EQUITABLE ELICITATION*

Anonymous authors
Paper under double-blind review

ABSTRACT

Individuals with similar qualifications and skills may vary in their demeanor, or outward manner: some tend toward self-promotion while others are more self-effacing. Comparing the self-descriptions of equally qualified job-seekers with different self-presentation styles is therefore problematic. We build an interactive AI for *skill elicitation* that provides accurate determination of skills *while simultaneously allowing individuals to speak in their own voice.* Such a system can be used, for example, when a new user joins a professional networking platform, or when seeking to match employees to needs during a company reorganization.

We articulate the equitable elicitation problem. We build an interactive AI to address the problem, providing accurate and equitable determination of skills. To this end, we train an LLM to act as synthetic humans to provide training data, and we provably enforce a mathematically rigorous notion of equitability, ensuring that the covariance between self-presentation manner and skill evaluation error is small.

1 INTRODUCTION

Garbage in, garbage out. Data are literally "the givens," and biases in data yield biased systems. Self-descriptions are not exempt from bias: data about individuals that are provided by the individuals themselves exhibit cultural biases. More broadly, individuals, even within the same culture, differ in their levels of modesty and tendency (or ability) to self-promote. These differences have measurable downstream effects on outcomes, such as recruitment Murciano-Goroff (2022).

If the problem were simply one of tone, existing language models could mitigate the effect by homogenizing the presentation of facts supplied by the individuals, for example, by using a prompt of the form "Rewrite this in the style of a white cisgender male in an individualistic culture." However, the LLM cannot compensate for relevant information that the user *fails to supply*, which studies show to exhibit pronounced cultural and gender disparities Lindström (2017); Murciano-Goroff (2022); Deschacht & Maes (2017) [1]

This is the subject of the current work. For concreteness, we focus on job-seekers' self-generated professional profiles, for example, as on LinkedIn, but our proof of concept can extend to many other settings. Our key contribution is an interactive tool for accurate and equitable elicitation of skill signals in job candidates across a cultural spectrum. That is, we determine whether or not a person has a specific skill based on answers to a few questions. Crucially – and this is a principal design goal – our tool allows individuals to speak in the style in which they are comfortable. The pressure, and even well-intended advice, to conform to a specific interaction style is widespread, burdening

*This work was facilitated by the Hire Aspirations Institute at Harvard, supported in part by Alfred P. Sloan Foundation grant G-2017-9890.

[1]For example, in their official blog, LinkedIn reports Lindström (2017): "In the U.S., women on average include 11% less skills than men on their LinkedIn profile, even at similar occupations and experience levels" (see also Ahuja (2024)). This number is matched in Murciano-Goroff's brilliant measurement of the gender gap in self-reporting of skills via individuals' actual previous coding work as seen on github and taking into account "the number of lines of code that a candidate has contributed to *others'* open source projects as an indication of the reputation of the candidate's coding work" (emphasis added). There is also cross-cultural research finding a self-promotion gap between collectivist and individualist cultures. For example, Deschacht & Maes (2017) finds a gap between rates of self-citations in individualist and collectivist cultures (and finds the gender gap firmly persistent across different cultures).

even highly competent but conversationally more restrained individuals Mohr (2014); Ahuja (2024). To our knowledge, ours is the first work to actively elicit information while respecting personal style.

For our purposes, a *predictor* is a function that maps a transcript to a score in $[0, 1]$. Speaking intuitively and informally, one can think of this score as a "probability" that the individual in the transcript has a stated skill [2].

We ensure a mathematically rigorous notion of "equitability" which is a form of multi-accuracy Hébert-Johnson et al. (2018) with respect to the Big 5 personality types, which roughly means that the covariance between the personality type and the prediction error is small Gopalan et al. (2021). The choice of Big 5 was expedient, as algorithms for these are readily available in the NLP literature. Indeed, multi-accuracy is completely general and multi-accurate learning algorithms provide the low-covariance guarantee with respect to any pre-specified collection of functions mapping individuals to a finite range Hébert-Johnson et al. (2018); Dwork et al. (2021); Gopalan et al. (2022).

Beyond formalizing and addressing the problem of equitable elicitation while preserving individual voice, our work has several notable features:

1. We provide a novel example of the use of LLMs for computational social science research. Drawing on a publicly available dataset of individual profiles, we build an LLM to create synthetic humans on which we train an accurate and equitable scoring function for skills[3]
2. In many commercial settings, the training of predictors is a dynamic process, with systems trained to improve themselves in a sequence of epochs. Let p_i denote the predictor in the ith epoch. Our system ensures that every one of these epochal predictors p_i satisfies our fairness constraint.

1.1 High-Level Overview

We use a structured approach to building an elicitation model based on Wang et al. (2025) and Sutton *et al.* style reinforcement learning Sutton et al. (1999).

We determine whether or not a person has a specific skill based on their answers to a series of adaptively selected questions. The question selection algorithm is trained to minimize uncertainty, with the questions belonging to a pre-decided and human created set: no user will ever be asked a question that was directly generated by an LLM, and the pool of possible questions can be audited beforehand by any relevant parties.

Given a question-and-answer conversation prefix, the selection of the next question is done via a fixed machine learning based function, which maps the transcript of user responses to a probability distribution on the next question. This *question selector* converts the user's responses into vectors using an embedding model, then feeds them into our pre-trained transformer based machine learning system. We use a transformer architecture because that allows us to input a variable length conversation prefix, *e.g.*, on the first question we only have the user's background information or profile, while on the fifth we have 4 previous answers. This system is static and reproducible: given the same answers it will produce the same distribution on outputs.

Once we have the transcripts of questions and responses, we construct our *skill predictor* using a technique similar to the one used to build the question selector. We formulate the skill prediction task as a binary classifier. The predictor deterministically maps a transcript of questions and responses to a score in $[0, 1]$. The skill predictor is post-processed to ensure *multi-accuracy* with respect to the Big-5 personality types using a technique from Gopalan et al. (2022)[4]. This guarantees that (1) the predictor is (nearly) correct in expectation on each personality type, and (2) the personality types have low convariance with the prediction errors.

[2] The meaning of a probability for a non-repeatable event is a deep and unresolved question in the theory of probability. For more on this in the context of the predictor technology used in this work, see Dwork et al. (2021).

[3] This is the only use of LLMs in our work; had we sufficient data LLMs would not have been necessary. See the discussion in Section 5.

[4] A stronger guarantee, multi-calibration, can also be ensured, but the algorithm is more complicated. We chose multi-accuracy for the proof of concept.

We remark that neither of the two functions is generative. The training process involves generative AI to bootstrap and expand the training data, but at runtime nothing generative is used.

2 Related Work

2.1 Algorithmic Fairness

The literature on algorithmic fairness is broad and rapidly growing, encompassing notions such as individual fairness (Dwork et al., 2012) and group fairness (Hardt et al., 2016). Protections defined with respect to a single group can be too weak (Dwork et al., 2012), while finding a metric for individual fairness is problematic. This lead to settings with multiple, potentially overlapping subgroups (Hébert-Johnson et al., 2018; Kearns et al., 2018; Kim et al., 2019). Most pertinent to our work are the frameworks of multi-calibration and multicalibration and multiaccuracy (Hébert-Johnson et al., 2018; Kim et al., 2019), which require predictors to maintain uniformly small error rates across a rich collection of subpopulations. By enforcing multi-accuracy with respect to presentation styles, we extend these fairness-as-accuracy goals into the domain of *elicitation*, adding pro-active interaction to ensure that stylistic variation does not systematically affect errors in skill estimation..

A wide range of related approaches highlight complementary fairness notions. Fair representation learning methods aim to construct latent spaces invariant to sensitive attributes Zemel et al. (2013); Wang et al. (2024). The problem of model multiplicity, as highlighted by Black et al. Black et al. (2024), emphasizes that among equally accurate models, one should select those that are less discriminatory. Additional perspectives include metric-based fairness, which requires that similar individuals receive similar predictions Dwork et al. (2012); Ilvento (2019); Rothblum & Yona (2018); Mukherjee et al. (2020), de-biasing Bolukbasi et al. (2016), corrective transformations Dwork et al. (2023), and work that seeks to move beyond explicit definitions of fairness altogether Jung et al. (2020).

2.2 Endogenous Biases: Empirical Evidence

A substantial body of empirical evidence shows that individuals' self-presentation styles introduce systematic disparities in evaluation outcomes. For example, East Asian employees tend to exhibit a modesty bias in self-ratings of job performance, leading to lower self-assessments despite comparable actual performance FARH et al. (1991); JIAYUANYU & Murphy (1993); Cho et al. (2023). Such disparities highlight how cultural norms of modesty versus self-promotion can shape the information available to evaluators, and thereby affect professional advancement.

Research on self-presentational behavior shows that applicants strategically adjust how they present themselves during hiring processes, often influencing evaluations in ways that are independent of their actual skills König et al. (2011). More recently, studies have explored how AI systems interact with such biases, showing both risks and opportunities. For instance, Hofmann et al. (2024) and Guenzel et al. (2025) analyze the role of AI in recruitment and assessment, noting that algorithmic tools can either amplify existing disparities or help to mitigate them if designed carefully.

Relatedly, Eloundou et al. (2024) investigates biases in chatbot responses linked to users' names, illustrating how seemingly innocuous identity cues can alter model behavior. Bolukbasi et al. (2016) shows that widely used word embeddings encode gender stereotypes, raising concerns that linguistic representations themselves can reproduce and amplify social biases. Most recently, Agarwal et al. (2024) demonstrates that Western-centric AI suggestions led Indian participants to adopt Western writing styles, homogenizing their work toward Western norms and diminishing cultural nuance.

2.3 LLMs as Interviewers

In recent years, AI has increasingly been employed as an interviewer across a range of contexts, including the elicitation of political opinions (Wuttke et al., 2024), qualitative interviews (Chopra & Haaland, 2023), quantitative telephone surveys (Leybzon et al., 2025), and even in assisting with questionnaire design (Adhikari et al., 2025). Wang et al. (2025) further proposes an adaptive elicitation framework that leverages LLMs to select questions sequentially so as to reduce uncertainty about latent traits. While these approaches emphasize scalability, efficiency, or data

quality, our work differs in explicitly addressing fairness by ensuring that elicited skill assessments are not systematically biased by individual differences in self-presentation style.

3 METHODS

3.1 PROBLEM SETUP

We study the *skill elicitation problem*: given an applicant j, the goal is to determine whether they possess a target skill. The ground-truth label is denoted $y^j \in \{0, 1\}$, and our predictor produces an estimate $\hat{y}^j \in [0, 1]$.

The elicitation is interactive. At each round i, the system asks a question q_i from a fixed bank $\mathcal{Q}$, and the applicant responds in free text s_i. The dialogue history up to round i is called the *transcript*:

$$x_i^j = (q_1, s_1, \ldots, q_i, s_i).$$

Let $\mathcal{X}$ denote the space of transcripts. Our system learns two components: 1). A **question selector** $q_\psi : \mathcal{X} \to \Delta(\mathcal{Q})$, parameterized by ψ, which maps the transcript x_i^j to a probability distribution over the remaining questions. 2). A **score function** $f_\phi : \mathcal{X} \to [0, 1]$, parameterized by ϕ, which outputs the predicted probability $\hat{y}^j$ of skill possession.

Each response s_i is embedded using a pretrained text encoder $\mathrm{BERT}()$, concatenated with its question index, and aggregated with a GPT-style attention mechanism $\mathrm{ATTN}()$. For example:

$$q_\psi(x_i) = \mathrm{ATTN}_\psi(q_1 \circ \mathrm{BERT}(s_1), \ldots, q_i \circ \mathrm{BERT}(s_i)),$$
$$f_\phi(x_i) = \mathrm{ATTN}_\phi(q_1 \circ \mathrm{BERT}(s_1), \ldots, q_i \circ \mathrm{BERT}(s_i)),$$

where $\circ$ denotes concatenation.

3.2 OPTIMIZATION OBJECTIVES

Our learning problem can be cast in an actor–critic framework (Sutton & Barto, 2018), where the question selector q_ψ plays the role of the *actor* and the score function f_ϕ serves as the *critic*. The actor is responsible for deciding which question to ask next, while the critic evaluates the quality of the interaction by predicting whether the applicant possesses the target skill. The two models are updated jointly and influence one another during training. Below we detail the objectives for each component.

Uncertainty-guided policy learning. A naive approach would be to update the question selector using the final prediction error $|y^j - \hat{y}^j|$ as the reward. However, this leads to very sparse and delayed feedback: a sequence of questions may only be judged good or bad once the entire interview is complete. To provide a more informative training signal, we instead optimize the policy for *information gain* (Wang et al., 2025).

For a transcript x_i^j and candidate next question q, we simulate M steps into the future using the response simulator $s()$ and compute the rollout-based certainty measure

$$Z(q) = 2\big|f_\phi(x_{i+M}^j) - 0.5\big|.$$

Intuitively, $Z(q)$ is large when asking q leads to states in which the score function f_ϕ makes more confident predictions. In other words, the policy is rewarded for selecting questions that are expected to reduce predictive uncertainty the most. Algorithm 1 outlines this rollout-based calculation.

Given this reward definition, we optimize the question selector using the REINFORCE algorithm. Specifically, the parameters ψ are updated by

$$\psi \leftarrow \psi + \alpha\gamma Z(q)\nabla_\psi \log q_\psi(q_i^j \mid x_i^j),$$

where α is the learning rate and γ is a discount factor (both set to 1 in our case). This update increases the probability of choosing questions that are expected to maximize certainty gain. This update encourages q_ψ to increase the probability of selecting informative questions.

Algorithm 1 Rollout-based Uncertainty Z

Input: Transcript x_i^j, candidate question q, depth M, models f_ϕ, q_ψ, simulator $s()$
Output: Z
Create x_{i+1}^j by appending q and its simulated response
for $m = 1$ to $M - 1$ **do**
  Sample next question q' from $q_\psi(x_{i+m}^j)$
  Append q' and simulated response to form x_{i+m+1}^j
end for
Compute $Z = 2|f_\phi(x_{i+M}^j) - 0.5|$
return Z

Supervised training of the score function. The critic f_ϕ is trained to approximate the ground-truth label y^j. Unlike the policy, the critic has access to the true skill label during training and can therefore be updated with a direct supervised loss. Specifically, we use mean squared error (MSE):

$$\mathcal{L}_{\text{MSE}}(\phi) = \frac{1}{B}\sum_{j=1}^{B}\big(f_\phi(x_i^j) - y^j\big)^2.$$

The gradient update for ϕ is

$$\phi \leftarrow \phi - \beta(f_\phi(x_i^j) - y^j)\nabla_\phi f_\phi(x_i^j),$$

with learning rate β. This ensures that f_ϕ gradually improves its skill predictions as more transcripts are observed.

The objectives of q_ψ and f_ϕ are interdependent. As the critic improves, its uncertainty estimates Z become sharper, which provides the actor with a more reliable reward signal. Conversely, as the actor learns to ask more informative questions, the critic observes richer transcripts and thus achieves higher predictive accuracy. This coupling drives a virtuous cycle: the actor guides exploration while the critic supplies feedback, leading to a jointly optimized interactive elicitation process.

3.3 Fairness Calibration

While the optimization objectives above ensure that the score function f_ϕ becomes accurate on average, they do not guarantee equitable performance across subgroups of applicants. In particular, skill evaluation may correlate with personality traits or communication styles that are irrelevant to the actual skill. If left unchecked, the model could systematically overestimate or underestimate candidates from certain personality groups, leading to biased outcomes.

Multiaccuracy constraint. To mitigate this issue, we require the predictor to be *multiaccurate* with respect to a family of auxiliary functions. Let $\mathcal{C}$ denote this family, which includes pretrained predictors of the Big-5 personality dimensions (openness, conscientiousness, extraversion, agreeableness, neuroticism) as well as the constant function $\mathbf{1}$. That is, $|\mathcal{C}| = 6$. Each $c \in \mathcal{C}$ maps a transcript x_i^j to a real-valued score in $[0, 1]$, representing, for instance, how strongly the applicant exhibits a certain trait.

A predictor f_ϕ is said to be $(\mathcal{C}, \epsilon)$-multiaccurate if

$$\sup_{c\in\mathcal{C}}\Big|\mathbb{E}_j\big[c(x_i^j)\cdot(f_\phi(x_i^j) - y^j)\big]\Big| \le \epsilon. \tag{1}$$

This condition ensures that the prediction error of the score function is approximately uncorrelated with all functions $c \in \mathcal{C}$: no personality dimension (or their linear combinations) can explain away the residual bias.

Constructing a calibrated predictor. To enforce multiaccuracy, we adapt the procedure of Gopalan et al. (2023). The idea is to correct the raw predictor f_ϕ by adding a linear combination of the auxiliary functions, followed by a transfer function. Concretely, we define

$$f_\phi^*(x) = \sigma\left(\sum_{c\in\mathcal{C}} l_c\, c(x) + \sigma^{-1}(f_\phi(x))\right),$$

Algorithm 2 Rollouts Algorithm

Inputs: x_i^j, N, M, $f_\phi()$ $q_\psi()$, $s()$
Output: Z_i^j, q
initialize $Z_i^j = 0$, $q = 0$
Assume for simplicity: $N \leq |Q|$
Calculate ordering of next questions $q_1, q_2, \ldots, q_N$ using $q_\psi(x_i^j)$ by taking the top N
for $n = 1$ to N **do**
  Create state x_{i+1}^j after q_n using $s()$
  for $m = 1$ to $M - 1$ **do**
    Select next question q_i using $q_\psi(x_{i+m}^{j,n})$
    Create state $x_{i+m}^{j,n}$ by adding q_i to $x_{i+m}^{j,n}$ then using $s()$
  end for
  Calculate $Z_{i+M}^{j,n} = 2|f_\phi(x_{i+M}^{j,n}) - .5|$
  if $Z_{i,n}^j > Z_i^j$ **then**
    $Z_i^j = Z_{i,n}^j$
    $q = q_n$
  end if
end for
return Z_i^j, q

where $\sigma(t) = 1/(1 + \exp(-t))$ is the sigmoid function, σ^{-1} is its logit, and $\{l_c\}$ are calibration weights. Intuitively, the correction shifts the logit of $f_\phi(x)$ depending on the personality scores $c(x)$ so that the residual correlation between errors and traits is minimized. The weights $\{l_c\}$ are obtained by solving a convex optimization problem derived from Theorem 5.6 of Gopalan et al. (2023) (see the theorem statement in the appendix Section A.3):

$$\begin{aligned}\mathcal{L}_g(y, x) &= -\big(y \log(x) + (1 - y) \log(1 - x)\big) \\ &= -\frac{1}{N} \sum_{j=1}^{N} \Bigg[y^j \log\Bigg(\sigma\Big(\sum_{c \in \mathcal{C}} l_c c(x_i^j) + \sigma^{-1}(f(x_i^j))\Big)\Bigg) \\ &\qquad + (1 - y^j) \log\Big(1 - \sigma\Big(\sum_{c \in \mathcal{C}} l_c c(x_i^j) + \sigma^{-1}(f(x_i^j))\Big)\Big)\Bigg] \\ &\quad + \epsilon \cdot \sum_{c \in \mathcal{C}} |l_c|\end{aligned} \tag{2}$$

This yields a predictor f_ϕ^* that is provably multiaccurate with respect to $\mathcal{C}$.

Integration with training. Fairness calibration is interleaved with the actor–critic updates described earlier. After a number of standard training steps, we perform a calibration phase: estimate $\{l_c\}$ on a batch of transcripts, update the calibrated predictor f_ϕ^*, we then use f_ϕ^* as f_ϕ for the next batch. This alternating procedure ensures that improvements in predictive accuracy are continually balanced with fairness guarantees. If the calibration condition is ever violated (i.e., the bound above exceeds ϵ), training reverts to the last feasible state.

3.4 Training Algorithm

Algorithm 3 summarizes the full training loop. Each iteration alternates between: (i) sampling transcripts, (ii) computing rollout-based uncertainty values Z, (iii) updating q_ψ and f_ϕ with stochastic gradients, and (iv) periodically recalibrating f_ϕ to enforce multiaccuracy.

3.5 Theoretical Guarantee

Finally, we establish that fairness is preserved during training.

Algorithm 3 Equitable Elicitation Algorithm - General Solution

initialize Fixed Values: $\mathcal{Q}$, $\mathcal{C}$, $s()$
initialize Learned Variables: $\psi \sim \mathcal{U}$, $\phi \sim \mathcal{U}$, $l_c = 0$, $Z = 0$
initialize Hyper-parameters: α, β, ϵ, γ, N, B, i_{max}
while ϕ and ψ not converged **do**
  for b^{fair} to B_{fair} **do**
    Initialize empty batches $\mathcal{B}_x = \{\}$, $\mathcal{B}_Z = \{\}$, $\mathcal{B}_y = \{\}$
    If not first run
    Initialize batches from $\mathcal{B}_x^{\text{fair}}$
    Pick random $i < i_{\text{max}}$
    for $b = 1$ to B **do**
      Sample j from $s()$
      Form x_i^j based on $s()$ with $q_\psi()$
      Calculate Z_i^j using Algorithm 2 with x_i^j, N, $f_\phi()$, and $q_\psi()$
      Append x_i^j to $\mathcal{B}_x$, Z_i^j to $\mathcal{B}_Z$, y^j to $\mathcal{B}_y$
    end for
    Update ψ using batch $\mathcal{B}_Z$ and $\mathcal{B}_x$ via Adam
    Update ϕ using batch $\mathcal{B}_y$ and $\mathcal{B}_x$ via Adam
  end for
  Initialize batches $\mathcal{B}_x^{\text{fair}} = \{\}$, $\mathcal{B}_y^{\text{fair}} = \{\}$
  Fill up $\mathcal{B}_x^{\text{fair}}$ and $\mathcal{B}_y^{\text{fair}}$ until they are big enough
  Calculate l_c for each $c \in \mathcal{C}$ using $\mathcal{B}_x^{\text{fair}}$ and $\mathcal{B}_y^{\text{fair}}$ via linear optimizer
  Create f_ϕ^* with l_c
  Verify f_ϕ^* satisfies Equation 1 using $\mathcal{B}_x^{\text{fair}}$, $\mathcal{B}_y^{\text{fair}}$ and l_c
  if not satisfied **then**
    Break, failure restart at previous good state
  end if
end while

Theorem 3.1 *Let $\mathcal{F} = \{f_\phi : \phi \in \Phi\}$ be the hypothesis class, and suppose the solution to the loss function* (2)*, f^*, lies within ϵ^* of $\mathcal{F}$. Then f^* is $(\mathcal{C}, \epsilon)$-multiaccurate at every iteration of the interactive procedure.*

Thus our method ensures that prediction errors remain approximately uncorrelated with applicant personality features, yielding an equitable elicitation process.

4 Specific Instantiation

We are interested in demonstrating the efficacy of this method in a real world scenario.

4.1 Problem Construction

We formulate our problem as skill identification, specifically given a Linked-In style profile and a small number of questions we seek to identify if the job candidate poses a specific skill.

Our procedure is based on existing interview practices, we consider the interview as a sequence of questions, where each question is picked from a fixed set of possible questions based on the current state of the interview. The interviewee then responds to each question, and at the end of the interview we attempt to determine if the interviewee possesses a specific skill. In practice this would be the first step in a hiring process, where the expectation is that many interviewees would "pass", and then be subject to further screening. This is not meant to replace a real interview, but rather to increase the fairness of the initial screening process.

While this work is motivated by the empirically observed biased in hiring practices against women and minorities in many technical fields. We will focus on a component of this problem, personality based biases. We attempt to create an interview process that is fair with respect to the Big 5 per-

Method	Step	Accuracy	Fairness loss
No Correction	137	.781	0.0163
With Correction	137	1.0	0.0048

Table 1: Accuracy and fairness loss for our model with and without the fairness correction during training. Accuracy is for a binary prediction task, and fairness loss is calculated across a batch using Equation 2

sonality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism). As some of these traits are correlated with gender and race, biases against people lacking Extraversion or Conscientiousness can lead to outcomes that are biased against women and minorities.

We start with a set of sample profiles collected from public profiles on a large professional networking site. These profiles were then filtered to remove entries that were missing fields or had too short of a summary. The remaining profiles were then split into training, validation, and test sets.

Then the profiles were converted into a structured text format, with a fixed set of key-value pairs, with each key representing a field in the profile (e.g., "full_name", "gender", "industry", "skills", "summary", etc.). This structured text format was then used to fine tune Meta-Llama-3.1-8B, creating a model that generates profile summaries based on the other fields in the profile. This is not a general purpose LLM, but rather a method of sampling from the space of valid profile summaries, that maintains the biases inherent in the dataset.

We structure our interviews as a series of implicit counterfactual questions. We also assume that the interviewee is being truthful, and that there is no deception, beyond that inherent to our dataset.

We start with a base profile, that represents the ground truth of the interviewee. We then ask questions by modifying the skill fields in the profile and take the generated summary as the response. Thus we can construct a series of self-summaries that represent the interviewee under different skill self-assessments. Our final objective is to determine if the interviewee possesses a specific skill based on this sequence of self-summaries.

Our interview format allows us to frame the interview process as the reinforcement learning problem described in section 3.2. Given a base profile it either does or does not possess a target skill. Our end goal is to create a fair score function that takes the sequence of self summaries and predicts if the interviewee has the target skill. We call this function f_ϕ.

Given f_ϕ we can then assign the correct answer to have a reward of +1, allowing us to attempt to learn a policy that maximizes the accuracy of the final answer. We use a simple policy gradient method, REINFORCE, to learn a policy (f_ψ) that selects the next skill, but use reducing uncertainty as the final reward instead of accuracy. These two functions combine to create an actor-critic model, where the actor is f_ψ and the critic is f_ϕ.

4.2 Results

For training we used a small (400k parameter) transformer based neural network for both f_ϕ and f_ψ with the transcripts being preprocessed via an embedding model[5]. We used the LLM to simulate the job applicants with half being initialized with real profiles from our validation set, and half being fully synthetic. The skill in question is leadership, with 20% of candidates having it during our simulations. We allowed the model to ask 2 questions after the initial self description, and our rollouts were for an additional 2 questions. As described above questions are posed as new skills the candidate may have. Each step was 16 interviews with the fairness calibration being done on 32 interviews. We ran the process until the loss on both $f_\phi()$ and f_ψ converged.

Table 1 shows the results of our full training procedure run with and without the fairness correction. Note that the fairness loss is given after the final fairness correction is applied, before it is applied the fairness loss was 0.0226 and the accuracy was 0.968. The final step of 137 was chosen as it was where both methods showed convergence, note we only conduct the fairness correction every

[5]We used `sentence-transformers/all-MiniLM-L6-v2` (Reimers & Gurevych, 2019) which is based on Wang et al. (2020)

16 steps. Note that the no correction model consistently converged faster, taking approximately 50 steps to converge.

5 Discussion

In this work we present results showing the fairness can be improved even in an interactive social setting. We prove that our algorithm will provide theoretical guarantees of fairness, and demonstrate that it can be used in a real world problem. While we focus on the interview framing there are many other potential areas that a fair score function would be beneficial. Chat bots are growing in usage (Stöhr et al., 2024), making medical or academic recommendations often rely on analysis of transcripts, the equitable elicitation process could aid deployment of scoring functions and improve question processes. Beyond fairness towards individuals this method could also be used to improve reinforcement learning by considering fairness as regarding the state space, allowing a learner to be unbiased with regards to the score given to different states reducing bias towards previously explored locations.

We also show that LLMs can be used for social science research, we construct a large language model that can accurately reproduce a complex dataset and allow use to test our methods without risking harm to individuals. That said we are not advocating for replacing human subjects research with LLMs and as we noted in section 4.1 these methods should be used in addition to real human interviewers analysis.

There are many possible directions that these results lead, the equitable elicitation process should work for multiple skills simultaneously, and with multi-calibration instead of multi-accuracy. Additionally we hypothesize that if the score function is neural network model distillation can be used to directly imprint the fairness constraints on the weights, reducing complexity and compute costs of the final model.

Ethics Statement

All authors have read and comply with the *ICLR* Code of Ethics. This work involves no human subjects or sensitive data (our training data is all from public sources), and we are unaware of any potential misuse, harm, or bias. No conflicts of interest or compromising sponsorships exist.

Reproducibility Statement

We have made efforts to ensure the reproducibility of our results. Our anonymized code is provided with the submission and will be made public when the paper is published. We do not plan to release our finetuned LLM as a part of the release out of an abundance of caution regarding the uncertainty of licensing public data for LLM training, but will maintain it in an archive and allow other researchers access upon reasonable requests.

References

Divya Mani Adhikari, Alexander Hartland, Ingmar Weber, and Vikram Kamath Cannanure. Exploring llms for automated generation and adaptation of questionnaires. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pp. 1–23, 2025.

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. AI suggestions homogenize writing toward western styles and diminish cultural nuances, 2024. URL http://arxiv.org/abs/2409.11360.

Kanika K Ahuja. Right byte or left out? gender differences in self-presentation among job-seekers on linkedin in india. *Discover Psychology*, 4(1):59, 2024.

Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. Less discriminatory algorithms, 2024. URL https://papers.ssrn.com/abstract=4590481.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016. URL http://arxiv.org/abs/1607.06520.

Inchul Cho, Biyun Hu, and Christopher M Berry. A matter of when, not whether: A meta-analysis of modesty bias in east asian self-ratings of job performance. *Journal of Applied Psychology*, 108 (2):291, 2023.

Felix Chopra and Ingar Haaland. Conducting qualitative interviews with ai. *CESifo Working Paper*, 2023.

Nick Deschacht and Birgitt Maes. Cross-cultural differences in self-promotion: A study of self-citations in management journals. *Journal of Occupational and Organizational Psychology*, 90 (1):77–94, 2017.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1095–1108, 2021.

Cynthia Dwork, Omer Reingold, and Guy N Rothblum. From the real towards the ideal: Risk prediction in a better world. In *4th Symposium on Foundations of Responsible Computing (FORC 2023)*, pp. 1–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2023.

Tyna Eloundou, Alex Beutel, David G Robinson, Keren Gu-Lemberg, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. First-person fairness in chatbots. *ICLR*, 2024.

JIING-LIH FARH, Gregory H Dobbins, and BOR-SHIUAN CHENG. Cultural relativity in action: A comparison of self-ratings made by chinese and us workers. *Personnel psychology*, 44(1): 129–147, 1991.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. *arXiv preprint arXiv:2109.05389*, 2021.

Parikshit Gopalan, Lunjia Hu, Michael P Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. *arXiv preprint arXiv:2210.08649*, 2022.

Parikshit Gopalan, Lunjia Hu, Michael P Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, pp. 60–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2023.

Marius Guenzel, Shimon Kogan, Marina Niessner, and Kelly Shue. AI personality extraction from faces: Labor market implications, 2025. URL https://papers.ssrn.com/abstract=5089827.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. AI generates covertly racist decisions about people based on their dialect. *Nature*, pp. 1–8, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07856-5. URL https://www.nature.com/articles/s41586-024-07856-5. Publisher: Nature Publishing Group.

Christina Ilvento. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*, 2019.

JIAYUANYU and Kevin R Murphy. Modesty bias in self-ratings of performance: A test of the cultural relativity hypothesis. *Personnel Psychology*, 46(2):357–363, 1993.

Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. An algorithmic framework for fairness elicitation, 2020. URL http://arxiv.org/abs/1905.10660.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Cornelius J König, Leifur G Hafsteinsson, Anne Jansen, and Eveline H Stadelmann. Applicants' self-presentational behavior across cultures: Less self-presentation in s witzerland and i celand than in the u nited s tates. *International Journal of Selection and Assessment*, 19(4):331–339, 2011.

Danny D Leybzon, Shreyas Tirumala, Nishant Jain, Summer Gillen, Michael Jackson, Cameron McPhee, and Jennifer Schmidt. Ai telephone surveying: Automating quantitative data collection with an ai interviewer. *arXiv preprint arXiv:2507.17718*, 2025.

Rachel Bowley Lindström. Women's equality day: A look at women in the workplace in 2017. *LinkedIn Official Blog*, 2017. https://www.linkedin.com/blog/member/career/womens-equality-day-a-look-at-women-in

TS Mohr. Why women don't apply for jobs unless they're 100% qualified'harvard business review. *August 25th*, 2014.

Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. In *International conference on machine learning*, pp. 7097–7107. PMLR, 2020.

Raviv Murciano-Goroff. Missing women in tech: The labor market for highly skilled software engineers. *Management Science*, 68(5):3262–3281, 2022.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Guy N. Rothblum and Gal Yona. Probably approximately metric-fair learning, 2018. URL http://arxiv.org/abs/1803.03242.

Christian Stöhr, Amy Wanyu Ou, and Hans Malmström. Perceptions and usage of ai chatbots among students in higher education across genders, academic levels and fields of study. *Computers and Education: Artificial Intelligence*, 7:100259, 2024.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Jimmy Wang, Thomas P Zollo, Richard Zemel, and Hongseok Namkoong. Adaptive elicitation of latent information using natural language. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, 2025.

Tianhao Wang, Zana Buçinca, and Zilin Ma. Learning interpretable fair representations. *arXiv preprint arXiv:2406.16698*, 2024.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020.

Alexander Wuttke, Matthias Aßenmacher, Christopher Klamm, Max M Lang, Quirin Würschinger, and Frauke Kreuter. Ai conversational interviewing: Transforming surveys with llms as adaptive interviewers. *arXiv preprint arXiv:2410.01824*, 2024.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Table 2: Variables and definitions

Variable	Type	Description
j	$\mathbb{N}$	applicant ID/index
y	$\{0,1\}$	Ground truth, 1 if the applicant has the skill, 0 otherwise
$\hat{y}$	$[0,1]$	Estimate of y value gives probability of y being 1
$s()$	str $\to$ str	Response simulator
$q_\psi()$	$\mathcal{X} \to \mathcal{Q}$	question selector, suggests questions based on user responses
$\mathcal{Q}$		Set of questions, (pre-defined)
q_i	$\mathcal{Q}$	Specific question, indexed by number asked
s_i	str	Specific response, indexed by number of responses
s_i^j	str	Specific response, indexed by number of responses (i) and person responding (j)
i	$\mathbb{N}$	index for responses/questions to the same person
$\mathcal{X}$		Transcript (questions and responses) type
x_i	$\mathcal{X}$	Specific Transcript (questions and responses) up to response i
x_i^j	$\mathcal{X}$	Specific Transcript (questions and responses) up to response i for person j
$f_\phi()$	$\mathcal{X} \to \mathbb{R}$	score function, outputs score of candidate based on responses
$f_\phi^*()$	$\mathcal{X} \to \mathbb{R}$	score function, with multi-accuracy
ψ		question selector weights (learned)
ϕ		score function weights (learned)
BERT()	str $\to \mathbb{R}^n$	Our pretrained embedding model, it maps text to vectors (`all-MiniLM-L6-v2`)
ATTN()	$\mathbb{R}^{n\times m} \to \mathbb{R}^n$	GPT style Attention mechanism, maps a sequences of vectors to single vector
Z	$\mathbb{R}$	Certainty, twice distance of $f(x_i^j)$ from .5
$\mathcal{C}$		Set of big 5 personality traits and constant function $\mathbf{1}$
$c()$	$\mathcal{X} \to \mathbb{R}$	Big 5 personality trait score function (`MBTI`), pre-trained neural network, or 1 for $c = \mathbf{1}$
γ	$\mathbb{R}$	Discount factor, set to 1 since we only care about the final evaluation
α	$\mathbb{R}$	Question loss coefficient, controls how quickly we change ψ
β	$\mathbb{R}$	Score function loss coefficient, controls how quickly we change ϕ
ϵ	$\mathbb{R}$	Fairness constraint
l_c	$\mathbb{R}$	Calibration constants
softmax()	$\mathbb{R}^n \to \mathbb{R}^n$	sigmoid function, maps a vector from $\mathbb{R}^n$ to a probability distribution, note n can be 1
sup()		Support
N	$\mathbb{N}$	Number of rollouts
M	$\mathbb{N}$	Rollout Depth
i_{max}	$\mathbb{N}$	Maximum number of questions
B	$\mathbb{N}$	Batch size
B^{fair}	$\mathbb{N}$	Batch size for fairness calibration
δ	$\mathbb{R}$	The failure probability (10^{-6})

A First appendix

A.1 Notation Table

A.2 Algorithm Design Notes

For this section we use $s \in S$ to refer to text vectors from the set of all possible texts (S). These are represented as embeddings in practice. See Table 2 for the definition of each variable used in this paper.

A.3 Proof of Theorem 3.1

We first recall a lemma from Gopalan et al. (2023). Let $\mathcal{T}$ denote the set of functions $g' : \mathbb{R} \to \mathbb{R}$ such that: 1). g' is continuous and monotonically increasing; 2). the range of g' $\text{Im}(g') \supseteq [0,1]$.

Lemma A.1 (Theorem 5.6 from Gopalan et al. (2023)) *For a function g and its derivative $g' \in \mathcal{T}$ whose range is [0,1], and let l_g be its matching loss. Let h^* be the optimal solution to the ℓ_1-*

regularized loss minimization problem:

$$\min_{h \in Lin(C)} [\ell_g(\mathbf{y}, h(\mathbf{x}))] + \epsilon \sum_c |w_c|$$

where $h(x) = \sum_c w_c c(x)$. *Then the function* $g' \circ h : \mathcal{X} \to [0,1]$ *is a* (c, ϵ)*-multiaccurate predictor.*

By Lemma A.1, f^*_ϕ satisifes (C, ϵ)-multiaccuracy:

$$\sup_{c \in \mathcal{C}} |\mathbb{E}_{j \in J}[c(x_i^j) \cdot (f^*_\phi(x_i^j) - y^j)]| \leq \epsilon.$$

Then, as for all $c \in \mathcal{C}$, we have $c \in [0,1]$, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}\sup_{c \in \mathcal{C}} |\mathbb{E}_{j \in J}[c(x_i^j) \cdot (f_\phi(x_i^j) - y^j)]| \leq & \sup_{c \in \mathcal{C}} |\mathbb{E}_{j \in J}[c(x_i^j) \cdot (f^*_\phi(x_i^j) - y^j)]| + \sup_{c \in \mathcal{C}} |\mathbb{E}_{j \in J}[c(x_i^j) \cdot (f_\phi(x_i^j) - f^*_\phi(x_i^j))]| \\ \leq & \epsilon + \epsilon^*.\end{aligned}$$

This implies that f_ϕ is $(C, \epsilon + \epsilon^*)$-multiaccurate.