transform videos into low-level data tracking body part positions over time. Our contribution lies in converting such panel data of human body part positions into quantifiable higher-level body language metrics, like hand gesture rate. We also leverage the capabilities of Large Language Models (GPT) to create linguistic style and adaptation-related variables such as collaborativeness, analytical, politeness, active listening and style matching (or mimicry) from the textual transcriptions of the video data. Thus, this is one of the early papers that illustrates how these tools can be used to generate theoretically meaningful higher-order metrics for persuasion (and more broadly marketing) research questions.

For the AI-human hybrid model, we propose a simple Bayesian method to enhance AI predictions with human insights. Post AI training, our model integrates human judgment into the hiring process. We suggest merging AI and human evaluations using a Bayesian weighted average, with weights determined by the estimated uncertainty in each score. The AI score's uncertainty for each interview is gauged using a Jackknife-after-Bootstrap method (Wager et al. 2014), while the variability across human evaluators' ratings provides the standard deviation for human scores.

Our key findings are as follows. Our AI model achieves good performance in predicting candidates' latent sales abilities, and subsequently selecting or screening candidates based on their rankings of predicted sales abilities. To assess model performance, we propose two types of managerially important metrics. First, using binary classification metrics for selection (screening), we assess the model's accuracy in identifying the right (or the wrong) candidates. Second, to assess the workforce quality loss to the firm due to errors in the model's predictions, we measure competency loss from selecting candidates that are below the bar, and the opportunity loss from not selecting candidates that are above the bar. In terms of the model's classification performance in selecting candidates above the 75th percentile, we achieved an AUC of 74%. We also find that our AI model improves workforce quality by 40% relative to the random baseline. With the addition of 1 human input, the AI-human hybrid achieves an AUC of 83%, and workforce quality gain of 67%.

In terms of relevant features that impact prediction, we find that all feature groups—content, interactivity, and both verbal and non-verbal styles–derived from text, audio, and visuals influence predicted sales abilities. Content and interactivity have the highest incremental impact on prediction. Specifically, conversational interactivity and adaptability (e.g., buyer-seller share of voice, active listening, style-matching) and optimal body language, like moderate hand movement, are positive indicators of latent sales skills. Previous studies on personal selling have been ambiguous about the effects of adaptation and style-matching, largely due to survey-based measurement errors (Lichtenthal and Tellefsen 2001). Finally, the most cost-effective AI-human hybrid model requires only one human evaluator, focusing on the initial two stages (or first 7 minutes) of the interview.

While the current application focuses on entry-level B2B salesforce hiring, our approach is generalizable to other hiring contexts. For instance, Table B.1 in the appendix outlines common entry-level jobs for recent graduates and the corresponding occupational requirements based on the US Department of Labor's classifications. Notably, many of these roles require skills like establishing rapport, needs identification, and handling objections, which align with the National Collegiate Sales Competition (NCSC)'s scoring rubric for identifying competent salespeople. Thus, our process of analyzing conversational video interviews can be applied to other hiring contexts as well. This would enable firms to find a cost-effective way of using AI to substitute the need for a large panel of human judges in the hiring process, and save managerial time and costs for candidates' evaluation.

The rest of the paper is organized as follows. §2 positions the paper with respect to the related literature. §3 describes the data, while §4 describes the AI and AI-human hybrid models. §5 discusses the main findings. Finally we conclude with a discussion of limitations and future research.

2. Related Literature

This paper is related to three strands in the literature: i) the recent literature on the use of video data in AI/ML applications; ii) the marketing and psychology literature on what drives persuasion and success in sales pitches; this provides guidance for the input variables in the AI model; and iii) the management and computer science literature on how AI and human judgment can be combined to improve decision making.

There is a small but growing machine learning literature in marketing using video data. However, unlike this paper, these papers do not focus on the dynamic content in videos such as human movement or the interactive elements of two-way conversation, which are more important in our setting i.e., evaluating video interviews. For example, Rajaram and Manchanda (2020) focuses only on the static frames from the beginning, middle, and end of 30-second influencer videos along with metadata; while Yang et al. (2021) focuses on known object recognition within each frame of the video to study whether pixels where objects are displayed have high engagement. Even when video content is studied, the focus is often on a single modality (text, voice) and one-sided communication (e.g., a video resume or crowdfunding pitch). For example, Wang et al. (2021) and Hwang et al. (2021) study persuasion in the context of crowdfunding pitches and influencer videos respectively focusing on the voice modality as other modalities are less important (or have no variation) in their setting.[10] Closest to our work, is a small and growing literature about automated interview scoring (e.g., Chen et al. (2016), Naim et al. (2015)). However, Chen et al. (2016) study monologue interviews (similar to those used by commercial systems like HireVue) where-in online participants submit deliberated responses to some standard behavioral interview questions. Our work adds to this nascent literature and is the first to study two-sided, multi-instance, multi-modal conversations leveraging state-of-the-art models in conversational interactivity and human body movement.

Second, we contribute to the understanding of drivers of persuasion in the sales interview context. The study of factors that contribute to successful persuasion in various settings has a long history in both marketing and psychology. Early literature is mostly conceptual (e.g, Sheth (1976)) and builds on Aristotle's original treatise that identifies three main channels of persuasion—logos (logic), pathos (emotions) and ethos (value system). The next wave of studies mainly included surveys (Frazier and Summers 1984, Dubinsky 1981, Spiro and Weitz 1990, Moncrief and Marshall 2005)

[10]Similarly, in computer science literature, Shmueli-Scheuer et al. (2019), Longpre et al. (2019) study persuasion in the context of textual debates.

where salespeople and their clients were interviewed post interactions to understand factors that influence the outcomes of the exchange. While survey-based approaches can identify some high-level influence tactics (as interpreted and recalled potentially with recall bias by respondents), they typically cannot characterize these tactics into relevant, specific objective behaviors that drive success. For example, there is no way to measure behaviors such as *rate of hand movements* that can be mapped to constructs such as confidence or nervousness or *real-time active listening* that reflects adaptability without the frame-by-frame recorded videos of the actual buyer-seller interaction. In the past literature, buyer-seller similarity are characterized mostly in terms of crude descriptors (e.g., demographics or dressing style) often resulting in inconclusive and weak findings (Lichtenthal and Tellefsen 2001, Churchill et al. 1975, Evans 1963).

Finally, our paper is one of the early empirical works to study the the application of hybrid AI-human models in the context of organizational decision making; current literature being largely theoretical. For example Shrestha et al. (2019) propose AI-human hybrid models of organizational decision making and Holstein et al. (2020) conceptualized such models in the domain of education. Hiring is an ideal setting to study AI-human hybrid models, given both the complex legal landscape and the need for time and cost efficient decisions. Yet, there is surprisingly little empirical work on the value of augmenting AI with human input in this domain. While there has been some recent research on the role of human intervention on AI recommentations, post model deployment (van den Broek et al. 2021), our paper is focused on integrating human input with AI predictions to aid decision making.

3. Data

We build the AI model using 195 videos of in-person sales interviews in the format of the National Collegiate Sales Competition (NCSC). Interviewees are students from a large research university in the US and each student participates in one interview only. The interviewers are B2B sales professionals (managers to VPs and CEOs) from 162 firms representing a wide range of sectors such as pharmaceutical, manufacturing, hospitality, insurance, transportation, technology, etc.[11]

[11]See Table B.2 in the Appendix for the list of companies.

Every interviewer conducts about 8 interviews on average. Every interview is a sales-pitch role-play where the student (interviewee) persuades the corporate buyer (interviewer) to buy a CRM Saas product. Both sides are motivated and incentivized to perform their roles well to closely mimic actual hiring scenarios – firms are seeking talent, while students are seeking to gain a foothold in the B2B selling industry.

Each sales interview is subsequently scored by a panel of 9 industry judges from the same pool as the interviewers. These serve to construct the supervisory labels of "latent sales ability" for the AI model.[12] Figure 1 shows a snapshot of the interview setting.

Figure 1 **Interview setting (Left: candidate (seller), Right: interviewer (buyer))**

Table 1 shows the distribution of demographics for interviewees, interviewers and judges. There is a good balance of gender and experience among the buyers and judges.

The interview is divided into 5 stages as per the NCSC format described in Table 2a. We can think of the first two stages (*Approach* and *Needs identification*) as precursors to selling, while the subsequent stages form the core of a salesperson's persuasion effort. A good *Approach* can "create mutual liking" (Cialdini 1987). The second stage of *Needs identification* is important because only by uncovering a buyer's needs can a salesperson tailor their pitch persuasively (Rackham 2020). As for the subsequent stages – in *Product/Service Presentation*, the objective is to demonstrate

[12]While the human interviewer (buyer) may directly judge the interviewee, the NCSC recommends an external panel of multiple judges to avoid variations in evaluations due to the particular interviewer's idiosyncrasies and biases.

Table 1 Interviewees, interviewers, and judges demographics

	N	Gender	
Candidate (Seller)	195	96 females, 99 males	
	N	Gender	Experience
Interviewer (Buyer)	29	13 females, 16 males	9 high, 20 low ($\leq$10 yrs)
	N	Gender	Experience
Judges	291	99 females, 192 males	134 high, 157 low
Evaluations		1752	

Table 2a Different stages of selling process

Stage	Approximate timestamp	Description
Approach	0-1.5 minutes	Seller greets buyer, builds rapport
Needs Identification	1.5-6 minutes	Seller digs deeper to uncover the buyer's needs
Product/service Presentation	6-10 minutes	Seller gives a demonstration of the product/service
Objections Handling	10-14 minutes	Buyer raises objections, seller presents counter arguments
Close	14-15 minutes	Seller asks for commitment from buyer or persuades buyer to sign a contract

the product's benefits in a way that matches the buyer's expectations. *Objection Handling* involves coming up with compelling arguments to counter buyers' concerns. The final stage is *Close*, where the seller has to take the initiative to understand where he/she stands with the buyer now and for the future, and be persuasive in presenting a reason to buy.

Each stage of the sales process is evaluated according to the NCSC scoring rubric (Table 2b).[13] There are 25 questions in the scoring rubric. Each individual question is scored on a 5 points likert scale, which is mapped numerically to $\{6, 7, 8, 9, 10\}$. Table 2c shows the summary statistics of these scores at the judge-candidate level. There is a large variation in scores, both between candidates and within candidates. Interestingly, we also see that judges become more stringent as the stage progresses – the average score gradually becomes lower going from "Approach" to "Closure." The total score for each stage is obtained by adding up the individual question in that stage, then dividing by the number of questions in that stage, and multiplying by 10. Finally, the overall Total Score is a weighted sum of these groups of questions according to the weights in Table 2b.

The structured interview format, the multi-stage scoring rubric, and the gender-balanced panel of judges help to minimize variability in judgment and potential biases towards different groups –

[13]The NCSC scoring rubric is well-established in B2B sales curricula among North American business schools (Loe and Chonko (2000), Widmier et al. (2007)).

(Kahneman et al. 2016, Campion et al. 1997, Pulakos and Schmitt 1995, Ashnai et al. 2020). It also helps firms to ensure that managers consider all factors deemed relevant by the firm as opposed to only the manager's preferences.[14]

Table 2b National Collegiate Sales Competition (NCSC) Scoring Rubric

	Questions	Weight
Approach	Professional Introduction Salesperson gains prospect's attention Effectively builds rapport Smooth transition into needs identification	5%
Needs Identification	Uncovered decision process Effectively determines relevant facts about the company and/or buyer Effectively uncovered needs of the buyer Used implication and needs payoff questions Gained pre-commitment to consider the product/service and smooth transition to presentation	25%
Demo	Presented benefits upon needs of buyer instead of features Logical, convincing presentation Used appropriate visual aids Effectively demonstrated the product/service Effectively involved the buyer in demonstration Effective use of trial closes	25%
Objection Handling	Initially gains better understanding of objection Effectively answers objections Confirms objection is no longer a concern	15%
Closure	Persuasive in closing the deal Asked for appropriate commitment	10%
Communication Skills	Effective verbal communication Effective non-verbal communication Verbiage (clear, concise, professional)	15%
Overall	Salesperson enthusiasm and confidence Product knowledge	5%

Table 2c Summary statistics of scores

				Stages			Entire interview
	Weighted Total	**Approach**	**Need**	**Demo**	**Objection**	**Close**	**Communication**
Mean	76.6%	77.4%	77.1%	76.2%	74.5%	73.7%	79.3%
Median	76.1%	77.5%	76%	76.7%	73.3%	70%	70%
Standard deviation	6.6	7.1	7.1	7.7	8.6	9.7	7.6
Min	60%	60%	60%	60%	60%	60%	60%
Max	100%	100%	100%	100%	100%	100%	100%

[14] Unlike in consumer choice, where the consumer's choice is the metric that matters to the firm, firms want to choose employees that have a holistic set of skills, determined is relevant for the firm, not merely a manager or evaluator's overall choice.

4. Model

Let θ_i be the metric that represents the candidate i's latent sales ability; and $\boldsymbol{X}_i$ be the set of multi-modal theory-relevant features extracted from the video interview data that can predict sales ability. The AI model seeks to predict θ_i as a function of $\boldsymbol{X}_i$.

$$\theta_i = f(\boldsymbol{X}_i) \tag{1}$$

For $f(\cdot)$, we choose the best fitting nonlinear machine learning model among Random Forest (RF), Support Vector Machine (SVM) and XGBoost. This best model is the AI model.

We first describe how we extract variables, $\boldsymbol{X}_i$, from the video data. Then, we describe how we convert judges' ratings into latent sales abilities, θ_i. Finally, we describe how we combine AI prediction and human judgment for the AI-human hybrid model.

4.1. Multi-modal features extraction

The selling and persuasion literature (Dubinsky 1981, Moncrief and Marshall 2005, Sheth 1976, Frazier and Summers 1984, Spiro and Weitz 1990)) suggests three distinct groups of features that matter for sales persuasion: (i) content (what is said), (ii) styles (how it is said), and (iii) interactivity/matching in conversations. In the previous literature, many of these were typically quantified using surveys, based on recall or through the judgment of observers. In contrast, we use videos to objectively construct these features.

Our content and verbal style features rely on text data of the conversation. To generate the text data from the video, we extract the audio component of the video and then convert it into a text transcript using a Speech-to-Text API.

For non-verbal styles such as body language and vocal, we use frame-by-frame audio and video data.[15] In all modalities, we extract separate features for the interviewer and interviewee. For interactivity and adaptation features, we also consider the interactions between the interviewer and interviewee.

[15] As our video recording only has side views of the participants, so we cannot extract facial features.

4.1.1. Content. The product to be sold in the sales role-plays is an enterprise CRM software. Given the product is the same in all role plays, there are similarities among interviewees in how the product is sold. However, they differ in how much time they allocate to various elements of the selling and persuasion process. Some candidates focus more on the product characteristics, some focus more on the client's business objectives and needs (growth, operational costs), while others focus on price negotiations, and how technology can address business challenges. Using a topic model (LDA), we identified these broad topical areas: *Greetings*, *Business*, *Product*, and *Pricing*.[16]

We calculate the proportion of spoken words that belong to these different topics. Table 3 shows the highest frequency words on each topic. We also count the number of words spoken by the candidate. Since every candidate has a fixed 15 minutes, the *Seller's Word Count* acts as a proxy for the information density of the content.

Table 3 Topic-wise high-frequency words

Topics	High-frequency words
Greetings	morning, weather, yankee, holidays, spring break, christmas, thanksgiving, family, alumni, reunion
Business	business, leads, growth, profits, employees, territory, potential client, profitability, goals, due diligence, operations, managers, teams, processes, travel, rep, salesperson
Product	data, cloud, digital, licensing, website, kiosk, technology, pipeline, mobile, database, app, tool, downtime, breakdown, learning curve
Pricing	pricing, discount, monetary, afford, cost-effectiveness, expenditure

4.1.2. Style. We consider three sets of style features: (i) verbal, (ii) body language, and (iii) vocal. We discuss each of these in turn.

Verbal Style. The literature on sales and persuasion has identified that being confident, collaborative, and polite are important traits for salespeople. Such traits can be conveyed through the text modality in the form of verbal styles. Overall, we identified 8 verbal styles that are important for persuasion, as in Table 4.

The challenge here is there are no established methods to capture these traits except for *politeness* and *language complexity.* Moreover, merely counting occurrences of words related to say,

[16]The unsupervised LDA does not produce topic labels automatically, we chose the labels for the top identified LDA classes based on input from our sales lab personnel using their domain knowledge. A supervised model such as CNN could also be used for topic identification as in Timoshenko and Hauser (2019), but even here, topic labels based on expert judgment are needed.

Table 4 Verbal Styles

Linguistic style/tone	Related salesperson's traits	Operationalization
Confidence	Certainty, Assertiveness, Competence (Erickson et al. 1978)	LLM-based text embedding (OpenAI's `text-embedding-ada-002`)
Analytical	Reasoning (Xiao and Khazaei 2019, Markowitz 2020), quantitative	LLM-based text embedding
Collaborative	Team player vs. taking ownership (Hawes and Rich 1998)	LLM-based text embedding
Optimism	Achieving targets (Rich 1999, Moncrief and Marshall 2005, Syam et al. 2013)	LLM-based text embedding
Caution	Highlighting risks (Moncrief and Marshall 2005)	LLM-based text embedding
Competitiveness	Winning spirit (Syam et al. 2013)	LLM-based text embedding
Politeness	Gratitude and Deference (Wilson et al. 1991)	LLM-based text embedding
Language Complexity	Ease of Understanding (Pogacar et al. 2018)	Flesch–Kincaid score, Type-Token Ratio, Percentage of words with 7 letters or more.

"collaborative," may not fully capture the semantic meaning of "speaking in a collaborative tone." Hence to measure the extent a salesperson uses a linguistic verbal style, we use the following two key ideas: i) an embedding (specifically, sentence embedding) captures the semantic meaning of a sentence (or spans of contiguous text) in a fixed-size numeric vector, without relying on a dictionary of words (Reimers and Gurevych 2019), ii) texts that have similar meanings are closer to each other in the embedding space. In our specific context, if an interviewee's speech demonstrates a certain verbal style, say collaborativeness, then it should have a higher cosine similarity in the embedding space with exemplar sentences that capture *collaborativeness*. By exemplar sentences, we mean sentences or phrases that convey a particular style. For example, an exemplar sentence for collaborativeness would be "I am open to your suggestions and feedback," and for a confident tone would be "I guarantee you won't be disappointed with this purchase."

For brevity, we describe the process for constructing the verbal style of "collaborativeness." The approach is similar for the other styles and the details can be found in Section E of the appendix. The process consists of the following steps:

1. *Generation of exemplar sentences.* We start by generating sentences that embody a particular verbal style using GPT-4. We prompt[17] GPT-4 in two ways. First, we use a generic prompt

[17]Prompting an LLM is an example of in-context learning. A prompt is defined by a template, which contains placeholders for the description and demonstrations of the inputs and outputs for the task that enables humans to interact with a Large Language Model using natural language (Arora et al. 2022)

"Could you provide us with 10 phrases/sentences that convey a collaborative tone?" Then, to get some context-specific exemplar sentences we ask GPT-4, "You are a salesperson from Salesforce.com meeting a prospective buyer. Could you provide 10 phrases/sentences specific to B2B selling that convey a collaborative tone to the buyer?"

2. *Validation of generated sentences.* We ask our human survey participants from the crowd-work platform, Prolific.com[18] to rate the sentences for relevance to a specific trait: "On a scale of 1-10, how would you rate [exemplar sentence] in conveying collaborative." In Table E.1, we illustrate the exemplar sentences generated for the style *collaborative* and the corresponding scores.

3. *Verbal style embedding.* To derive the verbal style embeddings, we derive an embedding for each of the exemplar sentences using OpenAI's `text-embedding-ada-002`. The advantage of using sentence embeddings is that we are not matching exact phrases/sentences, hence, the list of exemplar sentences need not be exhaustive, as long as the exemplar sentences capture the intended meaning. Recently, these embeddings have been used for semantic similarity tasks in a variety of domains such as company classification (Vamvourellis et al. 2023) and finding typicality measure (how similar a text document is to a concept) in political discourse (Le Mens et al. 2023). Although there could be alternative means of conveying a style, the semantic distance metric remains resilient to variations in words or phrases used to express the same concept.[19]

4. *Weighted average of cosine similarities.* We calculate the cosine similarity of the interviewee's transcript with each of the exemplar sentences in the embedding space. Then we take the weighted average across sentences, where the weights are obtained from the scores in the Prolific surveys. In the end, we have a measure that captures the presence of a linguistic verbal style in an interviewee's transcript, which is used as a feature in the AI model.

[18]We restricted crowd workers to people who self-report to have work experience in the sales domain. A minimum of three crowd-workers assessed each sentence, and we calculated the sentence's average score.

[19]To test this, we also use different subsets of the exemplar sentences to construct our features.

For *Politeness*, we use the same LLM-based sentence embedding approach but directly use the exemplar sentences in Danescu-Niculescu-Mizil and Lee (2011), as in Table E.7 in the Appendix. *Language Complexity* can be deconstructed into two fundamental components: readability, which gauges how easily a text can be comprehended, and lexical diversity, which assesses the range of vocabulary used in an individual's speech. These two aspects are operationalized using Flesch–Kincaid readability (Flesch 2007) score and Type-Token ratio (TTR) which are established metrics for capturing readability and lexical diversity.[20] We also constructed *LongWords*, which measures the fraction of a seller's words that have seven or more letters. Natural spoken text generally has a lower percentage of long words than formal text (Pennebaker et al. (2001)). *LongWords* is related to the Flesch–Kincaid readability score, since longer words are deemed less readable.

Body Language Style. Visual elements that can communicate style include hand movements (velocity and amplitude) as well as body postures. While the interpretation of body language can differ by context and culture, there is some convergence around what certain gestures communicate (Fast 1970, Krauss et al. 1991, Kendon 1994, Pease and Pease 2008). For example, wide open hands generally convey confidence and warmth whereas clenched fists and closed hands convey nervousness. Likewise, a relaxed posture is generally preferred to slumping or an uptight torso. Interpretation of frequency of body movements is ambiguous — while a moderate amount of hand and head movement may signify showing assent and active listening, too aggressive hand movements may be interpreted as threatening or nervous.[21]

We proceed in two steps to extract body language features. First, we use OpenPose (Cao et al. 2019) to extract low-level features from the videos (e.g., pixel coordinates of body parts at every frame). Figure 2 shows the keypoints of the body parts used by OpenPose. As an example, keypoint

[20]The Flesch–Kincaid formula presents a score as a U.S. school grade level. The formula for the Flesch–Kincaid readability score is $FK = (0.39 \cdot ASL) + (11.8 \cdot ASW) - 15.59$, where ASL is the average sentence length, and ASW is the average number of syllables per word. TTR is calculated by dividing the number of unique words (types) by the total number of words (tokens) in a text. A higher TTR indicates greater lexical diversity.

[21]We note that facial features would have certainly enhanced the predictive power of the AI model, but it was difficult to extract them correctly as our videos had side views and not frontal views of the two parties.

7 indicates the wrist, while keypoint 0 refers to the nose. We then convert these low-level features into higher-level body language features (e.g., rate of hand gestures) that are theoretically known to impact persuasion.

Figure 2 Key Points from OpenPose

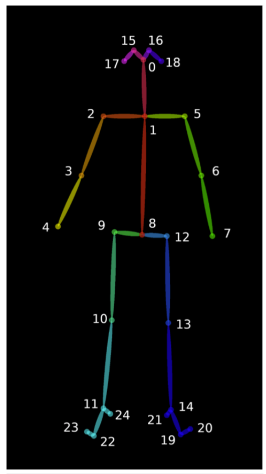


Figure 3 OpenPose in different interview stages

(a) Greet

(b) Identify needs

(c) Demo

(d) Objection Handling

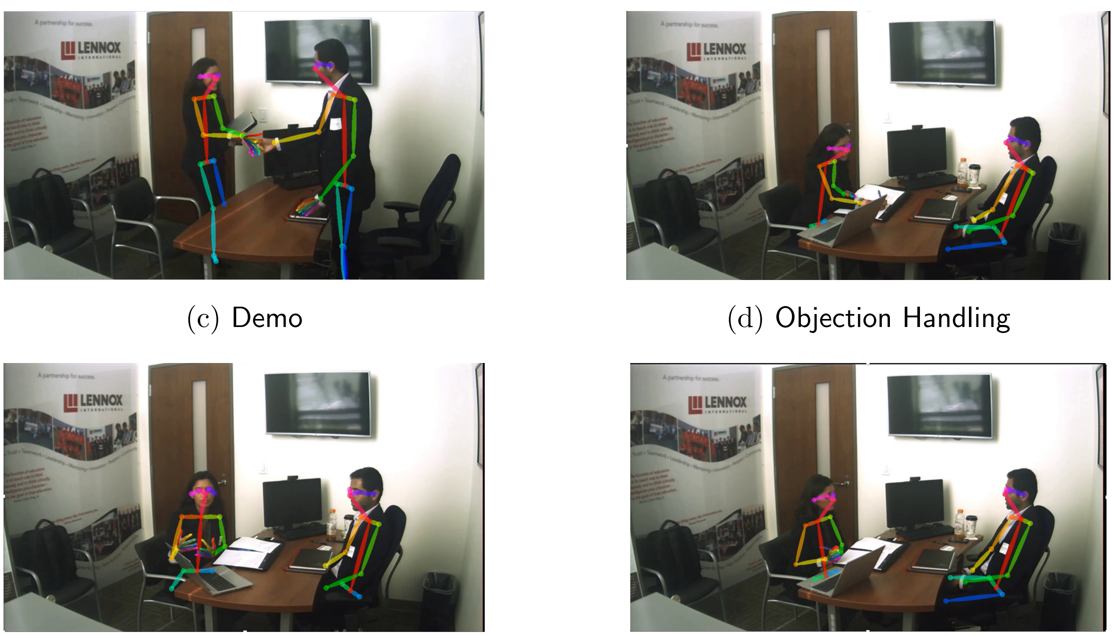


Figure 3 shows OpenPose in action across the different stages of the interview. OpenPose jointly detects and associates body parts to a particular person. The result is a wire-frame for every individual consisting of 24 identified keypoints or pixel locations of important body parts which includes — eyes, nose, ears, neck, shoulders, elbows, wrists, hip, legs and toes.

Using the low-level features extracted from OpenPose, we compute higher-level body-language features as described in Table 5 for each frame and then construct an average across frames. We illustrate this computation by describing in detail the "hand gesture rate" feature. Let $Wrist_t$ be the vector of coordinates of the seller's right wrist at frame t. This is the output of OpenPose, specifically, keypoint 7 in Figure 2. The coordinates are given in terms of the number of horizontal and vertical pixels from the lower left.

A measure of hand gesture rate at frame t is the distance traveled by the wrist across two consecutive frames. However, the unit of this distance is in pixels, which is affected by the resolution of the video or the distance of the camera from the objects. As such, we normalize the unit of distance with the length of the subject's forearm.

To measure the length of the seller's forearm, let $Elbow_t$ be the vector of coordinates of the seller's right elbow (OpenPose's keypoint 6) at frame t. The length of the seller's forearm at frame t is $\|Wrist_t - Elbow_t\|$, where $\|\cdot\|$ is the Euclidean norm. The length of the seller's forearm (in terms of pixels), is the average across $\|Wrist_t - Elbow_t\|$ for frames $t = 1, \ldots, T$.

Finally, we define the right-hand gesture rate as Equation 2 below.

$$\text{Right Hand Gesture Rate} = \frac{T \sum_{t=2}^{T} \|Wrist_t - Wrist_{t-1}\|}{(T-1) \sum_{t=1}^{T} \|Wrist_t - Elbow_t\|} \tag{2}$$

We averaged across all the frames to obtain an overall hand gesture rate. We repeat the above steps to obtain the hand gesture rate for the seller's left hand and then average the left and right-hand gesture rates to arrive at the hand gesture rate.

Similarly, from the low-level OpenPose output, we are able to construct the rate of head movement, hand openness, and torso angles. The complete formulas are detailed in Appendix G.

Vocal style. For vocal style, we extracted the audio signal and then used a variety of signal processing tools to transform these signals into meaningful features relevant to sales persuasion.

Table 5 Description of body language features extracted from videos

Variables	Formula
Left-hand gesture	Distance traveled by the left wrist from frame t to frame $t+1$, normalized by the length of the left forearm.
Right-hand gesture	Distance traveled by the right wrist from frame t to frame $t+1$, normalized by the length of the right forearm.
Distance between hands	Distance between left and right wrists, normalized by the width of the shoulder.
Head movement	Distance traveled by nose from frame t to frame $t+1$, normalized by the length of the neck.
Torso angle	Torso vector is the vector between neck and hip. Angle of torso is the angle between the torso vector and the horizontal vector $(1,0)$. A smaller angle means the person is leaning forward.

Research in psychology has found that an energetic, well-modulated, and non-monotonic voice shows confidence and increases persuasion (e.g., Van Zant and Berger (2020)). We operationalize "energy" as the sum of squares of the signal's amplitudes over each two-second window, and then take an average over all windows in the video. We proxy for voice modulation and monotonicity as "energy entropy," which measures the change in energy across windows. Further, research has found conflicting effects of voice brightness (sharpness) on persuasion – while it can increase perceived competence, it can also come across as less warm (Zoghaib 2019).[22] Following the literature, voice brightness is implemented as the "spectral centroid", which is obtained by first applying the Fourier Transform to the audio signal over each 2 seconds window, and then calculating the mean of the resulting spectrum distribution. We then take the average across all windows.

Our approach is similar to Chang et al. (2022), who also construct audio features such as energy, energy entropy, and spectral centroid to study their effects on persuasion. These features are also related to loudness, pitch, loudness variability and pitch variability used in Hwang et al. (2021).

4.1.3. Interactivity, Adaptation and Matching in conversations. Figure 4 shows a sample transcript of an interview conversation between the interviewer and the candidate (interviewee). We describe how we construct relevant constructs from these transcripts.

Interactivity: Conversations are intrinsically interactive and interactivity is critical to persuasion. We extract several interactivity elements from the transcripts (e.g., share of voice, turns, max length of monologues). For this, we extract time-stamped text transcripts from the video data.

[22]A bright voice is often perceived as being sharp, while a dull voice is perceived as being soft. Zoghaib (2019) found that speakers with dull (vs bright) voices are perceived to be warm and likeable though there are some gender effects.

Figure 4 Interview transcript

```
        Timestamp
     [00:00:06] Interviewer: Come on in.  } Utterance

     [00:00:12] Candidate: Hello from Salesforce. I am Jacob, I will be
Turn walking you through our product today. Almost didn't make it today.
     The traffic in Dallas can be horrendous on Fridays! Well, good news
     is it is the end of the week, any plans for the weekend?

     [00:00:28] Interviewer: Yeah. This weekend, I've got a few friends
     coming over. We're gonna play some Dungeons and Dragons.

     [00:00:32] Candidate: Oh, great, it is one of my favorite games
     too.

     [00:00:42] Interviewer: Yeah, I am very excited, haven't seen these
     folks 2 years since we graduated from college.
```

Then, using the time-stamp of the start and end times for every utterance of each speaker, we define the following interactivity features: (1) *Turns* is the number of times the conversation switches from the interviewer to the candidate. A higher number of turns is associated with more interactive conversations (Sacks et al. 1978, Hammer et al. 2004). (2) *Buyer-to-Seller Share of Voice* is the ratio of the total speaking times of the interviewer to the candidate, where the total speaking time for a speaker is the sum of the duration of all utterances associated with that speaker. (3) *Max Monologue* measures the longest continuous utterance of either the interviewer or the candidate.

Real-time adaptation: Real-time adaptation has been shown to impact persuasion (Spiro and Weitz 1990). We consider two specific types of adaptation– active listening and unconscious style matching.

Active listening involves effective paraphrasing, asking clarification questions, and showing interest in what is being said. Paraphrasing is a task fundamental to Large Language Models like GPT as its train/test corpus already consists of standard paraphrase datasets such as Microsoft Research Paraphrase Corpus, and Quora Question Pairs.[23] Hence, we use the expertise of GPT to help us score our conversations for active listening.

Using few-shot prompting, we provide GPT with some specific examples of what is good/bad active listening in our context. This approach is favored over zero-shot prompting (i.e., directly

[23]The performance of language models is evaluated on the basis of the General Language Understanding Evaluation (GLUE) benchmark (Wang et al. (2018)). One of the main tasks in GLUE is the ability to classify paraphrases accurately according to the Microsoft Research Paraphrase Corpus (MRPC) and Quora Question Pairs (QQP)

asking GPT to give a score for an interviewee's transcript).[24] We are able to identify noteworthy instances of good and bad active listening through the comments left by the judges in the comments section of the scoring sheets, as in Figure F.1 in the appendix.

We feed entire transcripts as examples to the GPT API during few-shot prompting. Each transcript has around 3,000 words. The maximum context we can provide to the `gpt-3.5-turbo-16k` API is 16k tokens, which corresponds to a limit of 2 example transcripts per prompt.[25] Thus, we are able to obtain a numeric score for active listening on a scale of 6-10 for every conversation transcript.

Unconscious style matching: Past research shows that when one party in a conversation tries to match the linguistic style of the other party, it can lead to better rapport building and more effective negotiations (Ireland and Pennebaker 2010, Taylor and Thomas 2008, Cialdini and Goldstein 2004, Chartrand and Bargh 1999). While excessive style matching can appear superficial and unnatural, doing it in the right proportion can enhance conversational outcomes. Following Ireland and Pennebaker (2010), we use classes of words that are known to be processed by humans in a non-conscious manner and hence are often used for subconscious style matching. Examples include auxiliary verbs (can, could, might), high-frequency adverbs (absolutely, certainly), conjunctions, impersonal pronouns, negations, personal pronouns, prepositions, and quantifiers (all, some, both).[26]

We operationalize style matching borrowing from the notion of coordination in Danescu-Niculescu-Mizil and Lee (2011) and Danescu-Niculescu-Mizil et al. (2012) – we first calculate the candidate's probability of using a word class immediately following the interviewer's usage of the

[24]There is evidence that GPT is a few-shot learner capable of in-context learning Brown et al. (2020), and few-shot prompting improves on zero-shot significantly.

[25]The cost of running through all the videos is as low as \$8 (\$0.035 per video). We also experimented with `gpt-4-32k` API for a smaller sample. The results remain similar while it costs 20 times more, totaling \$140 to run through the videos once.

[26]There are nine LIWC-derived categories with 451 words in total.

word. Then, positive adaptation occurs when this conditional probability is higher than the candidate's unconditional probability of using that word (i.e., the candidate's normal usage of that word).[27] We provide more details in Section F.2 in the appendix.

4.2. Outcome Variable

Our outcome variable is derived from a structured rubric consisting of the weighted scores from different stages of the standardized interview process. There are several benefits to using a structured rubric that takes into account all the stages of the evaluation process, rather than relying on judges' final hiring recommendations. Past literature has shown that increasing standardization – (i) using a structured interview format in asking questions, (ii) a rubric for scoring – improves the psychometric properties of the evaluations and produces more consistent results across evaluators, ensuring that all relevant performance factors are considered in hiring (Campion et al. 1997, Levashina et al. 2014). This structure can also reduce potential biases based on gender, race, and disability (Levashina et al. 2014).

One way to measure a candidate's latent sales ability is to take the average of the candidate's score across all the judges who evaluated the interview. However, such measure is influenced by whether an interviewer is tougher or more lenient, as well as the scoring styles of different judges. A unique aspect of our setting is that an interviewer interviews multiple candidates, and a judge evaluates multiple interviews. We use this to isolate the candidates' latest sales ability with a fixed effects regression, controlling for interviewer and judge fixed effects.

Let i denote the interviewee (seller). Let j denote a human judge from the panel of judges. Let k denote the interviewer (buyer). Let S_{ijk} denote the score given to the candidate i by a human judge j, when the interviewer is k:

$$S_{ijk} = \theta_i + \gamma_j + \delta_k + \epsilon_{ijk} \tag{3}$$

[27]The literature has also discussed the role of buyer-seller similarity, however, these studies mainly focus on demographic or observable physical attributes of similarity e.g, gender, age, height (Evans 1963, Churchill et al. 1975, Woodside and Davenport Jr 1974) or surveyed characteristics like political orientation or education levels. As a result, findings have been largely inconclusive.

The score S_{ijk} is the *Total Score*, the weighted sum of all the components of the NCSC scoring criteria described in Table 2b. Here, γ_j is the effect of judge j, δ_k is the effect of the interviewer k, and θ_i is the candidate i's latent sales ability, controlling for judges and interviewer fixed effects.

We estimate Equation 3 using ordinary least squares (OLS), to estimate $\hat{\theta}_i$. We treat this as a proxy for the candidate's latent sales ability and the outcome variable to be predicted by our AI models (Equation 1). Note that our measure of latent sales ability reduces to an unweighted average of judges' scores when the judge-specific or interviewer-specific fixed effects are zero.

A natural validity question in using the total score by judges as the outcome variable is whether it would be correlated with a judge's hiring recommendation. To assess this, for a new set of interviews conducted at the university during the Fall semester of 2022, we collected "hiring recommendations" in addition to the NCSC's scores. The average correlation between the scores and the hiring recommendations at the level of each judge is 0.78, indicating that rubric-based score correlates with the evaluator's hiring choice. However, given the benefits of the structured interview format and scoring rubric, we use the total score from the NCSC's rubric as the outcome metric.

4.3. Model Training and Testing

In the first step, we use the features extracted (as described in 4.1) to predict $\hat{\theta}_i$, the candidates' estimated latent sales ability (described in 4.2). Specifically, let $\hat{\theta}_i$ be the latent sales ability of candidate i and $\hat{\theta}_i^{AI}$ be the predicted salesperson ability from the AI model. We consider different machine learning algorithms (SVM, Random Forest, XGBoost) for predicting the continuous latent sales abilities. We train and test the model using 4-fold cross-validation, by splitting the dataset into training and testing. For all metrics, we report the average over the 4-fold cross-validations.

In the second step, we rank candidates based on their predicted scores $\hat{\theta}_i^{AI}$, and select those candidates that are above a certain *rank threshold*. Firms specify and set this threshold based on their hiring needs – for example, candidates above the 75th percentile.

Once a rank threshold is specified, the problem becomes a binary classification task, i.e. whether a candidate is correctly classified as selected versus not selected. Here, we set a threshold of 75th percentile for selection, and a threshold of 25th percentile for screening.[28] Thus, a candidate i is correctly classified if i belongs to the top 25th percentile (75th percentile) according to both $\hat{\theta}_i$ and $\hat{\theta}_i^{AI}$, $i = 1, \ldots, N$.

To evaluate the performance of this classification stage, we report standard metrics such as AUC (Area Under the Receiver Operating Characteristic Curve), Precision, Recall, Accuracy, and Balanced Accuracy.[29]

4.4. AI-human hybrid model

After the AI model has been trained, the AI-human hybrid model allows the firm to integrate human judgment in the hiring loop. For this, we use a standard Bayesian approach to augment and combine AI predictions with human inputs. Specifically, our AI-Human hybrid's interviewee score is a weighted average of the AI and human scores, where the weight is based on the relative precision of the AI versus human judgments.

Let $\hat{\theta}_i^{AI} = \mu(\boldsymbol{X}_i)$ be the AI-predicted score when the video feature is $\boldsymbol{X}_i$, while $h_i = \sum_{k=1}^{K} h_{ik}$ is the mean of K human scores solicited per candidate. Then the hybrid score $\hat{\theta}_i^H$ for interviewee i is a weighted average of $\mu(\boldsymbol{X}_i)$ and h_i, where the weight is based on the standard error of the AI prediction and the variance in human judgments, $\sigma_0(\boldsymbol{X}_i)$ and $\sigma_1^2(\boldsymbol{X}_i)$ respectively.[30] Specifically,

$$\hat{\theta}_i^H = \lambda(\boldsymbol{X}_i) \underbrace{\mu(\boldsymbol{X}_i)}_{\text{AI prediction from video features}} + (1 - \lambda(\boldsymbol{X}_i)) \underbrace{h_i}_{\text{Mean of human input}} \tag{4}$$

where $\lambda(\boldsymbol{X}_i) \in (0, 1)$ is the Bayesian weight on the AI prediction and is given by:

$$\lambda(\boldsymbol{X}_i) = \frac{\sigma_1^2(\boldsymbol{X}_i)/k}{\sigma_1^2(\boldsymbol{X}_i)/k + \sigma_0^2(\boldsymbol{X}_i)} \tag{5}$$

[28] We test the robustness of our results for different ranking thresholds.

[29] Since our hiring context involves an imbalanced proportion of true positives and true negatives, metrics such as AUC and Balanced Accuracy are preferred over the standard measure of accuracy, as a model that naively predicts only the majority class will have high accuracy. Let tp, tn, fp, fn be the true positive, true negative, false positive, and false negative respectively from the AI model predictions. Then, Accuracy $= \frac{tp+tn}{tp+tn+fp+fn}$. Balanced Accuracy $= \frac{\text{Sensitivity}+\text{Specificity}}{2}$, where Sensitivity $= \frac{tp}{tp+fn}$, and Specificity $= \frac{tn}{tn+fp}$.

[30] We refer to the Appendix for a detailed formulation of how our hybrid formula below arises. We note that it arises as the mean of the posterior distribution when the distribution of AI scores (the prior) is updated with human inputs drawn from a distribution of human scores. In this Bayesian updating formula, both the AI and human scores are assumed to be Normally distributed.

In general, one can use bootstrap-based inference to construct a Normally distributed confidence interval. For Random Forest (our preferred AI model), the inference distribution and $\sigma_0^2(\boldsymbol{X}_i)$ can be calculated using the Jackknife-after-Bootstrap method (Wager et al. 2014), or Bootstrap-of-Little-Bags approach (Athey et al. 2019). $\sigma_1^2(\boldsymbol{X}_i)$ is the variance of the human judgments specific to the interviewee i. Given that each interviewee is rated by multiple judges, we can identify this variance as a function of video features $\boldsymbol{X}_i$. Our hybrid model has some appealing and intuitive features; when the AI model is more confident in a prediction, human inputs will become less influential in terms of weight. This is because the weight on AI (Equation 5) becomes larger as the AI model becomes relatively more precise compared to human judgments.

To test the performance of the hybrid model, we divided our dataset into training and testing datasets. We obtained the trained AI model and hybrid weight, represented by $\mu(\cdot)$ and $\lambda(\cdot)$, respectively, from the training dataset. We then applied these to the testing dataset to obtain the AI prediction $\mu(\tilde{X}_i)$ and Bayesian weight $\lambda(\tilde{X}_i)$ for each $\tilde{X}_i$ in the test dataset.

In testing the hybrid model, we further require human inputs are drawn from judges' scores that are not part of the training or testing datasets. This is to ensure that human inputs in the hybrid do not influence the outcome variable during testing. To accomplish this, we randomly hold out k judges' scores per video (without replacement) from the testing dataset, and use the remaining judges as outcome variables. Here, k represents the number of human inputs per video. For example, in the 1-AI-human hybrid model, $k = 1$. As we have 9 judges per interview, we only considered $k = 1, 2, 3$. We used these held-out judges' scores as the human input, represented by h_i, and computed the hybrid score as $\hat{\theta}_i^H = \lambda(\tilde{X}_i)\mu(\tilde{X}_i) + (1 - \lambda(\tilde{X}_i))h_i$. Finally, we evaluated this hybrid score against the latent sales ability $\hat{\theta}_i$ in the testing dataset. This procedure ensures that the human input h_i does not influence the outcome variable, which is the latent sales ability $\hat{\theta}_i$.

We rank candidates according to their hybrid scores and calculate the accuracy and precision metrics as we did with the AI model. To account for the randomness in holding out judges' scores as human input, we repeat this process many times and take the average of the performance metrics.

Finally, we repeat this entire process over different training and testing datasets as part of our 4-fold cross-validation. We can compare directly with the pure AI model where candidates are ranked only on the basis of their AI-predicted scores.

4.5. Improvement in workforce quality from AI and AI-human hybrid models

Finally, we propose a metric to evaluate a model based on workforce quality loss to the firm when bad candidates are selected, and good candidates are not selected. We define *Average Quality Loss* as the loss in talent quality from selecting candidates that are below the bar (false positives), and not selecting candidates that are above the bar (false negatives). Specifically, the average quality loss metric is made up of the following two components, the *Average Competency Loss* and the *Average Opportunity Loss*.

i) *Average Competency Loss* is the loss in quality due to false positives (FP) —i.e., the set of selected candidates whose actual scores are below the bar (actual score threshold). Formally we write this as:

$$\text{Average Competency Loss} = \frac{1}{\#\text{Selected}} \sum_{i \in \text{FP}} \left(\text{Actual Score Threshold} - \text{Actual score}_i\right)$$

Actual Score is the candidate's latent sales ability, as defined in Section 4.2, while *Actual Score Threshold* is the 75th percentile (selection threshold) of the actual latent sales ability scores. Thus, if we plan to hire the top x percentile of candidates, this threshold will be the actual score of the top x percentile candidates ranked by their actual scores.

ii) *Average Opportunity Loss* is the loss in quality due to false negatives (FN) —i.e., the set of candidates who should have been selected i.e., those whose actual scores are above the bar (actual score threshold), but were not.

$$\text{Average Opportunity Loss} = \frac{1}{\#\text{Selected}} \sum_{i \in FN} \left(\text{Actual score}_i - \text{Actual Score Threshold}\right)$$

Overall, *Average Quality Loss* captures the total loss from the two components.

$$\begin{aligned} \text{Average Quality Loss} &= \frac{1}{\#\text{Selected}} \sum_{i \in FN \cup FP} \left|\text{Actual Score Threshold} - \text{Actual score}_i\right| \qquad (6) \\ &= \text{Average Competency Loss} + \text{Average Opportunity Loss} \end{aligned}$$

We do not need to know the predicted scores from the model to calculate these metrics, but only the ranking of these candidates according to the predicted scores. As such, we can easily benchmark our quality loss against a random permutation and ranking of candidates. For the benchmark, we randomly order the candidates (without repetition, each of the $N!$ possible ordering is equally likely, where N is the number of candidates), and calculate Equation 6 above.

5. Results

We present two sets of results. First, we compare the AI and AI-Human hybrid models on different performance dimensions. Next, we show explainability results that help understand what drives the decisions of the AI model and the performance improvement of the AI-Human hybrid.

5.1. Performance of the AI and AI-human models

We first report our models' performance in terms of binary classification metrics, which answers the managerial question: how many wrong candidates did we hire? Later in Section 5.2, we discuss the quality of the workforce selected by our models.

Table 6 compares the performance of the pure AI and AI-Human hybrid models with varying levels of human intervention. By the AI model, we refer to the best-performing machine learning model, which is the Random Forest. In Table C.1 of the appendix, we report the performance of other machine learning models such as XGBoost and SVM. We present metrics such as AUC (area under the ROC curve), accuracy, balanced accuracy, precision, recall, and "human cost for an additional good hire." The baseline column assumes a junk classifier that predicts "True" with a constant probability of q independently across candidates, where $q = 0.75$ for screening and $q = 0.25$ for selection. For example, by randomly classifying candidates into the top 25 or bottom 25, we can achieve both selection and screening accuracies of 62.5%, i.e., $(25 * 0.25 + 75 * 0.75)$.

For selection, our AI model achieves an AUC of 73.9% (baseline 50%), an accuracy of 72.2 % (baseline 62.5%), and precision and recall of 47.9% (baseline 25%). In terms of screening, the AI model achieves an AUC of 66.0%, an accuracy of 73.3%, and precision and recall of 82.4% (baseline 75%). These are significant improvements over the random benchmark. Our accuracy rates are

Table 6 Performance of AI and AI-hybrid models

Metric		Baseline	AI	AI–1 Human	AI–2 Human	AI–3 Human
AUC	Screening	50%	66.0%	81.6%	85.7%	87.8%
	Selection	50%	73.9%	82.5%	86.1%	87.8%
Accuracy	Screening	62.5%	73.3%	79.0%	81.6%	83.2%
	Selection	62.5%	72.2%	78.4%	80.9%	82.0%
Balanced Accuracy	Screening	50%	63.9%	71.6%	75.1%	77.2%
	Selection	50%	64.4%	72.4%	75.6%	76.9%
Precision	Screening	75%	82.4%	86.1%	87.8%	88.9%
	Selection	25%	47.9%	59.8%	64.3%	66.2%
Recall	Screening	75%	82.4%	86.1%	87.8%	88.9%
	Selection	25%	47.9%	59.8%	64.3%	66.2%
Cost per good hire		–	–	**$840**	$1219	$1639

Note: Precision and recall values are the same. This is because our binary classifier is derived from the thresholding of a continuous predicted variable, such that the numbers of false positives and false negatives are set to be equal.

comparable to other benchmarks within the class of ML problems seeking to predict "subjective human response" using video or other unstructured data, where accuracy is in the range of 55% to 70%.[31] This is in contrast to the class of objective "recognition" problems where the ground truth can be objectively defined and measured, and where accuracy is much higher, typically in the range of 80-95%.[32]

Further, in Figure C.1 in the appendix, we investigate how the size of the training dataset affects the predictive performance of the model. Companies will likely see more accurate predictions as they scale up the number of training videos. For instance, if the number of videos is increased 5-fold from 200 (currently available) to 1000, then our calculation shows that accuracy would increase from 74% to 82%.

[31] For these problems, the ground truth is a subjective human response such as persuasiveness (Chen et al. (2009)), informativeness (Timoshenko and Hauser 2019), personality traits (Chen et al. (2017)), fine-grained sentiment (Brahma 2018), and emotions (Fong et al. (2021)).

[32] These include, object detection in static images (Zou et al. 2023), action recognition from videos (e.g. different sports as in Karpathy et al. (2014), musical instruments being played as in Soomro et al. (2012)) and human pose recognition (e.g., AlphaPose, OpenPose, DeepLabCut).

We find that adding just one human expert rating in the loop greatly increases predictive accuracy for both selection and screening. Compared to the pure AI benchmark, adding just one human input increases the AUC by approximately 10 percentage points. From Table 6, we also see that increasing the number of human inputs generally increases the performance of the hybrid model. However, after the initial jump, performance increases very slowly with additional human input.

Although adding humans in the hiring loop improves performance, it is also costly. To quantify the cost-performance trade-off of the hybrid model, we calculate the *cost of human evaluation for an additional good hire*. Suppose a firm wants to fill 25 salesforce positions from 100 interviewees. With AI, 11.97 out of the 25 selected candidates would be good hires (selection precision of 47.9%). This is an improvement over random selection, where only 6.25 out of 25 are good hires. However, the hybrid model is even more effective, resulting in 14.95/25 good hires, or 3 additional good hires over AI. For the AI–1 Human hybrid model, the additional human cost of evaluating 100 candidates is $2,500;[33] resulting in a human cost of $840 for each additional good hire. Based on our calculations here, a company can trade off the cost of Type-II error (the candidate is not good but we hire him/her) with the marginal cost of the hybrid model.[34]

Overall, we see that the marginal human cost for the AI–1 Human hybrid model is the lowest. The optimal number of human inputs in the hybrid model is just one – the gains from increased accuracy with more human input are not worth the incremental cost. When we tried different levels of selection thresholds, from 70th to 85th percentiles, our results from Table 6 remained similar.

5.2. Workforce quality

Using the metrics we proposed in Section 4.5, we quantify the impact on workforce quality loss, competency loss (from false positives), and opportunity loss (from false negatives) in Table 7. Our

[33] Assuming a cost of $100 per hour for human evaluation and an average of 15 minutes to evaluate each interview video, the cost of human input per video is $25. This is a conservative estimate based on experienced sales professionals whose typical hourly salaries range from $100-$150. Thus, if we use k human inputs per candidate in the hybrid model (as in the AI–k Human hybrid model), the cost of evaluating 100 candidates would be $k \times 100 \times 25 = 2500k$.

[34] If the cost of Type-II error is high, then the firm may find it optimal to have more than 1 human input per video. Further, we do not consider the cost of Type-I error here (rejecting a good candidate).

AI model improves the quality of the workforce relative to random by about 40%. The addition of 1 human input in the AI-human hybrid improves quality relative to random by 67%. Overall, we find the magnitude of competency loss (false positives), exceeds the loss from opportunity loss (false negatives) for the AI model. The inclusion of humans in the hybrid model improves both losses, but also makes these two losses more comparable. Overall, we conclude that the models substantially improve workforce quality.

Table 7 **Quantifying losses to the firm in terms of the quality of salesforce.**

Metric	Random	AI	AI–1 Human	AI–2 Human	AI–3 Human
Average Quality Loss	5.376	3.188 (41% gain)	1.788 (67%)	1.512 (72%)	1.352 (75%)
Average Competency Loss	3.208	1.988 (38% gain)	0.944 (71%)	0.732 (77%)	0.644 (80%)
Average Opportunity Loss	2.172	1.200 (45% gain)	0.844 (61%)	0.780 (64%)	0.708 (67%)

Note: The random benchmark is based on a random permutation of the candidates.

5.3. Classifications under the AI and hybrid models

In Figure D.1a of the appendix, we plot the AI scores against the actual scores to visualize the classification. Since we performed 4-fold cross-validation, this scatterplot corresponds to the out-of-sample prediction, of a particular split of the dataset into 75% training dataset and 25% testing dataset. The horizontal line is the threshold for true selection; the vertical line is the model-predicted threshold for selection.

Figure D.1b shows the corresponding scatter plot for the hybrid model. As can be seen, the hybrid improves performance considerably. The false positives are pushed to the bottom-left quadrant and false negatives are pushed to the top-right quadrant. We explore the details of how the hybrid improves classification in the section on explainability in Section 5.4.2.

5.4. Model Explanations

We first discuss the drivers of performance of the AI model. Then we describe what drives the improvement of the hybrid model over the AI model.

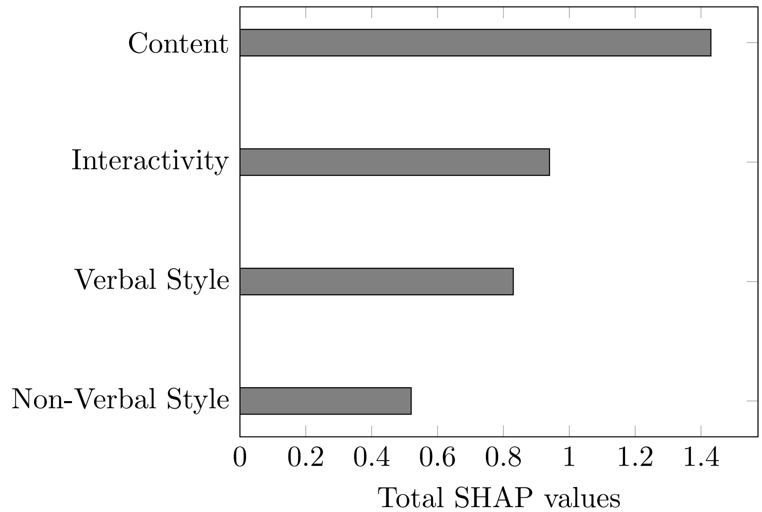
5.4.1. What drives the performance of the AI model? SHapley Additive exPlanations (SHAP) are used to explain supervised learning models, by quantifying the impact of a particular feature (or group of features) on the model outcome (Lundberg and Lee 2017). SHAP values are based on Shapley values, a concept from cooperative game theory, where it is used to quantify how much each player in the game contributes to the pie. In the context of machine learning, SHAP values help to explain how much each feature contributes to explain the prediction of a supervised model. The "game" here is at the level of a single observation and thus provides local (observation level) explainability. By aggregating SHAP values of a feature across observations (mean of absolute values), one can explain the global impact of the features across all data. To obtain a measure of the relative global impact of feature groups (e.g., content vs. interactivity), we aggregate SHAP values not only across observations but also across features within the group (sum of feature importances within the feature group).

We report the impact of feature groups in Figure 5. The *Content* group of features (e.g., information density, time spent on different topics) is the most important followed by *Interactivity* (e.g., active listening), and *Verbal style* (e.g., linguistic style and tones that reflect collaborativeness and politeness). While *Non-Verbal style*, which captures body movement is less important than others overall, certain features within this group are important e.g., hand gestures and distance between hands. These results suggest that capturing elements of interactivity in two-way conversations and body language adds significant incremental predictive value beyond just textual content.

Next, we present a scatter plot of SHAP values at the observational level for features found to be important, in Figure 6. For each of the four feature groups, we present the SHAP values for the three top features within each feature group. The scatter plots are overlaid with the histogram of feature values, to help see how the feature values are distributed. All features have been standardized and scaled, so a feature value of x means x standard deviation from the mean.

We discuss some of the managerial insights from Figure 6. Under content features, the most important is *Information Density* – the number of words spoken by the seller during the 15-minute interview. A below-average information density leads to a sharp decline in the predicted

Figure 5 Impact of Feature Groups



latent sales ability, suggesting that an unprepared salesperson with little to say tends to have poor performance. However, an increase in information density beyond the median does not have much incremental benefit. Comparing *Business* vs *Product*, we see that spending more time on business topics (customers' needs) increases the predicted latent sales ability, but a greater focus on the product's features reduces it.

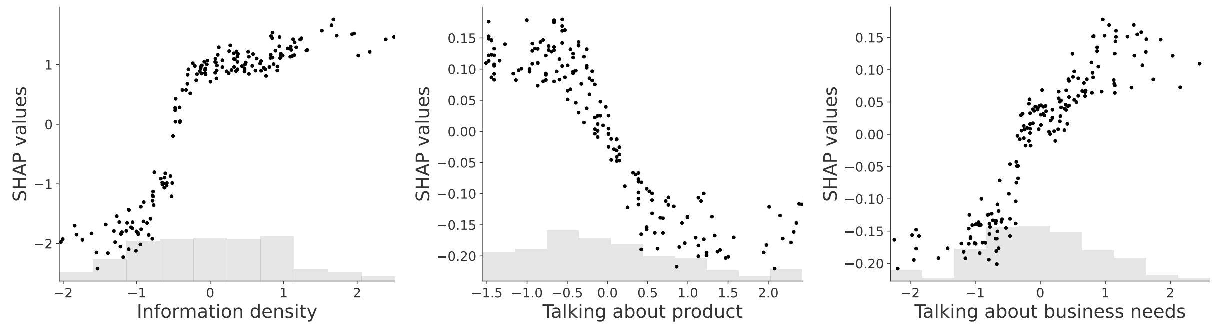
For features related to conversational interactivity, it is important for sellers to give buyers the opportunity to speak and provide their input, but not to the extent that buyers dominate the conversation and sellers do not contribute enough. This is evident in the graph of the ratio of *Buyer-to-Seller Share of Voice*. In addition, we find that *Active Listening* score is an important predictor of seller 's performance. A higher *Active Listening* score leads to stronger seller's performance. The plot also shows the discrete nature of GPT's score, for example, GPT tends to give scores rounded to the nearest integer or 0.5. Lastly, *Style matching* – the ability of the seller to adapt and mimic the buyer in terms of the languages used – is also an important feature that leads to more positive performance.

For verbal style, using more *Complex* words increases the predicted latent sales ability, but there is little additional benefit beyond a threshold.[35] Among the linguistic variables constructed from

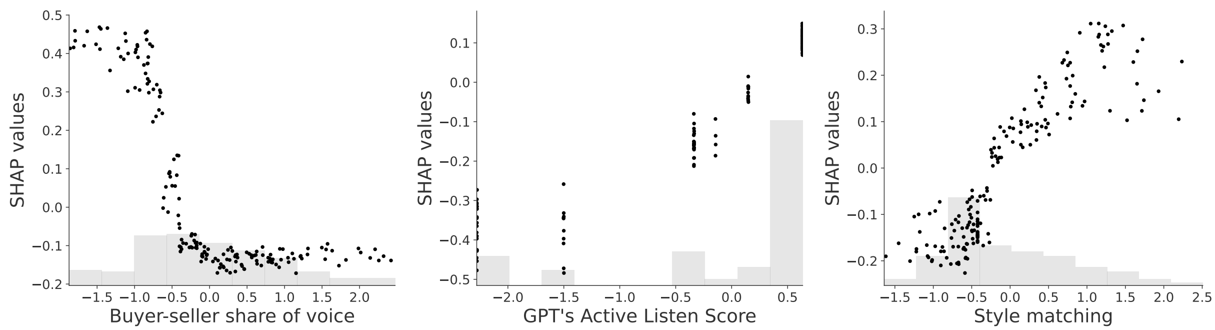
[35] Among the three measures of language complexity – Flesch–Kincaid readability, Type-Token ratio, and *LongWords* (the percentage of seller's words that have seven or more letters), we find that *LongWords* matters the most.

Figure 6 SHAP values for the most important features

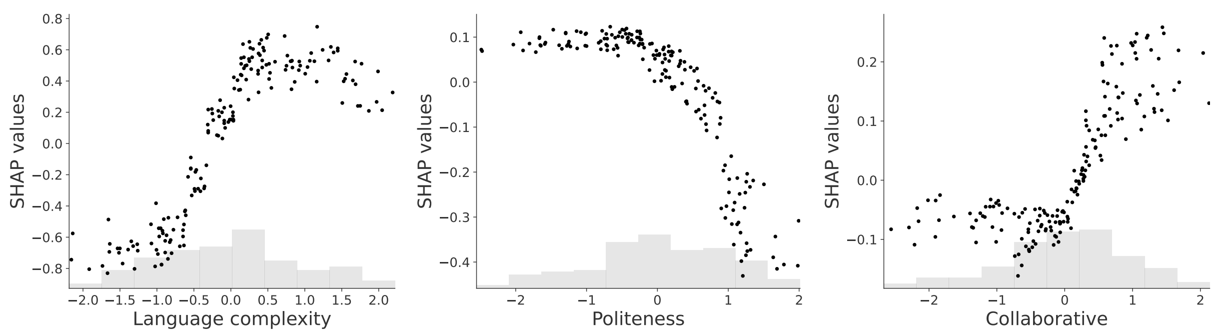
(a) Content features



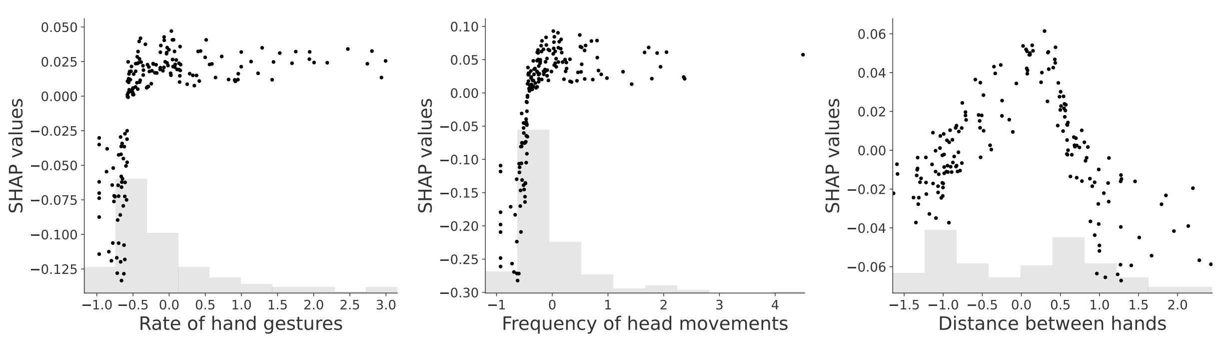
(b) Conversational Interactivity features



(c) Verbal style features



(d) Non-verbal style features



text embedding, *Politeness* and *Collaborativeness* stand out as more important than others. A more collaborative linguistic style explains better seller's performance. However, *Politeness* has a non-linear effect–being overly polite adversely affects seller performance. This may be because over-politeness can come across as artificial, and insincere; excessive politeness can also slow down the conversation too much, leading to less information information transfer.

Finally, for non-verbal style features that involve body language, a low *Rate of Hand Gestures* (being stiff and not animated) negatively impacts the seller. A high rate of hand gestures increases predicted sales ability up to a point, beyond which the impact becomes limited. So, a seller should use a moderate amount of hand gestures. Similarly, maintaining an active body language in the form of high *Frequency of Head Movement* has a positive impact, but too much movement does not help. Additionally, a moderate amount of *Distance between the left and right hands* (indicating openness) is optimal; too close or too wide a distance has a negative impact.

In terms of including demographic variables that we would usually find in a candidate's resume, such as the candidate's education and past industry experience – candidates in our dataset are business majors from the same university and of similar ages. In addition, they do not possess work experience in B2B sales. This is a typical scenario in campus recruiting and entry-level hiring.

Given the limitations of demographic variables, we examine only the role of one demographic factor: gender. To do this, we group the SHAP values by men vs. women. For each feature, we calculate the average absolute SHAP values across men vs. women and take the ratio of the two. This ratio measures the relative importance of a feature for males vs. females. A large ratio means that this feature is more important for males than for females. In Table C.2, we show that while head movement, active listening, and hand gestures are relatively more important for men; collaborativeness, politeness, and information density are relatively more important for women. Qualitatively, these features all have the same directional impact whether the candidate is a male or female.

5.4.2. What drives the improvement in the hybrid model? Earlier in Table 6, we see that the hybrid model increases the AUC of selection from 73.9% to 82.5%. How much of this improvement is due to the mere availability of human input, and how much of this improvement can be attributed to the interaction and synergy between human input and AI? To answer this question, we consider a hybrid model where we put zero weight on the AI, and all the weight on the human input. This is the "only human" scenario.

We summarize the decomposition of the improvement from human and the AI-human hybrid in Figure D.2 in the appendix. The darker-shade bar represents the incremental AUC of having one human input ("only human") over AI, while the lighter-shade bar represents the remaining incremental AUC gain of the hybrid model over AI, relative to the incremental accuracy of "only human." The numbers in each bar are the percentage incremental AUC attributed to each block.

"Only human" accounts for 78% of the improvement in AUC of the hybrid over AI; the remaining 22% of the improvement can be attributed to the interaction and combination of AI and human input in the hybrid model. Thus the AI model helps to augment the human input to improve accuracy. Not surprisingly, "only human" does quite well, but it is costly. The AI component helps *achieve an additional 22% improvement* by combining the costly human input with AI predictions, which come at no marginal human cost. Thus the hybrid model is far more cost-effective than "only human." Overall, across multiple metrics, around a quarter of the gain in hybrid over AI can be attributed to how the hybrid model combines AI and human input through Bayesian averaging.

Finally, we seek to explain what drives the gap between AI and the AI-human hybrid. For this, we see what factors drive the absolute difference between hybrid and AI scores using SHAP values. We calculated the hybrid and AI scores, along with their absolute difference, for each observation in the k-fold testing dataset. Based on the SHAP values, the most important feature explaining the difference is the distance between hands (Figure D.3 in the appendix). The hybrid model deviates more from the AI model when the distance between hands is smaller. Holding one's hands close together can have various interpretations – formality, listening attentively, or nervousness. While humans can easily interpret the cues and contexts, it may be difficult for AI to assess the meaning of holding one's hands close together.

5.5. Cost effective hybrids: Task-based and Sequential

Finally, we ask if we can make the models more cost-effective by reducing the need for human intervention while keeping accuracy at reasonable levels. We propose two types of cost-effective hybrid models in line with the taxonomy in Shrestha et al. (2019). The first is a *Task-based* hybrid model where we use human input to judge only certain tasks or stages of the conversation. The second is a *Sequential* hybrid model where AI is used in the first stage for screening followed by a hybrid model for selecting the top candidates amongst the screened pool.

To understand how each of these models reduces cost, let us suppose that there are 100 candidates and k human input per candidate, then the cost of human input becomes $100 \times k \times c$, where c is the cost of evaluating one interview. In Section 5.1, we have assumed a cost of $c = \$25$. Instead, if humans are only used for a fraction t of the overall time (as in task-based hybrid), then the cost of human input reduces to $100 \times k \times c \times t$. Now in a *Sequential* hybrid model where we use AI to screen out candidates below the p-percentile, and augment the screened candidates with human input, the cost of human input reduces to $(100 - p) \times k \times c$.

Task-Based hybrid Models. The interview has 5 stages as described in Table 2b. In the task-based hybrid model, we augment AI with human judges' scores from specific stages. We examine whether using only scores from the initial stages as human input to the hybrid model would be more cost-effective. We find that using the first two stages leads to the most cost-effective hybrid model, with the cost per additional good hire reducing from \$840 (Table 6) to \$648 (Table D.1).

Table D.1 in the appendix shows that most of the benefit of including humans in the loop is realized by the second stage i.e., Approach (effectively gains attention and builds rapport) along with Needs Identification (obtains a clear understanding of customer's situation). This finding is important as it means that without losing too much accuracy, we can reduce human evaluation time and cost by more than half as humans only need to evaluate the first 7 minutes of an interview, instead of the entire 15-minute conversation. Interestingly, asking humans to evaluate only the first stage (Approach) would lead to a huge loss in performance. This means that first impressions by humans can be very misleading, and it is important for the human judge to go beyond the introduction stage.

Sequential hybrid model In the Sequential hybrid model, we reduce the cost of human input by only soliciting human judgment for a subset of candidates, specifically, we solicit human judgment for candidates above a threshold. We first use AI to screen out the bottom candidates, then subsequently augment the AI predictions with human judgments for candidates remaining in the second stage. Since we only use the sequential hybrid model in the context of selection, we report precision and accuracy metrics pertaining to selection and not screening.

Table D.2 in the appendix shows the main result for the sequential hybrid model. We vary the screening thresholds from the 5th to 70th percentile. A screening threshold of κ-percentile means that human judgments are only required for candidates ranked above the κ-percentile according to AI. When the screening threshold is zero, the Sequential hybrid model is equivalent to the original hybrid model. There is no cost-saving when the screening threshold is zero.

As the screening threshold increases, cost-saving increases as human input is required only for candidates above the screening threshold. However, the concern is that there is a trade-off between cost savings and performance. Interestingly, Table D.2 shows that a sequential hybrid model with a screening threshold of 50th percentile is the most cost-effective. There is not much loss in performance when the threshold is 50%, i.e. human inputs are only required for 50% of the candidates. But above the screening threshold of 50%, the loss in performance is proportionally greater than the cost-saving. Under this Sequential hybrid model, the marginal human cost of a good hire is reduced from $840 to $505. Overall, the sequential hybrid is more cost-effective than the task-based hybrid, which in turn, is more cost-effective than the full hybrid.

6. Conclusion

AI is increasingly used in hiring decisions; but most work has been in the area of resume screening. As remote work and online hiring gain relevance, video interviews have become an integral part of the recruitment process. In this paper, we developed an AI and AI-human hybrid system for hiring using video recordings of conversational interviews. Our application is in the context of B2B salesforce recruitment at the college level, where the candidate pool is fairly homogeneous in terms of education and experience. As such, resume-based screening is insufficient for differentiating candidates, hence the need for interviews that are interactive in nature to assess sales skills at scale.

Since hiring is a “high risk” AI application and laws require explainability of the decisions made, we extract persuasion theory-relevant, objective features of candidates' sales performance from the visual, verbal, and voice modalities of video data. We then use these to predict candidates' latent salesforce skills. Unlike past machine learning applications using videos, a key contribution here is that we show how to capture features around *two-way conversational interactivity*, *adaptation*, and *body language*. Finally, to the extent that humans can identify aspects of ability not captured by the AI model, we illustrate how to incorporate human input in an AI-human hybrid hiring model.

Our AI model provides comparable performance relative to other human subjective response prediction AI models. While the content of what is spoken is most important in predicting sales skills, conversational interactivity, active listening, linguistic styles, and elements of body language also have good explanatory power.

We conclude with a discussion of limitations and suggestions for future work. Research on persuasion and sales influence has traditionally relied on survey-based recall. This paper demonstrates how we can construct objective, real-time metrics from recorded interactions with minimal measurement error. As such, our approach can be generally valuable for richer and more effective theory testing of factors impacting persuasion and sales influence.

While this is a first step towards understanding how conversational interactivity and body language impact persuasion, the fact that our study shows human input can significantly enhance AI suggests there is scope for improvement in the AI model. This finding echoes the prevailing sentiment that there is a gap in AI's current ability to comprehend nonverbal communications (hand gestures, body language, and tone of voice) which extends beyond mere textual interaction. There is a clear need for further development in AI with the goal of an Artificial General Intelligence that can comprehend nonverbal communications. We hope our work serves as an impetus for a research agenda around better use of video data in sales and marketing more broadly.

Finally, our training data is relatively small. Firms implementing a hiring tool like ours would have access to a larger number of videos. If firms scale up to a large number of videos, the accuracy of the AI model would increase and become more reliable. Potentially, the reliability of the AI model would be high enough to diminish the role of a AI-human hybrid model. Thus, firms face an interesting trade-off between collecting more training data early on versus relying on a hybrid approach of human in the loop. Either way, a higher accuracy would translate to greater cost savings for the firm as it scales up the implementation of our framework. Being able to run a large number of interviews efficiently and as cost-effectively as possible would lead to substantial long-term cost savings for the firm.

Funding and Competing Interests

Authors B and C gratefully acknowledge partial financial support from the NEC Foundation of America (https://www.necfoundation.org/) grant.

References

Arora, S., A. Narayan, M. F. Chen, L. Orr, N. Guha, K. Bhatia, I. Chami, F. Sala, and C. Ré (2022). Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*.

Ashnai, B., S. Mani, P. Kothandaraman, and S. Shekari (2020). Gender bias in the recruitment of entry-level b2b salespeople. *Journal of Business & Industrial Marketing 35*(8), 1335–1344.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Black, J. S. and P. van Esch (2020). Ai-enabled recruiting: What is it and how should a manager use it? *Business Horizons 63*(2), 215–226.

Brahma, S. (2018). Improved sentence modeling using suffix bidirectional lstm. *arXiv preprint arXiv:1805.07340*.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems 33*, 1877–1901.

Camp, R., E. Schulz, M. Vielhaber, and F. Wagner-Marsh (2004). Human resource professionals' perceptions of team interviews. *Journal of Managerial Psychology*.

Campion, M. A., D. K. Palmer, and J. E. Campion (1997). A review of structure in the selection interview. *Personnel psychology 50*(3), 655–702.

Cao, Z., G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chang, H. H., A. Mukherjee, and A. Chattopadhyay (2022). More voices persuade: The attentional benefits of voice numerosity. *Journal of Marketing Research*.

Chartrand, T. L. and J. A. Bargh (1999). The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology 76*(6), 893.

Chen, L., G. Feng, C. W. Leong, B. Lehman, M. Martin-Raugh, H. Kell, C. M. Lee, and S.-Y. Yoon (2016). Automated scoring of interview videos using doc2vec multimodal feature extraction paradigm. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, New York, NY, USA, pp. 161–168. Association for Computing Machinery.

Chen, L., R. Zhao, C. W. Leong, B. Lehman, G. Feng, and M. E. Hoque (2017). Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 504–509. IEEE.

Chen, X.-P., X. Yao, and S. Kotha (2009). Entrepreneur passion and preparedness in business plan presentations: a persuasion analysis of venture capitalists' funding decisions. *Academy of Management journal 52*(1), 199–214.

Churchill, G. A., R. H. Collins, and W. A. Strang (1975). Should retail salespersons be similar to their customers. *Journal of Retailing 51*(3), 29.

Cialdini, R. B. (1987). *Influence*, Volume 3. A. Michel Port Harcourt.

Cialdini, R. B. and N. J. Goldstein (2004). Social influence: Compliance and conformity. *Annual review of psychology 55*(1), 591–621.

Conway, J. M., R. A. Jako, and D. F. Goodman (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of applied psychology 80*(5), 565.

Criteriacorp (2022). video-interviewing-in-2022-and-beyond.

Danescu-Niculescu-Mizil, C. and L. Lee (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077*.

Danescu-Niculescu-Mizil, C., L. Lee, B. Pang, and J. Kleinberg (2012). Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pp. 699–708.

Danescu-Niculescu-Mizil, C., M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts (2013). A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Dubinsky, A. J. (1981). A factor analytic study of the personal selling process. *Journal of personal selling & sales management 1*(1), 26–33.

Erickson, B., E. A. Lind, B. C. Johnson, and W. M. O'Barr (1978). Speech style and impression formation in a court setting: The effects of "powerful" and "powerless" speech. *Journal of experimental social psychology 14*(3), 266–279.

Evans, F. B. (1963). Selling as a dyadic relationship–a new approach. *American Behavioral Scientist 6*(9), 76–79.

Fast, J. (1970). *Body language*, Volume 82348. Simon and Schuster.

Flesch, R. (2007). Flesch-kincaid readability test. *Retrieved October 26*(3), 2007.

Fong, H., V. Kumar, and K. Sudhir (2021). A theory-based interpretable deep learning architecture for music emotion. *Available at SSRN: https://ssrn.com/abstract=4025386*.

Frazier, G. L. and J. O. Summers (1984). Interfirm influence strategies and their application within distribution channels. *Journal of Marketing 48*(3), 43–55.

Hammer, F., P. Reichl, and A. Raake (2004). Elements of interactivity in telephone conversations. In *Eighth International Conference on Spoken Language Processing*.

Hawes, J. M. and G. A. Rich (1998). Selling and sales management in action: The constructs of sales coaching: Supervisory feedback, role modeling and trust. *Journal of Personal Selling & Sales Management 18*(1), 53–63.

Holstein, K., V. Aleven, and N. Rummel (2020). A conceptual framework for human–ai hybrid adaptivity in education. In *International Conference on Artificial Intelligence in Education*, pp. 240–254. Springer.

Hwang, S., X. Liu, and K. Srinivasan (2021). Voice analytics of online influencers—soft selling in branded videos. *Available at SSRN 3773825*.

Ireland, M. E. and J. W. Pennebaker (2010). Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology 99*(3), 549.

Kahneman, D., A. Rosenfield, L. Gandhi, and T. Blaser (2016). Noise. *Harvard Bus Rev*, 38–46.

Karpathy, A., G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732.

Kendon, A. (1994). Do gestures communicate? a review. *Research on language and social interaction 27*(3), 175–200.

Kocoń, J., I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, A. Kocoń, B. Koptyra, W. Mieleszczenko-Kowszewicz, P. Miłkowski, M. Oleksy, M. Piasecki, Ł. Radliński, K. Wojtasik, S. Woźniak, and P. Kazienko (2023, nov). ChatGPT: Jack of all trades, master of none. *Information Fusion 99*, 101861.

Krauss, R. M., P. Morrel-Samuels, and C. Colasante (1991). Do conversational hand gestures communicate? *Journal of personality and social psychology 61*(5), 743.

Le Mens, G., B. Kovács, M. T. Hannan, and G. Pros (2023). Uncovering the semantics of concepts using gpt-4 and other recent large language models. Technical report.

Levashina, J., C. J. Hartwell, F. P. Morgeson, and M. A. Campion (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology 67*(1), 241–293.

Lichtenthal, J. D. and T. Tellefsen (2001). Toward a theory of business buyer-seller similarity. *Journal of Personal Selling & Sales Management 21*(1), 1–14.

Loe, T. W. and L. B. Chonko (2000). Promoting sales programs: The national collegiate sales competition. *Journal of Personal Selling & Sales Management 20*(1), 11–13.

Longpre, L., E. Durmus, and C. Cardie (2019). Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*, pp. 167–176.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems 30*.

Maree, M., A. B. Kmail, and M. Belkhatir (2019). Analysis and shortcomings of e-recruitment systems: Towards a semantics-based approach addressing knowledge incompleteness and limited domain coverage. *Journal of Information Science 45*(6), 713–735.

Markowitz, D. M. (2020). Putting your best pet forward: Language patterns of persuasion in online pet advertisements. *Journal of Applied Social Psychology 50*(3), 160–173.

McDaniel, M. A., D. L. Whetzel, F. L. Schmidt, and S. D. Maurer (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of applied psychology 79*(4), 599.

Moncrief, W. C. and G. W. Marshall (2005). The evolution of the seven steps of selling. *Industrial Marketing Management 34*(1), 13–22.

Naim, I., M. I. Tanveer, D. Gildea, and M. E. Hoque (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, Volume 1, pp. 1–6. IEEE.

Pease, B. and A. Pease (2008). *The definitive book of body language: The hidden meaning behind people's gestures and expressions*. Bantam.

Pennebaker, J. W., M. E. Francis, and R. J. Booth (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates 71*(2001), 2001.

Pogacar, R., L. Shrum, and T. M. Lowrey (2018). The effects of linguistic devices on consumer information processing and persuasion: a language complexity× processing mode framework. *Journal of Consumer Psychology 28*(4), 689–711.

Posthuma, R. A., F. P. Morgeson, and M. A. Campion (2002). Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *Personnel Psychology 55*(1), 1–81.

Pulakos, E. D. and N. Schmitt (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology 48*(2), 289–308.

Rackham, N. (2020). *SPIN®-selling*. Routledge.

Rajaram, P. and P. Manchanda (2020). Video influencers: Unboxing the mystique. *arXiv preprint arXiv:2012.12311*.

Reimers, N. and I. Gurevych (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rich, G. A. (1999). Salesperson optimism: can sales managers enhance it and so what if they do? *Journal of Marketing Theory and Practice 7*(1), 53–63.

Rivera, L. (2015). Firms are wasting millions recruiting on only a few campuses. *Harvard Business Review*, 1–8.

Rivera, L. A. (2012). Hiring as cultural matching: The case of elite professional service firms. *American sociological review 77*(6), 999–1022.

Sacks, H., E. A. Schegloff, and G. Jefferson (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pp. 7–55. Elsevier.

Sheth, J. (1976). Buyer-seller interaction: A conceptual framework. *Advances in Consumer Rasearch 3*(3B), 382—386.

Shmueli-Scheuer, M., J. Herzig, D. Konopnicki, and T. Sandbank (2019). Detecting persuasive arguments based on author-reader personality traits and their interaction. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 211–215.

Shrestha, Y. R., S. M. Ben-Menahem, and G. Von Krogh (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review 61*(4), 66–83.

Smith, C. (2019). An employee's best friend? how ai can boost employee engagement and performance. *Strategic HR Review 18*(1), 17–20.

Soomro, K., A. R. Zamir, and M. Shah (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Spiro, R. L. and B. A. Weitz (1990). Adaptive selling: Conceptualization, measurement, and nomological validity. *Journal of marketing Research 27*(1), 61–69.

Syam, N. B., J. D. Hess, and Y. Yang (2013). Sales contests versus quotas with imbalanced territories. *Marketing Letters 24*(3), 229–244.

Tambe, P., P. Cappelli, and V. Yakubovich (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review 61*(4), 15–42.

Taylor, P. J. and S. Thomas (2008). Linguistic style matching and negotiation outcome. *Negotiation and conflict management research 1*(3), 263–281.

Timoshenko, A. and J. R. Hauser (2019). Identifying customer needs from user-generated content. *Marketing Science 38*(1), 1–20.

Vamvourellis, D., M. Toth, S. Bhagat, D. Desai, D. Mehta, and S. Pasquali (2023). Company similarity using large language models. *arXiv preprint arXiv:2308.08031*.

van den Broek, E., A. Sergeeva, and M. Huysman (2021). When the machine meets the expert: An ethnography of developing ai for hiring. *MIS Quarterly 45*(3).

Van Zant, A. B. and J. Berger (2020). How the voice persuades. *Journal of personality and social psychology 118*(4), 661.

Wager, S., T. Hastie, and B. Efron (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research 15*(1), 1625–1651.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, X., S. Lu, X. Li, M. Khamitov, and N. Bendle (2021). Audio mining: the role of vocal tone in persuasion. *Journal of Consumer Research 48*(2), 189–211.

Widmier, S. M., T. Loe, and G. Selden (2007). Using role-play competition to teach selling skills and teamwork. *Marketing Education Review 17*(1), 69–78.

Wilson, S. R., M.-S. Kim, and H. Meischke (1991). Evaluating brown and levinson's politeness theory: A revised analysis of directives and face. *Research on Language & Social Interaction 25*(1-4), 215–252.

Woodside, A. G. and J. W. Davenport Jr (1974). The effect of salesman similarity and expertise on consumer purchasing behavior. *Journal of Marketing Research 11*(2), 198–202.

Wotruba, T. R., E. K. Simpson, and J. L. Reed-Draznik (1989). The recruiting interview as perceived by college student applicants for sales positions. *Journal of Personal Selling & Sales Management 9*(3), 13–24.

Xiao, L. and T. Khazaei (2019). Changing others' beliefs online: Online comments' persuasiveness. In *Proceedings of the 10th International Conference on social media and Society*, pp. 92–101.

Yang, J., J. Zhang, and Y. Zhang (2021). First law of motion: Influencer video advertising on tiktok. *Available at SSRN 3815124*.

Zoghaib, A. (2019). Persuasion of voices: The effects of a speaker's voice characteristics and gender on consumers' responses. *Recherche et Applications en Marketing (English Edition) 34*(3), 83–110.

Zou, Z., K. Chen, Z. Shi, Y. Guo, and J. Ye (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*.

Appendix for "AI and AI-human based Salesforce Hiring using Conversational Interview Videos"

These materials have been supplied by the authors to aid in the understanding of their paper.

Appendix A: Details of the AI-human hybrid formula

This section shows how our hybrid formula is derived from a standard Bayesian approach. Let Equation 7 below describe the distribution of the AI predictions, which is assumed to be Normally distributed.

$$\hat{\theta}_i|\boldsymbol{X}_i \sim \mathcal{N}(\mu(\boldsymbol{X}_i), \sigma_0^2(\boldsymbol{X}_i)) \tag{7}$$

$\mu(\boldsymbol{X}_i)$ is the AI-predicted score when the video feature is $\boldsymbol{X}_i$ and $\sigma_0^2(\boldsymbol{X}_i)$ is the standard error or uncertainty of that prediction. We use bootstrap-based inference to construct a Normally distributed confidence interval. For our preferred AI model, the Random Forest, this inference distribution in Equation 7, can be calculated using the Jackknife-after-Bootstrap method of Wager et al. (2014), or Bootstrap-of-Little-Bags approach (Athey et al. (2019)). We can think of this as our prior distribution of the candidate's latent sales ability, and we wish to update this prior distribution given a new signal that is informative about the candidate's ability.

Similarly, Equation 8 below describes the distribution of human scores h_i, conditional on $\hat{\theta}_i$, $\boldsymbol{X}_i$

$$h_i|\hat{\theta}_i, \boldsymbol{X}_i \sim \mathcal{N}(\hat{\theta}_i, \sigma_1^2(\boldsymbol{X}_i)) \tag{8}$$

Since each candidate is rated by multiple judges, we can identify $\sigma_1^2(\boldsymbol{X}_i)$. Specifically, we take $\hat{\sigma}_i^2$ to be the sample variance of the observed judges' scores for each candidate i. We then obtain $\sigma_1^2(\boldsymbol{X}_i)$ from $\hat{\sigma}_i^2$ via an auxiliary random forest regression of $\hat{\sigma}_i^2$ on $\boldsymbol{X}_i$.

Given these distributions of AI and human scores (Equations 7 and 8 respectively), the hybrid score is:

$$\text{hybrid score} = \left(\frac{\mu(\boldsymbol{X}_i)}{\sigma_0^2(\boldsymbol{X}_i)} + \frac{h_i}{\sigma_1^2(\boldsymbol{X}_i)}\right) \Bigg/ \left(\frac{1}{\sigma_0^2(\boldsymbol{X}_i)} + \frac{1}{\sigma_1^2(\boldsymbol{X}_i)}\right) \tag{9}$$

This formula above arises from a standard Bayesian updating procedure where the AI prediction (Equation 7) is updated with human inputs realized from Equation 8, which results in a Normally distributed posterior distribution. The formula in 9 is exactly the mean of this posterior distribution. Let $f_{AI}(\hat{\theta}_i|\boldsymbol{X}_i)$ denote Equation 7 and $f_h(h_i|\hat{\theta}_i, \boldsymbol{X}_i)$ denote Equation 8. Upon obtaining a human input h_i drawn from $f_h(h_i|\hat{\theta}_i, \boldsymbol{X}_i)$, we update the AI predictions via Bayesian updating: $f(\hat{\theta}_i|h_i, \boldsymbol{X}_i) \propto f_h(h_i|\hat{\theta}_i, \boldsymbol{X}_i) f_{AI}(\hat{\theta}_i|\boldsymbol{X}_i)$.

This posterior distribution $f(\hat{\theta}_i|h_i, \boldsymbol{X}_i)$ is also a Normal distribution, whose mean is given by Equation 9. That is, $\mathbb{E}[\hat{\theta}_i|h_i, \boldsymbol{X}_i] = \left(\frac{\mu(\boldsymbol{X}_i)}{\sigma_0^2(\boldsymbol{X}_i)} + \frac{h_i}{\sigma_1^2(\boldsymbol{X}_i)}\right) \Big/ \left(\frac{1}{\sigma_0^2(\boldsymbol{X}_i)} + \frac{1}{\sigma_1^2(\boldsymbol{X}_i)}\right)$.

We can rewrite Equation 9 such that our hybrid score can be expressed as a weighted average of the AI prediction and the human intervention, i.e. our proposed hybrid score is:

$$\hat{\theta}_i^H = \lambda(\boldsymbol{X}_i) \underbrace{\mu(\boldsymbol{X}_i)}_{\text{AI prediction from video features}} + (1 - \lambda(\boldsymbol{X}_i)) \underbrace{h_i}_{\text{Human prediction}} \tag{10}$$

$\lambda(\boldsymbol{X}_i) \in (0, 1)$ is the Bayesian weight on the AI prediction and is given by:

$$\lambda(\boldsymbol{X}_i) = \frac{\sigma_1^2(\boldsymbol{X}_i)}{\sigma_1^2(\boldsymbol{X}_i) + \sigma_0^2(\boldsymbol{X}_i)} \tag{11}$$

More generally, if there are k independent human inputs, we can take the average to form h_i. Equation 8 becomes $h_i|\hat{\theta}_i, \boldsymbol{X}_i \sim \mathcal{N}(\hat{\theta}_i, \frac{\sigma_1^2(\boldsymbol{X}_i)}{k})$, and the Bayesian weight becomes $\lambda(\boldsymbol{X}_i) = \frac{\sigma_1^2(\boldsymbol{X}_i)/k}{\sigma_1^2(\boldsymbol{X}_i)/k + \sigma_0^2(\boldsymbol{X}_i)}$.

In addition, we could add a hyper-parameter to the prior distribution in Equation 7. In particular, $\hat{\theta}_i|\boldsymbol{X}_i \sim \mathcal{N}(\mu(\boldsymbol{X}_i), \tau\sigma_0^2(\boldsymbol{X}_i))$, where τ is an unknown hyper-parameter of the prior. We can tune τ using cross-validation. For example, we split the dataset into training and testing, the the trained hybrid model is then evaluated on the testing dataset, according to the relevant performance metrics. We can then tune τ by optimizing these performance metrics across multiple folds of the cross-validation. Note that this is akin to an Empirical Bayes approach.

Appendix B: Additional details about our data and institutional environment

B.1. Usefulness of the NCSC criteria for other jobs

The US Department of Labor maintains the O*NET database[36], which catalogs and classifies all occupations across the U.S. economy according to occupational characteristics and worker requirements.

Using the O*NET database, we illustrate that many other occupations require skills that are present in the NCSC scoring criteria for identifying competent salespeople. Table B.1 shows the typical skills needed for entry-level jobs in occupations such as procurement managers, health services, and business analytics. The first column lists the skill, the second maps each skill to the relevant NCSC criteria, the third column is the importance of that skill for the job.

B.2. A sample of participating companies.

Our judges and buyers are drawn from 162 companies. Below is a sample of 40 companies.

[36] https://www.dol.gov/agencies/eta/onet

Table B.1 Skills needed for various entry-level jobs

Procurement Managers		
Skills/Abilities	**NCSC criteria**	**Importance**
Resolving Conflicts and Negotiating with Others	Objection handling	Most Imp
Establishing and Maintaining Interpersonal Relationships	Rapport building	Top 5
Getting Information	Needs identification	Top 5
Selling or Influencing Others	Close	Top 20
Health Services		
Skills/Abilities	**NCSC criteria**	**Importance**
Communicating with Supervisors, Peers, or Subordinates	Communication	Most Imp
Getting Information	Needs identification	Top 5
Establishing and Maintaining Interpersonal Relationships	Rapport building	Top 10
Resolving Conflicts and Negotiating with Others	Objection handling	Top 20
Business Analysts/Entry-level Management Consulting		
Skills/Abilities	**NCSC criteria**	**Importance**
Getting Information	Needs identification	Top 5
Establishing and Maintaining Interpersonal Relationships	Rapport building	Top 5
Selling or Influencing	Close	Top 20
Resolving Conflicts and Negotiating with Others	Objection handling	Top 20

Table B.2 Pool of companies

IBM	Microsoft	AT&T	American Airlines
Johnson & Johnson	Fujitsu	HPE	Intuit
Gartner	Dish Network	Liberty Mutual	Eli Lilly
Workday	Trend Micro	SAS Institute	Paycom
Vivint	Sherwin-Williams	NCR Corporation	C.H. Robinson
Geico	Principal Financial Group	Rohde & Schwarz	Northwestern Mutual
Goodman Manufacturing	Match.com	Neopost	Digital Shadows
AMN Healthcare	Concentra	Cintas Corporation	Hyatt Regency Dallas
Signet	Solera Inc.	MIC Group	Associa
Lennar	Masergy	Rexel Energy Solutions	MuleSoft

Appendix C: Additional results

C.1. Sample size

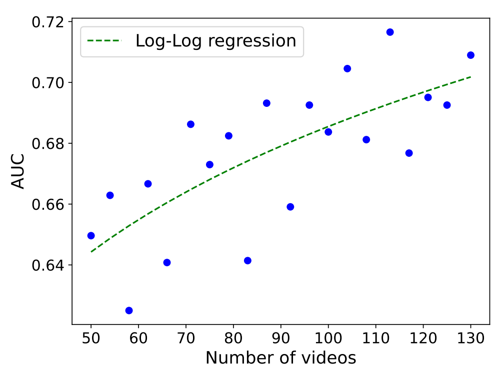


Figure C.1 Predictive performance increases as the number of videos in the training dataset increases.

Table C.1 Performance of other models

Metric		Baseline	Random Forest	XGBoost	SVM
Avg. competency loss	Selection	3.21	1.99	1.88	2.05
Avg. opportunity loss	Selection	2.17	1.20	1.38	1.99
Avg. quality loss	Selection	5.38	3.19	3.26	4.04
Balanced Accuracy	Screening	50%	63.9%	60.9%	65.4%
	Selection	50%	64.4%	61.6%	58.8%
Accuracy	Screening	62.5%	73.3%	71.1%	74.4%
	Selection	62.5%	72.2%	70.0%	67.8%
AUC	Screening	50%	66.0%	71.7%	70.1%
	Selection	50%	73.9%	67.7%	67.0%
Precision	Screening	75%	82.4%	80.9%	83.1%
	Selection	25%	47.9%	43.8%	39.6%

We explore how the size of the training dataset affects the predictive performance of the model. To do this, we vary the number of training videos from 50 to 130, while keeping the number of testing videos fixed. As before, we split the dataset into 4-folds to calculate the average AUC across the 4-folds. To reduce the number of videos in the training dataset, we randomly sample (without replacement) to obtain the needed number of training videos.

In Figure C.1, we show that as the number of videos in the training dataset increases, the model's predictive performance also increases The dotted curve shows a best-fit power function $y = Ax^b$,or equivalently a linear log-log regression. Extrapolating from this relationship, when there are 1,000 videos (750 in the training), the AUC could increase to 82% from the current best of 74%. At 2,000 videos, the AUC could increase to 87% from 74%.

C.2. Performance of other models

In Table C.1 here, we report the predictive accuracy of various AI models, such as XGBoost, Random Forest, and SVM. Random Forest performs best, although the differences among the models are small. We pick Random Forest as our desired AI model because: (i) its performance is comparable to other models, and (ii) statistical inference in the form of standard errors and confidence intervals can be easily calculated, which is needed in the hybrid model (see Section 4.4).

C.3. SHAP values by gender

The table C.2 below highlights the relative importance of a feature for male versus female interviewees as per SHAP plots

Table C.2 Relative importance of a feature for male vs female according to SHAP values.

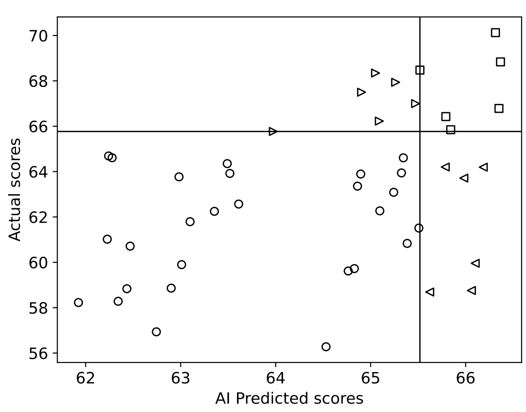
Feature	Relative importance for male vs female
Head movement	1.20
Active listening	1.20
Hand gestures	1.16
⋮	⋮
Information density	0.94
Collaborativeness	0.92
Politeness	0.89

Appendix D: Additional figures for the AI-Human hybrid

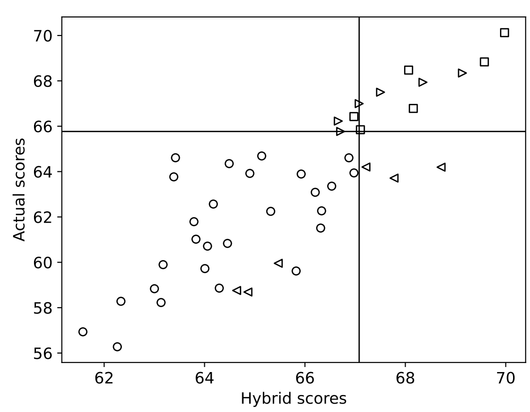
In Figure D.1a, we plot the AI scores against the actual scores to visualize the classification. Since we performed 4-fold cross-validation, this scatterplot corresponds to the out-of-sample prediction, of a particular split of the dataset into 75% training dataset and 25% testing dataset. The horizontal line is the threshold for true selection; the vertical line is the model-predicted threshold for selection.

Figure D.1b shows the corresponding scatter plot for the hybrid model. As can be seen, the hybrid improves performance considerably. The false positives are pushed to the bottom-left quadrant and false negatives are pushed to the top-right quadrant.

Figure D.1 Actual scores versus AI and hybrid scores



(a) AI scores versus actual



(b) Hybrid scores versus actual

Notes: □: True Positive, o: True Negative, ▷: False Negative, ◁: False Positive
The hybrid model corrects many false positives and false negatives from the AI model.

Figure D.2 Decomposing the improvement in hybrid over AI.

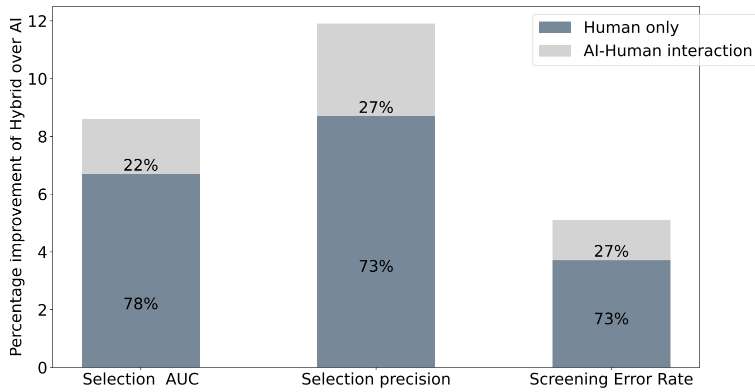


Figure D.3 Distance between hands explains the gap between hybrid and AI scores.

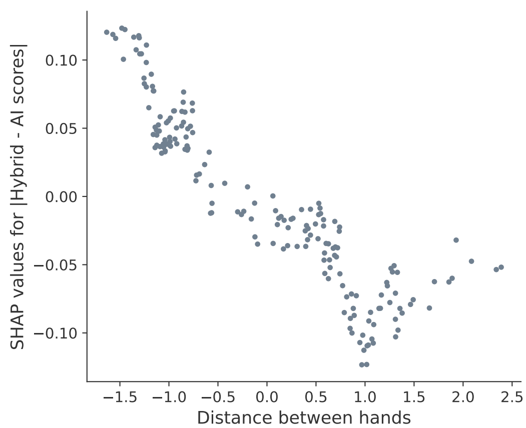


Table D.1 Task-based hybrid model.

	Selection AUC	Selection Precision	Cost per additional good hire	Mean competency loss	Mean opportunity loss	Mean quality loss
Stage 1 (Approach)	60.4%	38.6%	∞	2.29	1.46	3.75 (-17.6%)
Stages 1 to 2 (Needs ID)	79.3%	55.0%	**$648**	1.05	1.03	2.08 (34.8%)
Stages 1 to 3 (Product Demo)	81.3%	58.0%	$663	1.00	0.90	1.90 (40.4%)
Stages 1 to 4 (Obj Handling)	82.0%	59.0%	$838	0.98	0.90	1.88 (40.9%)
All stages (Close)	82.5%	59.8%	$840	0.94	0.84	1.78 (44.1%)

Note: In parentheses of the mean quality loss column, we calculate the percentage improvement of the task-based hybrid models from the pure AI model.

Table D.2 Sequential hybrid model.

Screening threshold	AUC	Precision	Cost per additional good hire	Competency loss	Opportunity loss	Quality loss
5	82.1%	58.8%	$872	0.929	0.853	1.782 (44.1%)
15	81.7%	59.0%	$766	0.931	0.853	1.784 (44.0%)
30	82.0%	59.2%	$619	0.922	0.824	1.746 (45.2%)
50	80.4%	57.8%	**$505**	0.923	0.855	1.778 (44.2%)
70	76.7%	49.2%	$2307	1.211	0.960	2.171 (31.9%)

Note: In parentheses of the mean quality loss column, we calculate the percentage improvement of the hybrid models from the pure AI model.

Appendix E: Verbal styles features

The tables below show the exemplar sentences generated for each verbal style using a simple and persona-based prompt in GPT-4. These two prompts together help capture both a general as well as a context-specific usage of a particular verbal style.

Table E.1 Exemplar sentences for collaborative style

Simple Prompt	Score (Mean)	s.d.
"What are your thoughts on this approach? I would love to integrate any suggestions you might have."	9.3	1.2
"Let's work on this together."	7.7	2.0
"I value your perspective; what do you think?"	7.3	2.5
"How can we combine our strengths to tackle this challenge?"	8.8	0.5
"I believe our collective expertise can lead to a great solution."	7.4	2.4
"Together, we can achieve more than working separately."	8.2	1.9
"I'm open to your suggestions and feedback."	8.1	1.2
"Let's brainstorm and build upon each other's ideas."	8.3	1.2
"Your input has been instrumental; let's continue this collaboration."	9.0	1.7
"By joining forces, we can ensure the best possible outcome."	7.8	1.5
Specialized Prompt		
"We are on the same page here. By working closely together, we can come up with a solution that benefits everyone. "	8.5	0.4
"Your team's input is essential to tailor our services to fit seamlessly with your operations."	8.0	0.0
"At Salesforce, we believe in collaboration, and I'm excited to work together to find the best solutions for your needs."	9.3	0.9
"We're not just selling a product; we're partnering with you to achieve your business goals."	8.0	1.0
"Let's bring in your key stakeholders to ensure a collective understanding and alignment."	9.7	0.6
"I'm here to facilitate collaboration between our teams, making sure we're all on the same page."	8.6	2.1
"Our success lies in our ability to work hand in hand, combining your industry expertise with our solutions."	7.5	1.3
"Your team's insights will guide our collaborative effort to streamline your processes."	8.7	1.0
"I appreciate your commitment to teamwork, which will lead us to a successful partnership."	8.3	1.4
"Together, we can leverage the power of Salesforce to drive innovation and efficiency."	8.2	1.3

Simple Prompt: Can you provide 10 phrases/sentences that convey a collaborative tone?
Specialized Prompt: You are a salesperson from Salesforce.com meeting a new prospective buyer. Can you provide 10 phrases/sentences that convey a collaborative tone to the buyer?

Table E.2 Sentences for analytical verbal style

Simple Prompt	Score (Mean)	s.d.
"Taking into account all available data, it follows that..."	7.5	1.2
"After systematically assessing the cost-benefit analysis, it stands to reason that our product delivers a higher ROI."	9.0	0.4
"From a logical perspective, considering all factors, we can infer that our product delivers greater incremental value."	8.3	1.0
"Given the patterns observed and analyzed, the most plausible conclusion is..."	7.3	1.0
"My goal is to provide you with solutions backed by solid data and a logical, step-by-step implementation plan."	8.8	0.6
"I believe in making decisions based on hard facts and empirical evidence rather than gut feelings."	8.2	1.6
"An analytical approach allows us to measure the impact of our solutions accurately."	8.0	0.8
Specialized Prompt		
"Comparing the ROI of our product to others in the market, it is evident that our offering delivers superior value."	9.5	0.4
We at Salesforce place a strong emphasis on quantitative analysis for accuracy	7.7	1.4
"Our approach will be data-oriented, ensuring we have solid evidence for every assertion."	8.5	0.4
"We excel at recognizing patterns and connections within data."	7.8	1.4
"Analyzing your business metrics will be a crucial part of tailoring our services for maximum efficiency."	8.3	1.0

Simple Prompt: Can you provide phrases/sentences that convey tones of analytical thinking and logical reasoning.
Specialized Prompt: You are a salesperson from Salesforce.com meeting a new prospective buyer.
Can you provide phrases/sentences that convey tones of analytical thinking and logical reasoning.

Table E.3 Sentences for confidence verbal style

Simple Prompt	Score (Mean)	s.d.
"I am certain that we can handle this challenge effectively."	8.5	1.2
"We have all the resources and skills we need to succeed."	8.5	0.7
"I have full faith in my abilities to accomplish this task."	8.7	1.5
"I'm absolutely confident in our team's capabilities to deliver results."	8.7	1.5
"I have no doubt that we'll achieve success in this endeavor."	8.3	2.1
"I'm sure I can provide a solution to this problem."	8.1	1.5
"I firmly believe in my skills and expertise."	8.2	2.5
"I have a strong conviction that we're on the right path."	7.7	0.6
"I'm fully committed to seeing this through."	9.0	1.4
"I know that I can make a positive impact in this situation."	8.4	1.4
Specialized Prompt		
"Our product consistently outperforms its competitors."	9.5	0.4
"I guarantee you won't be disappointed with this purchase."	8.8	0.6
"If you're looking for results, this is definitely the best choice."	9.0	0.4
"Our extensive track record of success across various industries gives me full confidence in the value we can bring to your organization."	8.2	1.5
"I've had the privilege of working with clients who have achieved exceptional results with Salesforce, and I believe your company can too."	8.1	1.4
"You're in capable hands; our team has a deep understanding of your industry's unique challenges and needs."	8.5	1.3
"You can trust that your data will be secure with Salesforce, as we maintain the highest standards of security and compliance."	8.5	2.1
"Our commitment to innovation ensures that you'll always stay ahead in an ever-evolving market."	7.6	1.4
"With Salesforce's comprehensive training and support resources, you'll have a strong partner guiding you every step of the way."	8.0	0.0
"I'm confident that choosing Salesforce is a strategic move that will help your business thrive, and I look forward to assisting you in achieving your goals."	10.0	0.0

Simple Prompt: Can you provide 10 phrases/sentences that convey a confident tone
Specialized Prompt: You are a salesperson from Salesforce.com meeting a new prospective buyer.
Can you provide 10 phrases/sentences that convey a confident tone to the buyer

Table E.4 Sentences for optimistic verbal style

Simple Prompt	Score (Mean)	s.d.
"I'm confident that we can overcome any challenge that comes our way."	7.9	2.3
"The future is bright, and I'm excited about the opportunities ahead."	9.0	0.8
"Let's focus on the possibilities and what we can achieve together."	8.0	0.0
"I see a lot of potential for growth and success in this endeavor."	8.5	2.1
"Even in difficult times, there's always a silver lining."	8.3	1.9
"I believe in our ability to make a positive impact on this situation."	8.3	1.6
"With determination and a positive attitude, we can achieve great things."	8.8	1.1
"Every setback is an opportunity for a comeback."	5.5	2.1
"I'm optimistic about the outcomes we can achieve with hard work and dedication."	8.7	1.6
"Our journey together will be marked by positivity and a can-do spirit."	7.0	1.0
Specialized Prompt		
"I'm genuinely excited about the possibilities our partnership can bring to your business."	8.0	1.4
"The future with Salesforce is bright, and I'm thrilled to be on this journey with you."	8.4	1.3
"Let's explore how Salesforce can transform your operations and drive success."	7.7	0.8
"I see tremendous potential for growth and improvement by implementing our solutions."	8.5	1.5
"Even in challenging times, there are opportunities for innovation and advancement."	9.0	0.0
"With Salesforce, you have the tools to achieve your goals and overcome obstacles."	7.7	1.2
"I firmly believe that together, we can achieve remarkable results."	8.6	0.9
"Every obstacle is a chance to demonstrate the resilience of your business."	7.3	2.1
"I'm optimistic about the positive impact Salesforce can have on your organization."	9.5	0.7
"Our collaboration promises a future filled with efficiency, growth, and success."	8.1	1.2

Simple Prompt: Can you provide 10 phrases/sentences that convey an optimistic tone
Specialized Prompt: You are a salesperson from Salesforce.com meeting a new prospective buyer. Can you provide 10 phrases/sentences that convey an optimistic tone to the buyer

Table E.5 Sentences for cautious verbal style

Simple Prompt	Score (Mean)	s.d.
"Before we proceed, I want to emphasize the importance of a thorough evaluation to mitigate any potential risks."	9.2	0.8
"Let's take a measured approach, ensuring that we address all concerns and minimize any possible pitfalls."	8.0	2.0
"Risk management is a key part of our strategy, and I'll be here to guide you through it."	7.0	1.4
"We'll work together to identify and mitigate any potential roadblocks in the implementation process."	8.2	1.5
"Our team is dedicated to your success, and that means being vigilant in risk assessment and management."	7.5	0.7
"Let's prioritize a thorough analysis of your needs and constraints to avoid any unforeseen challenges."	8.0	0.0
"Let's proceed with caution and carefully assess the potential risks."	10.0	0.0
"I prefer to take a conservative approach to avoid any unexpected setbacks."	8.3	1.7
"Before making a decision, I always consider the worst-case scenario."	9.5	0.7
"We should prioritize safety and minimize any exposure to potential risks."	8.7	0.8
Specialized Prompt		
"I tend to err on the side of caution to ensure the protection of our interests."	8.8	1.4
"It's essential to conduct a thorough risk analysis before moving forward."	8.7	1.6
"I believe in prudent planning to avoid unnecessary complications."	8.8	1.5
"I'm cautious about making hasty decisions without a full understanding of the risks involved."	9.3	1.0
"I believe in taking calculated steps to ensure the success and stability of your transition to Salesforce."	7.3	2.1
"I value a careful, risk-averse approach to safeguard your investment in Salesforce."	9.0	0.8
"I'm here to provide guidance and support every step of the way, ensuring a smooth and risk-free adoption of Salesforce."	6.8	1.5
"We should explore risk mitigation strategies to protect our investments."	8.6	1.3

Simple Prompt: Can you provide 10 phrases/sentences that convey a cautious tone
Specialized Prompt: You are a salesperson from Salesforce.com meeting a new prospective buyer. Can you provide 10 phrases/sentences that convey a cautious tone to the buyer

Table E.6 Sentences for competitiveness verbal style

Simple Prompt	Score (Mean)	s.d.
"We strive to outperform our competitors in every aspect of our business. "	9.0	1.2
"Our team is driven by a fierce desire to be the best in the industry."	9.4	1.1
"Competition fuels our innovation and pushes us to excel."	9.3	1.2
"We relish challenges and see them as opportunities to prove our capabilities."	8.3	1.2
"Our goal is not just to keep up but to set the pace in our market."	8.0	0.0
"We have a track record of surpassing our rivals with groundbreaking solutions."	7.7	1.7
"In a competitive landscape, we stand out as a market leader."	9.5	0.7
"We welcome competition because it motivates us to continuously improve."	8.0	1.4
"Our commitment to excellence is what sets us apart from the rest."	7.6	1.0
"We have a winning mindset, and we're always ready to go the extra mile to succeed."	8.5	1.3
Specialized Prompt		
"Salesforce.com is known for staying ahead of the competition with cutting-edge solutions."	8.4	1.6
"We thrive in competitive markets because we're driven to outperform and innovate."	8.2	1.8
"Our track record proves that we're not just in the game; we lead it."	9.7	0.6
"When it comes to CRM and sales automation, we're the top choice, and here's why."	7.5	3.4
"We're here to help you gain a competitive edge in your industry."	8.0	2.4
"Our tools and services are designed to ensure you stay ahead of your rivals."	9.3	0.8
"We embrace competition as an opportunity to showcase our excellence."	8.3	2.1
"With Salesforce, you'll have a partner that's always a step ahead of the competition."	8.4	0.8
"Our commitment to continuous improvement sets us apart in a fiercely competitive market."	7.8	1.8
"Let's strategize on how Salesforce can position your business for competitive success."	8.4	1.5

Simple Prompt: Can you provide 10 phrases/sentences that convey a competitive tone
Specialized Prompt: You are a salesperson from Salesforce.com meeting a new prospective buyer. Can you provide 10 phrases/sentences that convey a competitive tone to the buyer

Table E.7 For the politeness verbal style, we use well-established sentences from Danescu-Niculescu-Mizil et al. (2013)

Politeness strategies	**Sentences**
Gratitude	Thank you very much, I really appreciate it.
Deference	Wow, nice work!
Greeting	Good morning. Good afternoon. I hope you are having a wonderful day.
Apologizing	I am very sorry to bother you.
Please	Could you please...
Indirect	By the way, would you by any chance...; If it is not too much of a trouble, I was wondering if it would be at all possible..
Counterfactual modal	Could you...; Would you...

Appendix F: Interactivity features

We consider two dimensions of real-time adaptation — active listening and dynamic style matching. Active listening refers to paraphrasing, and asking clarifying interest such that it shows engagement and high empathy. Dynamic style matching is to measure whether one speaker (interviewee) is changing their linguistic style in response to the other (interviewer).

F.1. Active listening

Here, we describe the details of how we use GPT API to evaluate the seller's active listening skills.

There are two ways of accomplishing this; zero-shot prompting which means directly asking GPT to rate the conversation. However, there is evidence that GPT's performance on zero-shot prompting is often not as good as state-of-the-art in most complex classification tasks (Kocoń et al. 2023). However, the performance can be improved through few-shot prompting i.e. giving GPT some examples to learn.

We identify example transcripts where the salesperson demonstrates active listening by rephrasing and summarizing the buyers' concerns, and asking clarifying questions that confirm the buyers' needs. Figure F.1 shows how we obtain these examples. In the scorecards, judges can write down their comments. We found scorecards where judges specifically commented on the salesperson's active listening or the lack thereof.

More specifically, we use `gpt-3.5-turbo-16k` which has 16k context window. This means we are allowed to pass a total of 16K tokens including the interview transcripts and examples. Since our transcripts are relatively long, we can only give 2 examples. First, we provide a system message to the Chat Completion API: "You are given a conversation between a buyer and a seller. Your task is to score the seller's communication skills in terms of active listening on a scale of 6 to 10." Then we provide two examples to GPT, an example of a transcript that demonstrates active listening, and another transcript that demonstrates poor active listening. These example transcripts are labeled with a numeric response of 10 and 6 respectively. GPT then provides a numeric score on active listening to all other transcripts using these examples as references.

F.2. Dynamic style matching

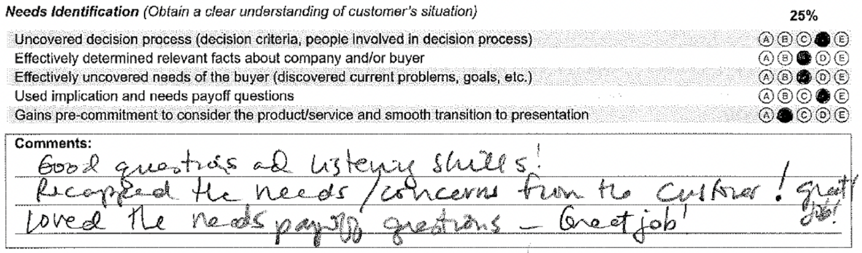
More precisely, let a and b denote the two speakers, C denote a word class (e.g. auxiliary verb or high-frequency adverbs). Let $T^{C}_{a,b\to a}$ denote the event that speaker a uses a word belonging to class C in a conversational turn with b, where b starts the turn. $P(T^{C}_{a,b\to a})$ is the probability of the event, which is calculated as the fraction of all a's conversational turns with b in which a's utterance contains a word in class C. Our linguistic adaptation measure with respect to class C is

$$LinguisticAdapt(C) = P(T^{C}_{a,b\to a}|T^{C}_{b,b\to a}) - P(T^{C}_{a,b\to a}) \tag{12}$$

$P(T^{C}_{a,b\to a}|T^{C}_{b,b\to a})$ is the conditional probability that a uses C given that b uses C during a conversational turn between a and b where b starts the turn. This conditional probability is calculated as the fraction of all conversational turns between a and b in which b's usage of C is immediately followed by a's usage of C. Finally, to arrive at a single overall adaptation for the seller in response to the buyer, we take an average across $LinguisticAdapt(C)$. Similarly, we also calculated the buyer's adaptation to the seller.

In addition to real-time adaptation, we also constructed linguistic and vocal similarity measures between the interviewer and interviewee. Our measure of linguistic similarity is based on Language Style Matching

Figure F.1 Examples of Active Listening comments from judges

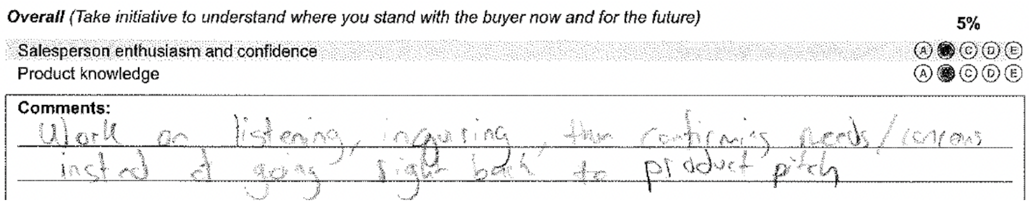


Needs Identification *(Obtain a clear understanding of customer's situation)* 25%

Uncovered decision process (decision criteria, people involved in decision process) Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ
Effectively determined relevant facts about company and/or buyer Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ
Effectively uncovered needs of the buyer (discovered current problems, goals, etc.) Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ
Used implication and needs payoff questions Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ
Gains pre-commitment to consider the product/service and smooth transition to presentation Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ

Comments:
Good questions and listening skills!
Recapped the needs/concerns from the customer! Great job!
Loved the needs payoff questions – Great job!

(a) Example of good active listening, judges commented: "Good questions and listening skills. Recapped the needs/concerns..."



Overall *(Take initiative to understand where you stand with the buyer now and for the future)* 5%

Salesperson enthusiasm and confidence Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ
Product knowledge Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ

Comments:
Work on listening, inquiring, then confirming needs/concerns
instead of going right back to product pitch

(b) Example of poor active listening, judges commented: "Work on listening, inquiring, then confirming needs/concerns, instead of going right back to product pitch"

(LSM) of Ireland and Pennebaker (2010), which measures the similarity between the interviewer and the interviewee's usages of categories of words considered important for style matching (e.g., articles, prepositions, conjunctions, etc). Vocal similarity is calculated using the cosine similarity between the interviewer and the interviewee's vectors of vocal features (as described above in *Vocal Styles*).

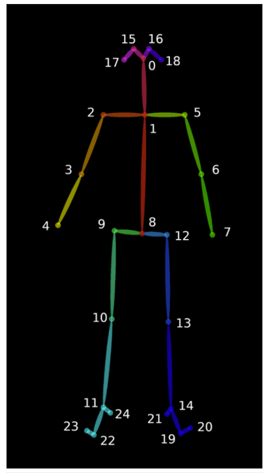
Appendix G: Formulas for constructing body language features

In the text, we described how to calculate the rate of hand gesture. In this section, we show how the other body language features are constructed from the OpenPose output.

Let $Nose_t$ be the vector of coordinates of the seller's nose at frame t. This is the output of OpenPose (keypoint 0 in Figure G.1). Let $Neck_t$ be the vector of coordinates of the seller's neck base (OpenPose's keypoint 1 in Figure G.1) at frame t. The length of the seller's neck at frame t is then $\|Nose_t - Neck_t\|$. In Equation 13, we define the rate of head movement as the distance traveled by the nose between two consecutive frames, normalized by the length of the seller's neck.

$$\text{Rate of Head Movement} = \frac{T\sum_{t=2}^{T}\|Nose_t - Nose_{t-1}\|}{(T-1)\sum_{t=1}^{T}\|Nose_t - Neck_t\|} \tag{13}$$

Figure G.1 Key Points from OpenPose



Let $LShoulder_t$ and $RShoulder_t$ be the vector of coordinates for the seller's left shoulder and right shoulder respectively, given by OpenPose's keypoints 2 and 5. To measure hand openness or the distance between hands, we take the average distance between $LWrist_t$ and $RWrist_t$, normalized by the width of the seller's shoulder. The width of the seller's shoulder can be calculated as $\frac{1}{T}\sum_{t=1}^{T}\|LShoulder_t - RShoulder_t\|$.

$$\text{Distance between hands} = \frac{\sum_{t=1}^{T}\|LWrist_t - RWrist_t\|}{\sum_{t=1}^{T}\|LShoulder_t - RShoulder_t\|} \tag{14}$$

Let Hip_t be the vector of coordinates for the seller's hip (Keypoint 8). As such, $Neck_t - Hip_t$ is the vector corresponding to the seller's torso. The angle of the torso is defined as the angle between the torso vector and the horizontal vector $(1,0)$.

$$\text{Torso angle} = \frac{1}{T}\sum_{t=1}^{T}\arccos\left(\frac{(Neck_t - Hip_t)\cdot(1,0)}{\|Neck_t - Hip_t\|}\right) \tag{15}$$

Therefore, the smaller the angle, the more attentive the posture is. 90 degrees means perfectly upright, greater than 90 degrees means the person is leaning backwards, and smaller than 90 degrees means the person is leaning forward.