Measuring Validity in LLM-based Resume Screening

Jane Castleman
Princeton University
Princeton, NJ
janeec@princeton.edu

Zeyu Shen
Princeton University
Princeton, NJ
zs7353@princeton.edu

Blossom Metevier
Princeton University
Princeton, NJ
bmetevier@princeton.edu

Max Springer
Princeton University
Princeton, NJ
maxspringer@princeton.edu

Aleksandra Korolova
Princeton University
Princeton, NJ
korolova@princeton.edu

Abstract

Resume screening is perceived as a particularly suitable task for LLMs given their ability to analyze natural language; thus many entities rely on general purpose LLMs without further adapting them to the task. While researchers have shown that some LLMs are biased in their selection rates of different demographics, studies measuring the validity of LLM decisions are limited. One of the difficulties in externally measuring validity stems from lack of access to a large corpus of resumes for whom the ground truth in their ranking is known and that has not already been used for LLM training. In this work, we overcome this challenge by systematically constructing a large dataset of resumes tailored to particular jobs that are directly comparable, with a known ground truth of superiority. We then use the constructed dataset to measure the validity of ranking decisions made by various LLMs, finding that many models are unable to consistently select the resumes describing more qualified candidates. Furthermore, when measuring the validity of decisions, we find that models do not reliably abstain when ranking equally-qualified candidates and select candidates from different demographic groups at different rates, occasionally prioritizing historically-marginalized candidates. Our proposed framework provides a principled approach to audit LLM resume screeners in the absence of ground truth, offering a crucial tool to independent auditors and developers to ensure the validity of these systems as they are deployed.[1]

1 Introduction

Resume screening has seen widespread AI adoption; faced with ever-increasing application volumes, roughly 90% of employers allegedly now rely on these tools for filtering and ranking candidates [1, 2]. The advent of Large Language Models (LLMs) has accelerated this trend as their capacity to parse unstructured text and generate human-readable reasoning makes them seem exceptionally suited for screening resumes at scale [3]. Such rapid integration into a critical part of the hiring pipeline, however, raises urgent questions about the validity and fairness of the LLMs' decisions.

While the propensity for LLMs to produce biased outcomes in hiring is a well-documented concern [4, 5, 6, 7, 8, 9], this body of research has a critical blind spot. Specifically, most studies focus on fairness, typically measured as the statistical parity of outcomes across demographic groups [4, 10, 11, 12,

[1] Our code can be found at: https://github.com/jane2620/validity-resume-screening, and an extended version of this paper can be found at https://arxiv.org/abs/2602.18550 (including all appendices).

Article presented at the Second IASEAI Conference (IASEAI'26).

13, 14], while overlooking the fundamental question of validity [15]. In this context, validity refers to whether a model's hiring recommendation is fundamentally correct—can it reliably identify the superior candidate based on job-relevant qualifications while ignoring irrelevant factors?

Previous work demonstrates the unreliability of LLMs in medical [16] and legal [17] domains, necessitating performance evaluations across high-stakes deployments, including hiring. Existing work that measures accuracy in LLM resume screening relies on correlations with human ratings [8] or overlap with previously-hired candidates [9]. However, human decision-makers may also make invalid decisions, and are similarly biased [18], resulting in a poor baseline for valid decision-making. Additionally, evaluating validity is notoriously difficult even at the human-level since real-world data lacks a definitive ground truth. There is a need for evaluations that move beyond human agreement to assess the validity of these decision-making systems [15].

Generative AI is reshaping recruiter workflows. LinkedIn released its "Hiring Assistant" in 2026, an automated recruitment tool powered by generative AI, including third-party models from OpenAI [19]. The AI agent evaluates candidates based on recruiter-specified qualifications and a candidate's LinkedIn profile and resume [20]. Similarly, Greenhouse's "AI Screened" [21] feature automates resume evaluation by scoring applicants against required and "nice-to-have" qualifications. Anecdotal evidence also suggests that individual recruiters use "out-of-the-box" LLMs, even if not sanctioned by their employer [22]. At the same time, it is unclear whether the AI companies provide any independently verifiable validity guarantees. Thus, it is important for entities deploying such models to perform independent evaluations of the suitability of the models to their hiring tasks [23, 24]. Furthermore, it is crucial for researchers to conduct independent evaluations in order to inform policymakers and the public about the potential trade-offs of LLM resume screeners [22].

However, conducting such evaluations faces two methodological hurdles. First, the evaluator may have limited ground-truth data available for testing, making it difficult to draw statistically robust conclusions. Second, evaluations using publicly-available data are vulnerable to train-test contamination, where models may have been trained on this data, thus affecting evaluation scores [25, 26, 27, 28].

Thus, our work addresses an urgent need for new frameworks which can generate novel, reliable evaluation scenarios at scale. Specifically, we introduce a framework for systematically constructing novel evaluation tasks with a known ground truth to measure both validity and fairness in resume screening. Our work is inspired by software engineering testing principles that provide a systematic way to evaluate system behavior under controlled conditions. From the perspective of metamorphic testing [29], our framework asks whether the LLM's reasoning follows expected logical rules. From the perspective of mutation testing [30], the LLM is treated as a quality-assurance test that must detect meaningful changes in candidate resume profiles. To our knowledge, ours is one of the first frameworks to provide principled, reproducible, ground-truth-controlled auditing of both validity and fairness in LLM-based decision making.

Concretely, in this work we address the following critical research questions in the face of mounting LLM deployment in resume screening:

RQ1 To what extent are LLM decision-makers valid in their assessments of resumes with a clear ground truth?

RQ2 To what extent are LLMs valid decision-makers in assessing equally-qualified resumes that differ in the explicit or implied demographic information of the applicant?

Statement of Contributions. We introduce a novel framework enabling the simultaneous evaluation of both validity and fairness of LLM decision-making in resume screening. By treating them as distinct, measurable properties, we refocus evaluations towards reliability, revealing shortcomings along each axis. It provides a methodology for generating previously unseen pairs of resumes for whom the ground truth is known, that we then use to conduct a large-scale evaluation of numerous LLMs, providing a clear picture of their current capabilities and shortcomings for resume selection tasks. Our evaluation framework incorporates live job descriptions, detailed in Section 2.1, and can therefore be reused and updated over time without the threat of train-test set contamination [25, 26]. In contrast to prior fairness work that relies on surface-level comparisons of model selection rates, our paired, ground-truth-controlled design is both intellectually principled and practically scalable. Thus, this framework enables independent auditors to measure validity and fairness in automated hiring tools, as required by regulations such as New York's Local Law 144 [31].

2 A Framework for Evaluating LLM Decision-Making

Any resume screening system, human or automated, operates as a measurement instrument attempting to assess candidate qualifications for a job. To evaluate whether such a system works correctly, we draw on established measurement validity concepts from psychometrics [32] and system evaluation literature [33, 34, 35, 36]. Specifically, we focus on two fundamental validity requirements.

Criterion Validity requires that relevant variables impact hiring decisions. In resume screening, this means job-relevant qualifications must influence who gets selected (e.g., technical skills or years of experience). **Discriminant validity** requires that irrelevant variables do not impact hiring decisions. Characteristics like demographic attributes (race, gender) or unrelated hobbies should not influence selection for most jobs. These two requirements capture what it means for a hiring system to be valid: it must respond appropriately to relevant information while ignoring irrelevant information.

An Idealized Scenario for Ground Truth. To operationalize these validity concepts, we consider an idealized scenario where we can establish definitive ground truth. Consider the setting where all candidate attributes can be (disjointly) divided into relevant and irrelevant qualifications for job performance. Under this idealized scenario, validity has clear implications: a candidate who possesses strictly more relevant qualifications than another should be preferred by any valid decision maker, and candidates who possess identical relevant qualifications should be treated as equally qualified. While real-world hiring decisions involve more nuanced complexity, these principles represent necessary conditions that any rational resume screening process must satisfy.

Formal Setup. Let $\mathcal{C} = \{c_1, \ldots, c_n\}$ be a set of candidates and $\mathcal{X}$ be the universe of all possible attributes a candidate can possess (skills, degrees, years of experience, name, demographics, etc.). We map any candidate c to their attributes via $f : \mathcal{C} \to 2^{\mathcal{X}}$.

For any specific job j, attributes partition into:

- **Relevant Attributes** (X_j^+): Qualifications meaningful for job performance
- **Irrelevant Attributes** (X_j^-): Characteristics that should not influence hiring (demographics, name, unrelated hobbies)

These sets are mutually exclusive and exhaustive: $X_j^+ \cup X_j^- = \mathcal{X}$ and $X_j^+ \cap X_j^- = \emptyset$.

Based on this partition, we define when one candidate should be preferred over another, establishing axioms that any valid resume screening process must satisfy.

Definition 1 (Axioms for Valid Resume Screening)**.** *For any two candidates $c, c' \in \mathcal{C}$ and job j, we define a preference relation $\succeq_j$ that ranks candidates based only on relevant attributes:*

1. ***Strict Preference*** *($c \succ_j c'$): Candidate c should be strictly preferred over c' if c possesses all of c''s relevant attributes plus additional ones:*
$$\big(f(c) \cap X_j^+\big) \supset \big(f(c') \cap X_j^+\big) \Rightarrow c \succ_j c'.$$
2. ***Indifference*** *($c \sim_j c'$): Candidates c and c' should be treated as equally qualified if they possess identical relevant attributes:*
$$\big(f(c) \cap X_j^+\big) = \big(f(c') \cap X_j^+\big) \Rightarrow c \sim_j c'.$$

We crucially note that our definitions use implication ($\Rightarrow$) rather than equivalence as these conditions are sufficient to establish preference (indifference) but may not be the only way to do so in complex settings. Our framework herein treats all relevant attributes as equally important, but the axioms represent necessary principles any rational system must follow regardless. Moreover, we note that because Definition 1 relies on an idealized subset-nesting of qualification, our metrics should be interpreted as conservative estimates rather than exact measures of real-world validity.

2.1 Constructing Test Cases at Scale

Our framework enables generating pairs of resumes with known ground truth that have not been used in LLM training. This addresses a critical challenge for independent auditors, regulatory bodies,

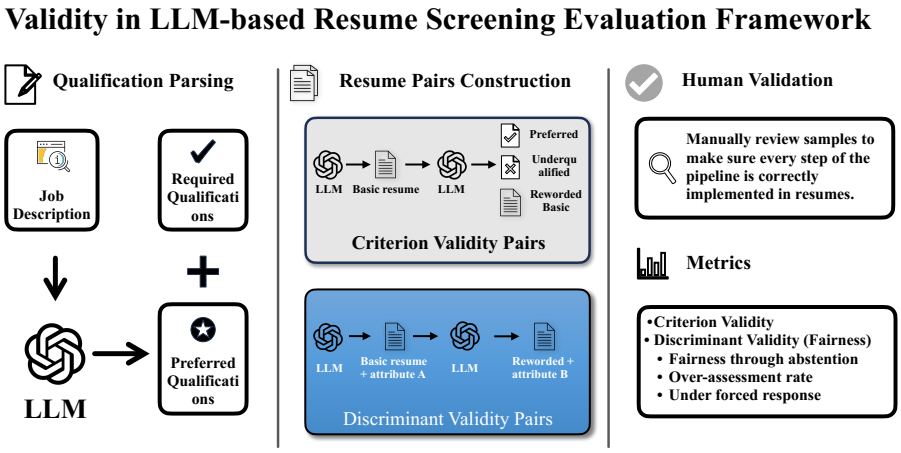


Figure 1: Based on a job description, we create a base resume c that meets all required qualifications. Then, we use LLMs to generate more-qualified candidates $c^+ \succ_j c$, less-qualified candidates $c \succ_j c^-$, and equally-qualified candidates with varying demographic information.

and organizations evaluating out-of-the-box LLM resume screeners. More specifically, this enables validity testing without access to proprietary training data or historical hiring records. Our approach works with any set of job descriptions, including an organization's own internal postings, and can be re-run over time to avoid train-test contamination.

High-Level Generation Procedure. Our construction pipeline, illustrated in Figure 1, consists of four stages. Given a job description j, we first parse it into structured lists of required qualifications ($Q_j^{\texttt{req}}$) and preferred qualifications ($Q_j^{\texttt{pref}}$), defining $X_j^+ = Q_j^{\texttt{req}} \cup Q_j^{\texttt{pref}}$. We then generate a base resume c that satisfies exactly the required qualification, such that $(f(c) \cap X_j^+) = Q_j^{\texttt{req}}$. From this base resume, we construct unequal pairs for testing criterion validity by creating a less-qualified variant c^-, formed by removing k randomly selected required qualifications, and a more-qualified variant c^+, formed by adding k randomly selected preferred qualifications. This yields pairs where $c^+ \succ_j c \succ_j c^-$. Finally, to test discriminant validity, we duplicate a base resume, reword it while maintaining its qualifications, and append different demographic signals to each copy. These pairs satisfy $c_A \sim_j c_B$, differing only in irrelevant attributes. The order of resumes is randomized.

Implementation Details. We instantiate this framework using publicly available job descriptions from Greenhouse,[2] but the approach generalizes to any job board or internal postings. We scraped 186 job descriptions across 25 categories. While we instantiate the framework using Greenhouse, we find that our results transfer to roles sourced from LinkedIn and Indeed, as well as to real-world resumes to further confirm the framework's generalizability across occupational domains and resume styles. For scalable qualification extraction and resume generation, we use LLMs as data processing tools. To mitigate model-specific biases, we generate variants using both `Gemini-2.5-Pro` and `Claude-Sonnet-4`, particularly important given that LLM evaluators may favor their own outputs [37, 38].

Following established bias measurement methodologies [4, 6, 22], we append demographic signals to each base resume for $|G| = 4$ groups: {Black, White} $\times$ {man, woman}. We distinguish between *implicit signals*, which are demographic associations irrelevant to hiring, such as using names with high racial/gendered correlation, and *explicit signals*, which directly communicate demographic characteristics through identifiers, such as "National Association of {Demographic Group} Professionals: {Job Title} Emerging Leader Award" [39, 40]. These signals are designed to be of equal relevance to the candidate's professional qualifications across demographic groups.

[2] `https://www.greenhouse.com/`

2.2 Measuring Validity with Test Cases

Given the test cases constructed above and an LLM-based resume screener to evaluate, we now define metrics that quantify validity. We model the screener as a pairwise comparison function $P_j : \mathcal{C} \times \mathcal{C} \to \mathcal{C} \cup \{\perp\}$ that either selects one candidate or abstains ($\perp$). While many real-world systems are configured to be decisive [41, 42, 43], we first analyze systems that can abstain, then address forced-choice scenarios. We focus on pairwise comparisons for clarity; Section 4 discusses generalization to ranking larger pools.

A perfectly valid screener would satisfy: $P_j(c, c') = c$ when $c \succ_j c'$, $P_j(c, c') = c'$ when $c' \succ_j c$, and $P_j(c, c') = \perp$ when $c \sim_j c'$. Our metrics measure deviations from this ideal behavior, directly connecting to our research questions.

Criterion Validity on Unequal Pairs (RQ1). Consider a test set S containing pairs (c^+, c^-) where $c^+ \succ_j c^-$— one candidate is demonstrably more qualified than the other. A valid decision-maker should choose the more qualified candidate in every case. We measure the fraction of instances across the test set in which P_j correctly selects the more qualified candidate:

$$\texttt{CriterionValidity}(P_j, S) = \frac{1}{|S|} \sum_{(c^+, c^-) \in S} \mathbb{1}(P_j(c^+, c^-) = c^+). \tag{1}$$

When the model fails ($P_j(c^+, c^-) \neq c^+$), we distinguish two error types. In an *unjustified selection*, the decision-maker chooses the less qualified candidate, and in an *unjustified abstention* it fails to make a decision despite clear candidate superiority. We quantify these as proportions of all errors:

$$\texttt{UnjustifiedSelection}(P_j, S) = \frac{\sum_S \mathbb{1}(P_j(c^+, c^-) = c^-)}{\sum_S \mathbb{1}(P_j(c^+, c^-) \neq c^+)}, \tag{2}$$

$$\texttt{UnjustifiedAbstention}(P_j, S) = \frac{\sum_S \mathbb{1}(P_j(c^+, c^-) = \perp)}{\sum_S \mathbb{1}(P_j(c^+, c^-) \neq c^+)}.$$

We here note that when candidates come from different demographic groups, some fairness literature requires errors to be group-independent [10]. Within our framework, let $S_{A,B}$ be pairs where the more qualified candidate is from group g_A and the less qualified from group g_B (i.e., $c_A \succ_j c_B$ with $c_A \in g_A$, $c_B \in g_B$). The correct decision is $P_j(c_A, c_B) = c_A$. Any other outcome represents *over-assessing* candidate c_B relative to their actual qualifications. We measure the fraction of instances across the test set in which P_j incorrectly chooses a less qualified candidate or abstains from choosing the more qualified candidate:

$$\texttt{OverAssessment}(P_j, S_{A,B}, g_B) = 1 - \texttt{CriterionValidity}(P_j, S_{A,B}). \tag{3}$$

A fair system should exhibit low and approximately equal over-assessment rates across all demographic groups. Significant disparities (e.g., consistently over-assessing White candidates relative to Black candidates or vice versa) indicate systematic, group-dependent errors rather than random mistakes.

Discriminant Validity on Equal Pairs (RQ2). Now consider test set $E_{A,B}$ containing equally-qualified pairs (c_A, c_B) where $c_A \sim_j c_B$ but the two possess different irrelevant attributes: $(f(c_A) \cap X_j^-) \neq (f(c_B) \cap X_j^-)$. A valid decision-maker should recognize their equivalence and abstain, as there is no qualification-based reason to prefer either candidate. Any non-abstention suggests the decision was influenced by irrelevant information. We measure the fraction of instances across the test set in which P_j abstains from arbitrarily choosing one of two equally qualified candidates (higher is better):

$$\texttt{DiscrimValidity}(P_j, E_{A,B}) = \frac{1}{|E_{A,B}|} \sum_{(c_A, c_B) \in E_{A,B}} \mathbb{1}(P_j(c_A, c_B) = \perp). \tag{4}$$

When the model selects one candidate over an equally-qualified peer, that candidate is over-assessed. As for unequal pairs, we measure the fraction of instances across the test set in which P_j incorrectly chooses a candidate of a certain demographic over an equally qualified candidate of another demographic:

$$\texttt{OverAssessment}(P_j, E_{A,B}, g_B) = \frac{1}{|E_{A,B}|} \sum_{(c_A, c_B) \in E_{A,B}} \mathbb{1}(P_j(c_A, c_B) = c_B). \tag{5}$$

Forced-Choice Scenario. In practice, many systems prohibit abstention ($P_j : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$), requiring a definitive ranking to select top-k candidates. For unequal pairs, we still measure `CriterionValidity` (Equation 1). For equal pairs where either candidate could reasonably be chosen, fairness requires approximately equal selection rates across demographics:

$$\texttt{SelectionRate}(g_A) = \frac{1}{|E_{A,B}|} \sum_{(c_A, c_B) \in E_{A,B}} \mathbb{1}(P_j(c_A, c_B) = c_A).$$

In the pairwise setting, each group should have $\texttt{SelectionRate}(\cdot) \approx 0.5$ if the system is unbiased.

3 To What Extent Are LLMs Valid Decision-Makers?

Using our framework, we evaluate popular LLMs and find that in many cases models do not have `CriterionValidity` that exceeds 0.95 when given differently-qualified resumes, and struggle to show indifference between equally-qualified resumes. While our experiments show that validity generally improves with model scale, we caution that high performance on our metrics is a necessary but not sufficient condition for real-world validity. Additionally, we observe unequal selection rates in certain occupations, such as Software Engineering, where some models favor Black and women candidates, suggesting evidence of over-alignment.

Experimental Setup. We focus on frontier models ranging in size, release date, and developer to understand how validity varies with parameter scale and over time. Therefore, we include `Claude Sonnet 4` [44], `Deepseek Chat v3.1` [45], `Gemini-2.0-Flash` [46], `Gemini-2.5-Pro` [47], `Gemma-3-12B` [48], `GPT-4o-mini` [49], `GPT-5` [50], `Llama-3.1-8B` [51], and `Llama-3.3-70B` [52] in our evaluation. For all open models, we used instruction-tuned and aligned versions, as we are most interested in evaluating validity after alignment and these are the consumer-facing models typically adopted out-of-the-box. To control for prompt sensitivity, we use three variations of system prompts and resume comparison prompts, and find our results are similar across the prompts we test (average standard deviation is approximately 4%).

Validation of LLM-Generated Resumes. First, we manually validated our LLM-generated resumes (Section 2.1), reviewing a random subset of 54 differently-qualified and 26 reworded resume pairs. During this process, we identified and removed four pairs with minor errors, e.g., cases where the skill change only occurred in the resume summary. We also filtered out pairs of differently-qualified resumes with changes containing only trivial edits, defined as those with fewer than 120 characters difference between resumes. After applying changes, the maximum absolute change to `CriterionValidity` was 0.03 for GPT-4o-mini, with an average change of 0.004.

3.1 Evaluating Criterion Validity of LLM Decision-Making

We first investigate the extent to which LLMs are valid decision-makers in their assessments of resumes with a clear ground truth, answering **RQ1**. We calculate `CriterionValidity` (Eq. (1)), which measures the proportion of correct decisions given a clearly better-qualified candidate.

In Table 1, we aggregate `CriterionValidity` across the job titles we study, taking the average `CriterionValidity` over all unequal pairs. Here, the random baseline is $\texttt{CriterionValidity} = 1/2$, because the best naive model should randomly choose one of the two candidates and never abstain. When disaggregating results by job, we find that `CriterionValidity` changes with job but broader trends remain the same: models struggle to choose the more qualified candidate when resumes are more similar and models do not always select the more qualified resume.

We find that larger, newer models have higher, though still not perfect, `CriterionValidity`. For example, `Claude-Sonnet-4` has an average `CriterionValidity` of 0.96, meaning the model successfully selects more-qualified candidates in 96% of pairs we test with different relevant qualifications. Similarly, `GPT-5`, `Gemini-2.5-Pro` and `Gemini-2.0-Flash` have a $\texttt{CriterionValidity} > 0.90$ when averaged over the number of differing relevant qualifications ($k \in 1, 2, 3$). On the other hand, both `Llama-3.1-8B` and `Llama-3.3-70B` struggle to distinguish between differently-qualified candidates even at $k = 3$.

Error Rate Breakdown. For each model, we categorize its incorrect decisions into unjustified abstentions (`UnjustifiedAbstentions`, Eq. (2)) and unjustified selections (`UnjustifiedSelection`,

Table 1: `CriterionValidity` scores conditioned on the generating model (C: Claude, G: Gemini) and the number of differing relevant qualifications ($k \in \{1, 2, 3\}$). Cells with values lower than 0.95 are colored maroon (higher is better).

	$k = 1$		$k = 2$		$k = 3$	
Model (Evaluator)	C	G	C	G	C	G
Llama 3.1 8B	0.64	0.67	0.74	0.77	0.73	0.81
Llama 3.3 70B	0.44	0.50	0.66	0.77	0.76	0.88
Gemma 3 12B	0.74	0.82	0.93	0.97	0.96	0.99
Gemini 2.0 Flash	0.73	0.84	0.90	0.96	0.95	0.99
Gemini 2.5 Pro	0.82	0.88	0.94	0.96	0.98	0.98
GPT-4o-Mini	0.50	0.59	0.74	0.90	0.86	0.97
GPT-5	0.83	0.90	0.93	0.98	0.98	0.99
Claude Sonnet 4	0.87	0.93	0.97	0.99	0.99	0.99
DeepSeek 3.1	0.66	0.73	0.84	0.93	0.91	0.95

Eq. (2)), shown in Table 2. We find that for the majority of models, most errors stem from over-abstention. While this increases the cost of human intervention if manual review of ambiguous cases is required, it significantly reduces the cost of false negatives where qualified candidates are unjustly denied opportunities. Consequently, we argue that in high-stakes decision-making, models should prioritize abstention to minimize the potential for unfair harm against candidates.

Effects of Generating Model on Validity. We find that Gemini and Claude do not necessarily perform better on resumes generated using the same model. However, the models we test have slightly lower validity on resumes generated by Claude than Gemini, with the largest performance gap at -16% for `GPT-4o-mini` ($k = 2$). While validity gaps are relatively small on average, the significant differences for certain models emphasizes the need to use multiple models when constructing synthetic evaluations.

Table 2: Breakdown of error rates when models fail to select the more-qualified candidate. Proportions are averaged across Claude- and Gemini-constructed variants.

Model	Unjustified Selections	Unjustified Abstentions
Llama 3.1 8B	0.65	0.35
Llama 3.3 70B	0.06	0.94
Gemma 3 12B	0.12	0.88
Gemini 2.0 Flash	0.16	0.84
Gemini 2.5 Pro	0.06	0.94
GPT-4o Mini	0.07	0.93
GPT-5	0.10	0.90
Claude Sonnet 4	0.07	0.93
DeepSeek 3.1	0.06	0.94
Median	0.07	0.93

Implications. Our results emphasize that the functionality of AI systems should not be assumed [15]. While the majority of models have a `CriterionValidity` > 0.95 when $k = 3$, many have a validity < 0.90 at $k = 1$ and < 0.95 at $k = 2$. The performance at $k = 1$ is particularly significant as it represents the model's ability to distinguish between candidates with minor differences in qualifications, which we hypothesize is common in hiring for competitive roles. In real-world deployments where firms screen hundreds of thousands of resumes for a single posting [1], many applicants likely cluster near this decision boundary. If models struggle to maintain validity in these $k = 1$ scenarios, they could incorrectly deny thousands of qualified applicants, harming both firms and applicants.

The performance disparities across models necessitate validity audits for any commercial LLM hiring tool. Lacking both minimum standards and transparent testing data, customers cannot meaningfully assess a tool's reliability. Our methodology offers a path forward by allowing customers to generate job-specific evaluation scenarios on demand, informing pre- and post-deployment monitoring.

3.2 Evaluating Discriminant Validity of LLM Decision-Making

Next, we examine whether LLMs maintain discriminant validity when demographic signals are introduced while qualifications are held constant. In particular, we measure `DiscrimValidity` and `SelectionRates` in the presence of explicit demographic information (e.g., awards or organizational

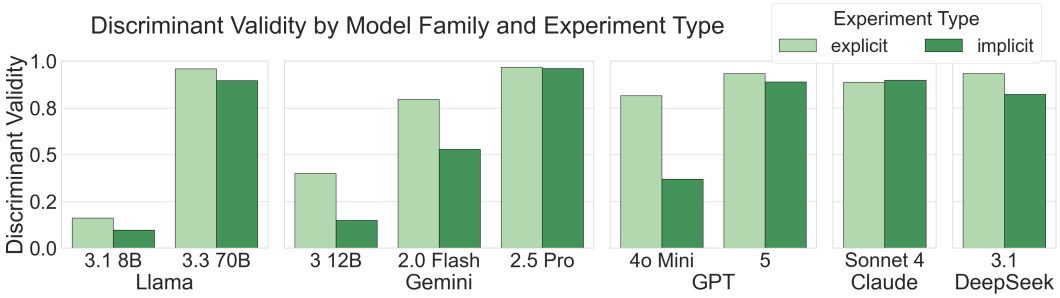


Figure 2: `DiscrimValidity` by model and candidate demographic information type, measuring model abstention rates in deciding between equally-qualified candidates. Model error occurs when a model selects one of the two candidates rather than abstaining.

affiliations that signal race or gender) and implicit demographic information (e.g., names that signal race or gender, Section 2.1). This setup allows us to assess whether models abstain or select candidates differently when only irrelevant attributes vary.

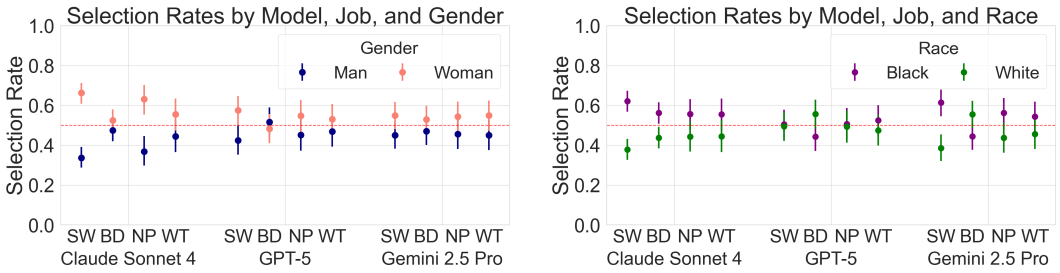
Discriminant Validity Through Abstentions. We find that `DiscrimValidity` (Eq. (4)) varies significantly, shown in Figure 2. We aggregate `DiscrimValidity` across the job titles we study, taking the average `DiscrimValidity` over all equal pairs. As with `CriterionValidity`, the results are similar across jobs. `DiscrimValidity` is lower than `CriterionValidity`, in line with related work showing that abstention is challenging for models [41, 53]. Notably, `DiscrimValidity` generally decreases when demographics are indicated explicitly through awards and organizations rather than implicitly through name.

For the majority of models, `DiscrimValidity` < 0.9 in the presence of demographic information, meaning that more than 10% of candidates are chosen arbitrarily over an equally-qualified candidate. For smaller, open models, `DiscrimValidity` is comparable to or worse than when randomly guessing an outcome, emphasizing the need to improve abstention in LLMs. We manually inspected model outputs, including the decisions and brief justification. We find that model justifications sometimes acknowledged equal qualifications but relied on untrue rationale to ultimately justify candidate selection, such as "better formatting" when resumes are formatted similarly. We hypothesize that these arbitrary decisions are due to model training for decisive question-answering [54, 41].

Over-Assessment by Demographic Group. We measure the `OverAssessment` (Eq. (3) and (5)) by demographic group of the selected candidate. When candidates are not equally qualified, we find that candidates of different races and genders are over-assessed at similar rates. In equally qualified candidate pairs, we find that White men are less likely to be over-assessed (rate of 0.14) than candidates of other demographics (average rate of 0.24) we tested when candidate demographic is indicated explicitly, such as through extracurriculars.

Selection Rates Under Forced Decisions. As in previous work measuring unfairness in AI resume screening systems [12, 4, 6], we find that not all models select candidates at equal rates, shown in Figure 3. However, we find that models select White men at the lowest rates, and differences in `SelectionRate` do not uniformly improve with model size, particularly for `Claude-Sonnet-4`. On the other hand, `GPT-5` shows negligible differences in selection rates. Disparities in `SelectionRate` are less pronounced, or directionally reversed, for the Business Development Representative - German Speaking (BD), suggesting some effect of the language requirement explicit in the job title. Allowing models to abstain changes relative selection rates between equally qualified candidates of different demographics. As with forced decisions, women are more likely than men to be chosen, while Black candidates are more likely than White candidates.

Implications. Low model discriminant validity, measured through abstentions, reflects the difficulty of model decision-making under uncertainty [41]. We echo arguments that abstentions can be a crucial tool for fairness by reducing arbitrary decisions in high-stakes contexts [55], and our framework offers a tool to measure fairness with and without abstentions. Unequal selection rates, in this case biased towards minority groups, may suggest an over-correction to address bias against minority groups found in previous works [4, 7]. Still, some models, such as GPT-5, show more balanced selection rates. Our results are limited to binary gender and Black and White applicants, but our



(a) Gender (b) Race

Figure 3: We plot models' `SelectionRate` for Software Engineer (SW), Business Development Representative, German Speaking (BD), Nurse Practitioner (NP), and Wind Turbine Technician (WT). The expected `SelectionRate` given pairwise comparisons is 0.5.

framework could be used to generate candidates varying in other protected attributes or with more diverse gender and racial identities. Given the heterogeneity across models and jobs, it is crucial for downstream users to evaluate the discriminant validity of out-of-the-box LLMs for their job context.

3.3 Detailed Analysis of Model Error Cases

To better understand model performance, we conduct validation studies and investigate common error cases. A detailed analysis of errors can be found in the extended paper [56].

Performance Across Job Types. Table 3 presents the results of two validation studies examining the generalizability of our findings beyond the primary evaluation setting. Across broader occupations, sourced from LinkedIn and Indeed, criterion validity remains broadly consistent, with a modest average decrease of 0.06 and an increase in unjustified abstentions by 0.05 relative to the CS roles of this section. We further evaluate on anonymized real-world resumes from Reddit's `r/resumes` community, which introduce the linguistic diversity and contextual ambiguity characteristic of real hiring settings. Here, criterion validity decreases by 0.09 and unjustified abstentions increase by 0.10 on average. Importantly, the relative ordering of models and the direction of key findings remain stable across both settings, though absolute validity scores should be interpreted here as optimistic relative to real hiring conditions.

Table 3: Model performance in Original, Software Engineering (SWE); Non-CS, Nurse Practitioner (Nurse); Real-World, Software Engineering (RW). Best score for each category is **bolded**. We average over the number of different relevant qualifications ($k = [1, 2, 3]$) for brevity.

Model ID	Crit. Validity ↑			Unjust. Abstent. ↓			Disc. Validity ↑		
	SWE	Nurse	RW	SWE	Nurse	RW	SWE	Nurse	RW
Llama 3.1 8B	0.74	**0.78**	0.66	0.12	**0.12**	0.21	0.18	0.23	**0.50**
Llama 3.3 70B	**0.72**	0.59	0.43	**0.26**	0.40	0.57	0.96	**0.99**	0.98
Gemma 3 12B	**0.96**	0.84	0.81	**0.04**	0.14	0.18	0.37	0.61	**0.66**
Gemini 2.0 Flash	0.89	**0.90**	**0.90**	**0.09**	**0.09**	**0.09**	0.75	0.54	**0.82**
Gemini 2.5 Pro	**0.94**	0.86	0.90	**0.06**	0.12	0.10	0.99	0.98	**1.00**
GPT-4o-Mini	**0.80**	0.74	0.70	**0.20**	0.25	0.30	0.72	0.46	**0.83**
GPT-5	**0.94**	0.90	0.87	**0.04**	0.07	0.13	**0.98**	0.84	0.92
Claude Sonnet 4	**0.97**	0.93	0.92	**0.03**	0.05	0.08	0.95	0.94	**0.82**
DeepSeek 3.1	**0.90**	0.82	0.83	**0.09**	0.16	0.17	0.90	0.93	**0.98**

Errors by Qualification Type. Job descriptions frequently require both "soft" and "hard" skills. Soft skills are unmeasurable traits such as social awareness and passion while hard skills are verifiable competencies with specific systems [57]. To better understand errors by type, we filter examples based on the changed qualifications. For soft skills, we use keywords such as "passion" and "curiosity." For educational credentials, we use keywords including "Bachelor's" and "PhD". We find that average criterion validity is higher when resumes differ in educational credentials rather than soft skills (0.85

vs. 0.81), while the proportion of high consensus errors (>50% of models making the same error) is about the same for both qualification types (0.13 vs. 0.12).

4 Discussion & Conclusion

A central insight of this work is that validity and fairness are analytically separable: a model can be valid yet unfair, or consistent yet invalid. Many prior studies collapse these into a single moral question, whereas our framework treats them as distinct, measurable properties. This shifts the conversation from ethical aspiration to empirical reliability, which is precisely what high-stakes hiring systems require. Our findings show that validity in LLM-based resume screening should not be assumed, even for frontier models. While performance improves with model scale and with clearer qualification differences, both criterion and discriminant validity vary across settings, illustrating the necessity for more controlled testing. Our work contributes a principled framework for evaluating validity under controlled conditions and varying levels of task difficulty. These template-based methods provide a crucial tool for on-demand, scalable assessment across downstream use cases. In the rest of this section, we examine implications of our findings, discuss how to extend our framework for top-k rankings and longitudinal evaluations, and end with limitations and future directions.

Validity and Over-Alignment. Our results reveal a complex tension between model validity and alignment. While we observe that validity generally scales with model size, our evaluations of discriminant validity uncover evidence of unequal selection rates that persists despite high criterion validity (Figure 3). In contrast with previous work [4], unequal selection rates favor Black and women candidates when candidates are equally qualified, suggesting that current post-training techniques designed to mitigate bias may be inducing a new form of invalidity where demographic signals override relevant qualifications [58]. Future evaluations must therefore treat validity and fairness not as orthogonal metrics, but as coupled objectives, ensuring that bias mitigation efforts do not compromise the fundamental reasoning capabilities required for accurate decision-making.

Pairwise to Global Rankings. While our framework evaluates pairwise comparisons, practical deployment often requires ranking larger pools to identify the top-k candidates. Pairwise validity is sufficient for this, as it ensures the model's preferences form a transitive structure (specifically a DAG), guaranteeing a coherent total ordering via topological sorting. Without pairwise validity, preference cycles make a top candidate mathematically undefined. Once pairwise validity is established, the LLM becomes a reliable comparator for efficient sorting algorithms or can be integrated into continuous scoring systems like Elo ratings [59] or tournament structures [60].

Longitudinal Evaluations. Our framework supports longitudinal evaluations by enabling repeated testing under evolving job descriptions and model versions. Because static benchmarks are vulnerable to rapid train-test contamination [61], they provide limited insight into how model behavior changes over time. By sourcing live job descriptions and constructing controlled counterfactual resume pairs, our approach mirrors metamorphic [29] and mutation testing [30] to generate novel, contamination-free evaluation sets. Furthermore, our framework explicitly controls task difficulty (via the number of qualification edits, k), incorporating principles from software and standardized testing [62, 30, 63].

Limitations. Our approach provides necessary but not sufficient conditions for validity; models that perform well under our metrics may still fail on subtler, real-world distinctions. Moreover, our study is limited to four demographic groups varying in race and binary gender. Studying broader demographics or attributes such as religion could reveal more nuanced effects on validity. Our ground truth captures discrete qualification differences under controlled conditions, which may not reflect the full complexity of real resumes or demographics. To maintain scalability and adaptability, we rely on LLMs to parse and generate resumes, which could introduce errors that propagate downstream. While this enables consistent comparisons across job descriptions and time, synthetic resumes may differ subtly from human-written ones, and results can vary across generation models. Consequently, synthetic evaluations should not be interpreted as lower bounds on real-world performance.

Future work should apply our evaluation to broader model setups, including those with confidence threshold re-weighting, fine-tuning, or abstention enforcement under uncertainty, and measure their effects on validity and fairness. By doing so, our framework serves as a rigorous foundation for the automated compliance testing of real-world decision-making systems.

Acknowledgments

We thank the IASEAI 2026 anonymous reviewers for their thoughtful feedback and constructive suggestions, which helped improve the clarity, presentation, and empirical evaluation of this work. This work was supported in part by the National Science Foundation grants CNS-1956435, CNS-2344925, and by the Alfred P. Sloan Research Fellowship for A. Korolova.

References

[1] U. Amitabh and A. Ansari, "Hiring with AI doesn't have to be so inhumane." `https://www.weforum.org/stories/2025/03/ai-hiring-human-touch-recruitment/`, 2025.

[2] E. Gorelick, "Recruiters Use A.I. to Scan Résumés. Applicants Are Trying to Trick It.," *The New York Times*, Oct. 2025.

[3] T. Korbak, M. Balesni, E. Barnes, Y. Bengio, J. Benton, J. Bloom, M. Chen, A. Cooney, A. Dafoe, A. Dragan, S. Emmons, O. Evans, D. Farhi, R. Greenblatt, D. Hendrycks, M. Hobbhahn, E. Hubinger, G. Irving, E. Jenner, D. Kokotajlo, V. Krakovna, S. Legg, D. Lindner, D. Luan, A. Mądry, J. Michael, N. Nanda, D. Orr, J. Pachocki, E. Perez, M. Phuong, F. Roger, J. Saxe, B. Shlegeris, M. Soto, E. Steinberger, J. Wang, W. Zaremba, B. Baker, R. Shah, and V. Mikulik, "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety," 2025.

[4] K. Wilson and A. Caliskan, "Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, p. 1578–1590, Oct. 2024.

[5] K. Wilson, M. Sim, A.-M. Gueorguieva, and A. Caliskan, "No Thoughts Just AI: Biased LLM Hiring Recommendations Alter Human Decision Making and Limit Human Autonomy," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 8, p. 2692–2704, Oct. 2025.

[6] K. Glazko, Y. Mohammed, B. Kosa, V. Potluri, and J. Mankoff, "Identifying and Improving Disability Bias in GPT-Based Resume Screening," in *The 2024 ACM Conference on Fairness Accountability and Transparency*, FAccT '24, p. 687–700, ACM, June 2024.

[7] H. Iso, P. Pezeshkpour, N. Bhutani, and E. Hruschka, "Evaluating bias in LLMs for job-resume matching: Gender, race, and education," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)* (W. Chen, Y. Yang, M. Kachuee, and X.-Y. Fu, eds.), (Albuquerque, New Mexico), pp. 672–683, Association for Computational Linguistics, Apr. 2025.

[8] S. Vaishampayan, H. Leary, Y. B. Alebachew, L. Hickman, B. A. Stevenor, W. Beck, and C. Brown, "Human and LLM-based resume matching: An observational study," in *Findings of the Association for Computational Linguistics: NAACL 2025* (L. Chiruzzo, A. Ritter, and L. Wang, eds.), (Albuquerque, New Mexico), pp. 4808–4823, Association for Computational Linguistics, Apr. 2025.

[9] E. Anzenberg, A. Samajpati, S. Chandrasekar, and V. Kacholia, "Evaluating the Promise and Pitfalls of LLMs in Hiring Decisions." `https://arxiv.org/abs/2507.02087`, 2025.

[10] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[11] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Comput. Surv.*, vol. 54, July 2021.

[12] C. Wilson, A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, and F. Polli, "Building and auditing fair algorithms: A case study in candidate screening," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, (New York, NY, USA), p. 666–677, Association for Computing Machinery, 2021.

[13] Y. Hu, Z. Lyu, L. Bai, and L. Cui, "FairWork: A Generic Framework For Evaluating Fairness In LLM-Based Job Recommender System," in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Padua Italy), p. 3964–3968, ACM, July 2025.

[14] P. Seshadri, H. Chen, S. Singh, and S. Goldfarb, "Small Changes, Large Consequences: Analyzing the Allocational Fairness of LLMs in Hiring Contexts," in *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[15] I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst, "The Fallacy of AI Functionality," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, (New York, NY, USA), p. 959–972, Association for Computing Machinery, 2022.

[16] A. Gourabathina, W. Gerych, E. Pan, and M. Ghassemi, "The medium is the message: How non-clinical information shapes clinical decisions in llms," in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, (New York, NY, USA), p. 1805–1828, Association for Computing Machinery, 2025.

[17] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho, "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools," *Journal of Empirical Legal Studies*, vol. 22, no. 2, p. 216–242, 2025.

[18] M. Bertrand and S. Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, vol. 94, p. 991–1013, Sept. 2004.

[19] LinkedIn, "Meet Hiring Assistant." `https://business.linkedin.com/hire/hiring-assistant`, 2026.

[20] LinkedIn, "LinkedIn Hiring Assistant Data Security & Compliance." `https://business.linkedin.com/content/dam/lem/business/en/hire/ai-transparency/liha-white-paper.pdf`, 2026.

[21] Greenhouse, "AI Screened integration." `https://support.greenhouse.io/hc/en-us/articles/29122364864539-AI-Screened-integration`, Apr. 2025.

[22] L. Yin, D. Alba, and L. Nicoletti, "OpenAI's GPT Is a Recruiter's Dream Tool. Tests Show There's Racial Bias," *Bloomberg*, 2024.

[23] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435, 2019.

[24] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: evaluating claims and practices," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, (New York, NY, USA), p. 469–481, Association for Computing Machinery, 2020.

[25] O. Sainz, J. Campos, I. García-Ferrero, J. Etxaniz, O. L. de Lacalle, and E. Agirre, "NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark," in *Findings of the Association for Computational Linguistics: EMNLP 2023* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 10776–10787, Association for Computational Linguistics, Dec. 2023.

[26] M. Riddell, A. Ni, and A. Cohan, "Quantifying contamination in evaluating code generation capabilities of language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), (Bangkok, Thailand), pp. 14116–14137, Association for Computational Linguistics, Aug. 2024.

[27] Y. Dong, X. Jiang, H. Liu, Z. Jin, B. Gu, M. Yang, and G. Li, "Generalization or memorization: Data contamination and trustworthy evaluation for large language models," in *Findings of the Association for Computational Linguistics: ACL 2024* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), (Bangkok, Thailand), pp. 12039–12050, Association for Computational Linguistics, Aug. 2024.

[28] S. Balloccu, P. Schmidtová, M. Lango, and O. Dušek, "Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 2024.

[29] T. Y. Chen, S. C. Cheung, and S. M. Yiu, "Metamorphic testing: A new approach for generating next test cases." `https://arxiv.org/abs/2002.12543`, 2020.

[30] A. J. Offutt and J. Pan, "Automatically detecting equivalent mutants and infeasible paths," *Software testing, verification and reliability*, vol. 7, no. 3, pp. 165–192, 1997.

[31] NYC Consumer and Worker Protection, "Automated Employment Decision Tools (AEDT)," 2021. LOCAL LAWS OF THE CITY OF NEW YORK FOR THE YEAR 2021, No. 144.

[32] S. Messick, "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning," *ETS Research Report Series*, vol. 1994, no. 2, p. i–28, 1994.

[33] A. Z. Jacobs and H. Wallach, "Measurement and fairness," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, p. 375–385, ACM, Mar. 2021.

[34] A. Balagopalan, A. Z. Jacobs, and A. J. Biega, "The Role of Relevance in Fair Ranking," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, (New York, NY, USA), p. 2650–2660, Association for Computing Machinery, 2023.

[35] K. L. Truong, A. Zimmermann, and H. Heidari, "Toward Valid Measurement of (Un)fairness for Generative AI: A Proposal for Systematization Through the Lens of Fair Equality of Chances," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 8, p. 2535–2549, Oct. 2025.

[36] H. Wallach, M. Desai, N. Pangakis, A. F. Cooper, A. Wang, S. Barocas, A. Chouldechova, C. Atalla, S. L. Blodgett, E. Corvi, P. A. Dow, J. Garcia-Gathright, A. Olteanu, S. Reed, E. Sheng, D. Vann, J. W. Vaughan, M. Vogel, H. Washington, and A. Z. Jacobs, "Evaluating Generative AI Systems is a Social Science Measurement Challenge," 2024.

[37] A. Panickssery, S. R. Bowman, and S. Feng, "LLM evaluators recognize and favor their own generations," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[38] J. Xu, G. Li, and J. Y. Jiang, "AI Self-preferencing in Algorithmic Hiring: Empirical Evidence and Insights," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 8, p. 2757–2758, Oct. 2025.

[39] "Blk Men in Tech - About the Organization." `https://www.blkmenintech.com/about-us`, 2025.

[40] "SWE Awards Program." `https://swe.org/awards/swe-awards-program/`, 2025.

[41] P. Kirichenko, M. Ibrahim, K. Chaudhuri, and S. J. Bell, "AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions," 2025.

[42] B. Wen, B. Howe, and L. L. Wang, "Characterizing LLM abstention behavior in science QA with context perturbations," in *Findings of the Association for Computational Linguistics: EMNLP 2024* (Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, eds.), (Miami, Florida, USA), pp. 3437–3450, Association for Computational Linguistics, Nov. 2024.

[43] B. Wen, J. Yao, S. Feng, C. Xu, Y. Tsvetkov, B. Howe, and L. L. Wang, "Know your limits: A survey of abstention in large language models," *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 529–556, 2025.

[44] Anthropic, "Claude Opus 4 & Claude Sonnet 4 — System Card," system card, Anthropic, May 2025.

[45] DeepSeek-AI, "DeepSeek-V3.1 — Model Card." `https://huggingface.co/deepseek-ai/DeepSeek-V3.1`, Sept. 2025.

[46] Google, "Gemini 2.0 Flash — Model Card," model card, Google, Apr. 2025.

[47] Google, "Gemini 2.5 Pro — Model Card," model card, Google, May 2025.

[48] Google, "Gemma 3 (12B) — Model Card." `https://huggingface.co/google/gemma-3-12b-it`, Aug. 2025.

[49] OpenAI, "GPT-4o mini: advancing cost-efficient intelligence." `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, July 2024.

[50] OpenAI, "GPT-5 System Card," system card, OpenAI, Aug. 2025.

[51] M. AI, "Llama 3.1 8B — Model Card." `https://huggingface.co/meta-llama/Llama-3.1-8B`, July 2024.

[52] M. AI, "Llama 3.3 70B — Model Card." `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`, Dec. 2024.

[53] T. Zhang, P. Qin, Y. Deng, C. Huang, W. Lei, J. Liu, D. Jin, H. Liang, and T.-S. Chua, "CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), (Bangkok, Thailand), pp. 10746–10766, Association for Computational Linguistics, Aug. 2024.

[54] J. Leng, C. Huang, B. Zhu, and J. Huang, "Taming overconfidence in LLMs: Reward calibration in RLHF," in *The Thirteenth International Conference on Learning Representations*, 2025.

[55] A. F. Cooper, K. Lee, M. Z. Choksi, S. Barocas, C. De Sa, J. Grimmelmann, J. Kleinberg, S. Sen, and B. Zhang, "Arbitrariness and social prediction: the confounding role of variance in fair classification," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24, AAAI Press, 2024.

[56] J. Castleman, Z. Shen, B. Metevier, M. Springer, and A. Korolova, "Measuring validity in llm-based resume screening," 2026.

[57] Jamie Birt, "Hard skills vs. soft skills: What's the difference?." `https://www.indeed.com/career-advice/resumes-cover-letters/hard-skills-vs-soft-skills`, 2025.

[58] X. Bai, A. Wang, I. Sucholutsky, and T. L. Griffiths, "Explicitly unbiased large language models still form biased associations," *Proceedings of the National Academy of Sciences*, vol. 122, no. 8, p. e2416228122, 2025.

[59] A. Gray, A. A. Rahat, T. Crick, S. Lindsay, and D. Wallace, "Using Elo rating as a metric for comparative judgement in educational assessment," in *Proceedings of the 6th International Conference on Education and Multimedia Technology*, pp. 272–278, 2022.

[60] J.-F. Laslier, *Tournament solutions and majority voting*, vol. 7. Springer, 1997.

[61] Y. Oren, N. Meister, N. S. Chatterji, F. Ladhak, and T. Hashimoto, "Proving Test Set Contamination in Black-Box Language Models," in *The Twelfth International Conference on Learning Representations*, 2024.

[62] H. Zhu, P. A. Hall, and J. H. May, "Software unit test coverage and adequacy," *Acm computing surveys (csur)*, vol. 29, no. 4, pp. 366–427, 1997.

[63] N. Jo and A. Wilson, "What Does Your Benchmark Really Measure? A Framework for Robust Inference of AI Capabilities." `https://arxiv.org/abs/2509.19590`, 2025.