

Through the Looking Glassdoor: AI-Generated Employer Portrayals in Open-Weight Language Models

SCOTT SEIDENBERGER, Cornell University, USA

TAYLOR WISE, University of Oklahoma, USA

ANINDYA MAITI, University of Oklahoma, USA

CHRISTOPHER COLLINS, Cornell University, USA

Job seekers increasingly encounter employer information through large language models (LLMs), yet little is known about the portrayals these systems generate. We audit AI-generated employer portrayals using 915,000 controlled prompt-response observations spanning 100 U.S. companies and 61 open-weight models (0.5B–120B parameters), benchmarked against 1.2 million Glassdoor reviews. We find a pattern we call *calibration with compression*. On average, the models moderately agree with Glassdoor about which companies are better places to work, but they compress the differences between companies. The LLMs’ portrayals pull poorly rated employers toward the middle. When portrayals are pooled across models, the spread between companies is only 62% of that in Glassdoor reviews, compared with 85% for the typical individual model. The models also rate career opportunities 0.22 stars higher than the real reviews from 2020 to 2026. Larger models give more stable answers and match real employee reviews more closely (correlation rises from .20 for models of 3B parameters or fewer to .47 for models between 20B and 70B), but large models also tend to agree with one another ($r = .80$ above 20B). This means that consulting another large model often provides less independent evidence than job seekers may assume. How a model is prompted also matters: asking the model to act as a “neutral” assistant instead of a “helpful” one lowers ratings of companies. Taken together, these results suggest that while AI tools can help job seekers get oriented to the information environment around employers (such as by functioning as a type of search engine), they do not function as a truly independent source of information or as a purely faithful proxy for widely accepted social information sources (like Glassdoor), because they can sound well calibrated while hiding differences that may matter most.

Additional Key Words and Phrases: large language models, employer image, employer branding, recruitment information sources, algorithm audit

1 Introduction

A brand is a promise, and employer brands carry real consequences for people making major life decisions about where to work. For decades, organizations sought to manage that promise through tightly controlled messaging in recruiting and corporate communications. The rise of social platforms and review sites rebalanced that control by elevating the voices of employees and job seekers. A new evolution is underway, one where large language models (LLMs) increasingly serve as cognitive intermediaries between job seekers and a vast, fragmented information environment. When a candidate asks “Is company X a good place to work?” an LLM returns a single, often confident answer. This harmonization of multiple sources into one narrative may be convenient, but it also raises foundational questions about accuracy, bias, and whose interests are reflected in responses. Users of LLM-powered apps report encountering factual errors and fabricated information in these tools (Massenon et al., 2025).

Now that AI conversational systems are acting as intermediaries that synthesize and serve decision support directly to end users, a natural new question arises. When job seekers ask LLMs about companies, in the context of a job search, what kind of answer do they get? Job seekers are increasingly using generative AI tools to make all sorts of life

Authors’ Contact Information: Scott Seidenberger, Cornell University, Ithaca, New York, USA, scotty@cornell.edu; Taylor Wise, University of Oklahoma, Norman, Oklahoma, USA, taylor.j.wise-1@ou.edu; Anindya Maiti, University of Oklahoma, Norman, Oklahoma, USA, am@ou.edu; Christopher Collins, Cornell University, Ithaca, New York, USA, cjc53@cornell.edu.

decisions. Yet, the answers these systems provide may differ systematically from traditional sources of information due to heterogeneity in training data and methods, potentially altering the employer information job seekers receive before any perception forms.

Existing evidence establishes this use as a practical motivation for the audit, rather than as an outcome the experiment is designed to estimate. In a nationally representative panel, 34% of U.S. adults (and 58% of those under 30) had used ChatGPT (Sidoti and McClain, 2025). Among recently hired job seekers, 53% used generative AI in their job search, and among those who used generative AI in their job search at all, 19% used it for researching career options and 16% for researching pay (ZipRecruiter, 2024). Among workers who had personal conversations with ChatGPT, 38% reported using it for “tough decisions” and 29% for career guidance. In the same survey, 49% of Gen Z respondents said they definitely or probably rely more on ChatGPT than on their boss (Resume.org, 2025). We also examined 89,678 English conversations sampled from the consented, deidentified WildChat corpus (Allen Institute for AI, 2025; Zhao et al., 2024). Transparent keyword indicators classified 1.3% as job-search related and 0.3% as specifically evaluating an employer as a place to work. Because these indicators lack human-coded precision and recall validation, they establish that such use occurs but are not precise population-prevalence estimates (Appendix B.7).

This paper reports an audit of **AI-generated employer portrayals**, which are the quantitative evaluative judgments (binary and numeric ratings) and descriptive language that LLMs produce about specific employers in response to job-seeker-style queries. We characterize portrayals in terms of established employer brand attributes, namely instrumental attributes (e.g., pay, benefits, work conditions) and symbolic associations such as prestige, sincerity, and innovativeness, following Lievens and Highhouse (2003), but we study the portrayals themselves, not job seekers’ perceptions of them. Conceptually, the portrayal occupies the middle of a chain where the employer information environment (corporate signaling, press coverage, employee reviews) feeds model training, alignment, and prompting. In turn, these produce the AI-generated portrayal of an employer that a job seeker encounters. That portrayal, then filtered through the job seeker’s trust and information processing, can shape employer image and application intentions (Cable and Turban, 2001; Collins and Stevens, 2002). Our study tests the middle link: how portrayals vary with model, scale, and framing, and how they correspond to the aggregate human evaluations they appear to summarize. We treat the downstream links as established mechanisms of applicant information processing to be tested in future work. Unlike social media, which aggregates diverse voices (albeit mediated through personalized algorithmic feeds), LLMs generate synthetic consensus as they compress underlying signals into a single response whose tone and content can vary systematically by the model, the prompt, and the context.

Three sources of heterogeneity motivate our analysis. First, model families differ in their training data, technical architecture, and alignment objectives. Second, prompt topic and system-role framing matter. The free-text prompts vary jointly in topic and wording, so those contrasts are descriptive, whereas the provider-set system role is experimentally manipulated. System-level role prompts can shift emphasis between symbolic and instrumental language, and these prompts are typically opaque to users, reflecting the design choices and business goals of the company providing access to the particular LLM. Third, model size is systematically associated with LLM output properties, and therefore with what job seekers are told.

Scaling-law studies motivate testing whether model size is associated with how consistently models integrate the available signal under standardized sampling perturbation (Kaplan et al., 2020; Muennighoff et al., 2023). Some products restrict access to more capable models by subscription, but consumer access tier does not map cleanly onto parameter count. If access channels disproportionately route some users to smaller models, the size-associated

differences documented here could create access stratification. This study tests those differences, not their distribution across commercial products.

Motivated by the usage evidence above, we organize the audit around three research questions.

RQ1 (Model Differences). Do portrayals differ systematically across model families and model size, net of other factors?

RQ2 (System-Role Framing). How does experimentally imposed system-role prompting change the valence of employer portrayals?

RQ3 (Correspondence and Departure). How closely do AI-generated employer portrayals correspond with aggregate human evaluations on a large review platform (Glassdoor), and in what ways do they systematically depart from them?

RQ1 maps onto hypotheses H1a and H1b, RQ2 onto H2, and RQ3 onto H3a and H3b. Prompt-topic contrasts in the free-text analysis are descriptive because topic and wording vary together.

We address these questions with a large-scale controlled experiment that crosses 61 open-weight models (spanning 22 families and 0.5B–120B parameters), system role, sampling temperature, and prompt topic for all Fortune 100 companies, producing 915,000 prompt-response observations. We standardize prompts, record model metadata, and manipulate inference temperature through a prespecified five-level sweep. We then conduct two analyses on these data. The first examines model, size, and system-role differences in binary evaluations, numeric ratings, and open-ended language responses. The second adds an external dataset from Glassdoor to compare model ratings with aggregate human responses.

Previewing the results, we show clear between-model differences, but not where one might first look. Model size is only weakly associated with the *level* of evaluations. It is much more clearly associated with temperature-sweep robustness. Smaller models give evaluations that swing under standardized sampling perturbation, while larger models answer more consistently. Model families show modest, measure-dependent differences in the language they use to describe the same employers (which families appear more symbolic or instrumental depends on the dictionary), and language differs much more across prompt topics. System-role prompting shifts numeric evaluations, while prompt-topic variation is used descriptively for text-side analyses rather than as an estimable numeric effect in the canonical ordinal data. Furthermore, benchmarked against more than a million Glassdoor reviews, LLM evaluations show moderate rank-order correspondence but compress the range (pulling lower-rated firms up toward the middle), and the apparent size of any mean difference depends strongly on the vintage of the human benchmark. Two further patterns sharpen the picture. Models, especially the largest, converge on a single shared answer rather than offering independent readings, and the positivity that benchmark vintage explains away elsewhere persists on the career-opportunities dimension. Our results hold broadly across the diverse range of companies included in the study, although the direction of the consensus–benchmark difference is sector-patterned. Taken together, the results indicate that what a job seeker gets from an LLM depends on which model they use and how the system prompt frames the exchange. Within this open-weight sample, larger models are more robust across the temperature sweep and more closely aligned with human consensus.

Our contributions are fourfold. Conceptually, we define AI-generated employer portrayals, locate them in the chain above, and extend theories of signaling and social information processing to an AI-mediated recruitment information environment in which a synthesizing intermediary, rather than identifiable senders or third parties, delivers the signal.

Empirically, we provide large-scale evidence on model-family, model-size, prompt-topic, and system-role differences in employer portrayals and quantify their correspondence with, and departures from, an existing social baseline. Practically, we surface immediate guidance for job seekers. Portrayals compress differences between employers, show

benchmark-vintage sensitivity, and keep career-opportunities ratings elevated relative to the 2020–2026 benchmark. The experimental system-role change also shifts evaluations. Methodologically, we contribute an inference-aware pipeline for auditing model-generated portrayals at scale.

Ultimately, this paper frames a timely question (what do job seekers encounter when they ask an LLM about an employer?) and answers it with empirical evidence. LLMs are powerful tools that can offload significant cognitive load from their users, but their outputs can systematically transform the employer information environment. The remainder of this paper develops the theoretical basis and hypotheses, details the experimental and analytical methods, reports results, and discusses implications for both job seekers and future research.

2 Theoretical Background

2.1 Employer Brand Equity and AI-Mediated Employer Information

Employer brand equity refers to the value attached to a company as a desirable employer, analogous to, but distinct from, consumer brand equity in marketing (Cable and Turban, 2001; Theurer et al., 2018). Early definitions emphasized that an employer’s brand encompasses a mix of functional, economic, and psychological benefits (Ambler and Barrow, 1996). This early understanding has been refined into an established framework that identifies the instrumental and symbolic employer brand attributes that job seekers evaluate (Lievens and Highhouse, 2003; Lievens et al., 2007). Job seekers evaluate employers based on objective attributes such as pay and benefits as well as subjective, identity-based associations such as culture and organizational values. Instrumental aspects appeal to utilitarian needs, while symbolic facets fulfill self-expressive and emotional needs by allowing individuals to identify with the company’s image. A strong employer brand thus harmonizes tangible job features with a compelling narrative about “who” the organization is, creating a promise of a valued work experience (Love and Singh, 2011; Theurer et al., 2018). Job seekers must then be made aware of these brands so they can identify a company as a potential employer, and then associate those brands with their attributes based on job seekers’ formed perceptions (Collins and Stevens, 2002). The pathways through which job seekers become aware of employer brands and form associations with them have shifted over time, reflecting changes in who controls the narrative and how brand signals are communicated.

Corporate-centered era. In the years before social media, employer brand communications were largely controlled by employers. Organizations strategically projected their desired image through formal channels such as recruitment ads, brochures, career websites, and targeted publicity (Cober et al., 2004; Collins and Stevens, 2002). These efforts were essentially top-down signals broadcasting what the company offered to employees.

Independent information sources for job seekers were limited, aside from personal word-of-mouth, which was known to influence perceptions (Van Hoya and Lievens, 2009) but typically spread only through personal networks. Most candidates had to rely on the company’s own messaging or third-party accolades. Even ostensibly neutral awards like “Best Company to Work For” rankings were often leveraged (and sometimes shaped) by firms to reinforce a positive employer image (Lievens et al., 2023). In this era, the employer’s voice carried the most weight in defining its brand equity, as corporate narratives dominated.

Social media era. The rise of platforms like Glassdoor, LinkedIn, Facebook, and Twitter in the late 2000s shifted the balance of power. Employer branding entered a social era in which narratives became crowd-sourced and democratized (Van Hoya and Lievens, 2007). Current and former employees, as well as prospective candidates, could now publicly share uncensored reviews and discussions about their work experience. Websites dedicated to employer reviews emerged as influential hubs of information that exist outside the direct control of companies (Dineen et al., 2019). Consequently,

job seekers today rely heavily on these channels. In one survey, 86% of job seekers report looking at a company's reviews and ratings before deciding where to apply (Glassdoor, 2021). This abundance of peer-generated content means that an employer's brand equity is increasingly shaped by social information streams. Importantly, the social media era introduced new complexity, in that candidates must interpret a mix of positive and negative commentary of uncertain credibility. Studies have shown that job seekers grapple with the signal-to-noise challenge of deciphering the online review space, which can be overly optimistic (potential or real plants by the company) or harshly negative, yet these reviews influence perceptions (Höllig et al., 2025; Lievens and Slaughter, 2016). Social platforms have pluralized employer brand narratives, forcing companies to engage with an empowered audience that can amplify or undermine the official brand messages in real time.

AI era emerges. The mid-2020s have ushered in yet another shift with the advent of LLMs capable of producing human-like answers to practically any question. Job seekers are increasingly turning to AI assistants to research companies, and are given a single, synthesized narrative in response (Rumage, 2026). LLMs encode patterns from their training corpora and compress viewpoints into one coherent answer. In effect, the diffuse voices of the social era are being re-centralized through an opaque, algorithmic intermediary. The convenience of an all-in-one AI-generated summary is high, as it saves candidates from wading through pages of search results or myriad Glassdoor posts.

However, this also means that AI's portrayal of an employer can supply salient inputs to first impressions. The prevailing positive or negative themes are those captured in the AI model's training data, and certain important aspects may be underrepresented based on the size of the model and how it is trained (Bommasani et al., 2022; Narayanan Venkit et al., 2025). Thus, while LLMs harmonize perspectives by blending many inputs, they also become a new center of influence curating job seekers' associations with an employer brand.

Theoretical anchors. The evolution above can be understood through two complementary theoretical lenses. Signaling theory posits that in markets with information asymmetry, such as the job market, better-informed parties send signals to convey unobservable attributes (Connelly et al., 2011; Spence, 1973), and organizations likewise signal their attributes to prospective employees (Celani and Singh, 2011). Personnel selection itself has been analyzed as a signaling game in which applicants and employers exchange and interpret such signals (Bangerter et al., 2012). Employer brand communications serve as signals, or promises, of what a candidate can expect from the employment experience.

These signals are effective only if perceived as credible and consistent (Collins and Martinez-Moreno, 2022; Dineen et al., 2019). More generally, we argue that they are effective only insofar as receivers judge them diagnostic, that is, verifiable, consistent across encounters, and distinctive across employers (building on Cable and Turban, 2001; Evertz et al., 2021). At the same time, social information processing theory (SIP) holds that individuals rely on cues from their social environment when forming attitudes and decisions, especially under uncertainty (Salancik and Pfeffer, 1978). Rather than evaluate an employer in isolation, job seekers actively look to how others (friends, current employees, media, online communities) describe and react to that employer. They integrate those social cues into their own perception of the company. In practice, candidates build their perception of a company's brand equity by observing available narratives, or "social proof." These include, but are not limited to, in-person word-of-mouth, online ratings, social network feeds, and increasingly now AI-provided answers (Evertz et al., 2021; Panagiotidou and Mihail, 2024). Employer brand equity in the digital era is now fundamentally different because AI is serving as a new entrant that mediates how signals reach job seekers, and how those job seekers seek, filter, and process information.

Two scope notes discipline how we use these lenses. We draw on signaling theory for its receiver-side account of what makes a signal usable, namely credibility and consistency (Connelly et al., 2011), which we extend to distinctiveness, recency, and perceived consensus, rather than for its model of sender incentives. We also theorize the LLM not as

a new sender emitting signals of its own, but as a **channel** that transforms the properties of other senders' signals as it aggregates them. At the receiver-facing level, the conversational answer functions as what we call a fourth-person informant. Thus, *channel* denotes the production-side intermediary and *informant* denotes the interface the user encounters. Neither term makes the LLM a truly independent source of social evidence. Our thesis, in these terms, is that LLMs alter the recruitment information environment not by adding one more voice but by changing the properties receivers use to weight voices. These properties are the **source** of the signal (an aggregate rather than an identifiable sender), its **credibility cues** (no observable incentives or track record), its **consistency** (an evaluation that can change on re-asking), its **distinctiveness** (a summarizing channel compresses differences between employers), its **recency** (content may not reflect the current information environment), and the **perceived consensus** it conveys (one synthesized answer presented as if it were many). These six properties organize the hypotheses that follow. They map to informant identity (H1a), signal consistency (H1b), covert source control over framing (H2), and the correspondence and departure of the consensus the signal appears to carry (H3a, H3b).

2.2 Conversational AI as a Fourth-Person Informant

Foundation models learn reusable statistical representations from broad training corpora and adapt them across downstream tasks (Bommasani et al., 2022). At query time, conversational systems synthesize learned patterns (and, in retrieval-augmented settings, potentially external sources) into a single response. Prior work shows that LLMs can express diverse perspectives, but eliciting that diversity may require deliberate prompting (Hayati et al., 2024). A prompt to an LLM can therefore deliver a convenient summary that spares users the work of information synthesis. In doing so, AI systems have become sociotechnical intermediaries that influence how people access information. Where the social era of the web was characterized by many individual voices and platforms, the LLM acts as an opaque central node that filters and presents one output. Narayanan Venkit et al. (2025) find that answer engines narrow users' exploration of sources, trading autonomy and diversity of viewpoints for convenience.

Classic SIP research in organizational settings demonstrated that employees look to their colleagues, peers, and the broader social context to shape how they interpret their jobs and environment (Salancik and Pfeffer, 1978). Similarly, in the employer brand equity space, job seekers do not form their perceptions in a vacuum. They adapt based on the social cues available to them (Collins and Kanar, 2013).

A new class of informant has entered this environment, and there is direct evidence that users treat it socially. Surveys show that workers, especially in younger generations, routinely consult tools like ChatGPT for advice, ideas, and support. In a survey of more than 8,600 full-time U.S. workers, over one-third of those who have personal conversations with ChatGPT report having asked it for help with tough decisions (Resume.org, 2025). Many users describe these chatbots as a "coworker or assistant," some as a "companion" (Resume.org, 2025), and providers lean into this anthropomorphization to increase engagement (Cheng et al., 2022; Nikolov et al., 2026). Job seekers, then, are positioned to solicit employer opinions from conversational AI as they would from a social source, without the cues they would normally use to weigh one.

In SIP terms, conversational AI is best treated as a new class of informant rather than as a person. Classic treatments distinguish the job seeker's own experience, word-of-mouth from identifiable others, and third-party sources such as review platforms and rankings (Cable and Turban, 2001; Dineen et al., 2019). An LLM-based assistant fits none of these categories. It is not a participant in the employment relationship, not an acquaintance with observable credibility cues, and not an identifiable third party with inspectable incentives. It synthesizes the recorded utterances of many sources (employer messaging, employee reviews, journalism) into a single response whose provenance the user cannot audit,

filtered through training and alignment choices made by the model’s developer. We therefore label this receiver-facing interface a fourth-person informant, an aggregated, machine-generated voice that delivers socially sourced content in the conversational form of individual testimony. Four properties distinguish this informant class and motivate our hypotheses. Its content varies with the model consulted (H1a). Its consistency varies with technical attributes, such as scale, that users cannot observe (H1b). Its stance can be tilted by provider-set framing the user never sees (H2). And its correspondence with the human consensus it appears to summarize is an empirical question, not a definitional property (H3a, H3b).

A job seeker’s impression of a company may thus be informed not only by word-of-mouth or online reviews but also by an LLM’s answer to a prompt like “Is [Company] a good place to work?” That answer is produced by a centralized, opaque intermediary whose evaluations may amplify certain informational cues and mute others. Whether those evaluations are consistent, representative, and aligned with the human consensus they appear to summarize is precisely what an audit of the portrayals themselves can establish. We next derive hypotheses by combining recruitment theory, which supplies the constructs (source identity, credibility, consistency, distinctiveness) and the stakes, with prior work on generative model bias, scaling, and conversational framing, which supplies the generating mechanisms.

2.3 Hypotheses

We derive hypotheses by treating AI-generated portrayals as signals in the recruitment information environment, whose influence on receivers depends, in our synthesis of this literature (Cable and Turban, 2001; Celani and Singh, 2011; Dineen et al., 2019), on source identity, credibility, consistency, distinctiveness, and recency. Machine learning literature supplies the generating mechanisms. Training corpora and alignment objectives differentiate model families (source identity), scaling makes representations more stable (signal consistency), prompting alters context (framing), and alignment training rewards outputs raters judge helpful and harmless (Askell et al., 2021; Ouyang et al., 2022), and may encourage agreeable, sycophantic responses (Malmqvist, 2025); LLMs also display social-desirability bias when they infer they are being evaluated (Salecha et al., 2024). In the SIP account developed above, if the conversational assistant is a new informant class, then model family determines *which informant* is speaking, model scale is associated with how *consistent* that informant is, framing determines the *context cues* surrounding the exchange, and correspondence with human reviews determines whether the informant’s view tracks the social consensus it appears to summarize.

2.3.1 H1a. Employer portrayals differ across model families. Different model families are trained on different corpora with different fine-tuning objectives (Wang et al., 2025), so each effectively encodes a different “worldview.” In SIP terms, each family is a different informant. Just as different human informants emphasize different aspects of an employer when asked the same question (Salancik and Pfeffer, 1978; Van Hove and Lievens, 2009), we expect families to differ systematically in the content and rhetoric of their employer descriptions, particularly in the balance of symbolic (identity, culture, belonging) versus instrumental (pay, work, achievement) language that job seekers use to form brand associations (Lievens and Highhouse, 2003). If family signatures exist, then “what the AI thinks” of an employer is not a single fact about the employer but a property of the informant a job seeker happens to consult.

H1a. The symbolic-versus-instrumental framing of employer descriptions differs systematically across model families.

2.3.2 H1b. Larger models exhibit evaluations more robust to sampling perturbation. Scaling-law studies show that language-model loss improves predictably with model size and training compute, including in data-constrained regimes (Kaplan et al., 2020; Muennighoff et al., 2023). We test whether this general scaling relationship extends to employer-evaluation robustness under standardized sampling perturbation. For a job seeker, the practical risk of a small model is

not only a different answer but an unreliable one, an employer portrayal that swings with trivial changes in sampling. Signaling theory makes the same point from the receiver’s side. A signal influences decisions only insofar as it is credible and consistent (Celani and Singh, 2011; Collins and Martinez-Moreno, 2022), and an informant whose evaluation changes on re-asking transmits little usable information. We therefore expect evaluation **robustness to standardized sampling perturbation** to increase with model size. Scale may also shift the *level* of evaluations, though theory is less decisive about direction. We treat the level effect as secondary to the robustness effect.

H1b. Larger models exhibit employer evaluations that are more robust to standardized sampling perturbation than those of smaller models (with a secondary, non-linear association between size and evaluation level).

2.3.3 H2. Conversational framing shifts evaluations. Minor changes in prompts can substantially change an LLM’s output (Liu et al., 2023; Salinas and Morstatter, 2024), and persona assignment induces systematic shifts in model reasoning (Gupta et al., 2024), echoing classic framing effects in human judgment (Tversky and Kahneman, 1981). In our setting, two framing levers are distinct. The first is the system role (a provider-set persona the user never sees), and the second is the user’s own wording. The system role is the invisible context cue set by the platform rather than the job seeker, and that may tilt the informant’s stance before the conversation begins.

H2. System-role framing shifts the valence of numeric employer evaluations.

We additionally examine whether the user’s own wording shifts the symbolic/instrumental register of free-text answers. In our prompt battery, the valence of user wording is intentionally varied only within the free-text items and is therefore confounded with prompt topic, so we treat user-side framing as a descriptive language-side analysis (reported under H1a) rather than a formal hypothesis test.

2.3.4 H3a and H3b. Correspondence and departure of LLM evaluations relative to human reviews. How should AI-generated portrayals relate to the aggregated human evaluations they appear to summarize? If models absorb the public sentiment record (including the employer reviews, press coverage, and rankings that circulate online), their company-level evaluations should track aggregated human reviews of the same employers, in the spirit of electronic word-of-mouth aggregation (Dineen et al., 2019; Evertz et al., 2021). This correspondence is what makes an LLM a *usable* informant at all.

H3a (Correspondence). Company-level LLM employer ratings are positively associated in rank order with aggregated human reviews of the same employers. We separately estimate the mean calibration gap and range compression. Neither is treated as an equivalence test.

Next, in what ways do portrayals systematically depart from the human benchmark? The summarizing character of LLMs predicts systematic departure. A channel that compresses many voices into one answer should mute differences between employers (range compression, reducing signal distinctiveness). Content that is not continuously updated may correspond differently with review benchmarks from different periods. We call this benchmark-vintage sensitivity. By itself, it does not identify temporal lag or training-data provenance. Uneven coverage and valence across employer brand content dimensions should displace some dimensions more than others (dimension-specific displacement). A further question is whether these departures take the specific form of mean positivity. Alignment training rewards outputs that raters judge helpful and harmless (Askell et al., 2021; Ouyang et al., 2022) and may encourage agreeable, sycophantic responses (Malmqvist, 2025). LLMs also display social-desirability bias when they infer they are being evaluated (Salecha et al., 2024). In SIP terms, this is a systematically rosy informant whose social proof overstates the consensus it summarizes, precisely the property a decision maker most needs to know about a source they cannot cross-examine.

H3b (Departure). Relative to aggregated human reviews, LLM employer ratings exhibit systematic departure, including reduced range, benchmark-vintage sensitivity, and dimension-specific displacement. We additionally examine, without a directional hypothesis, whether the mean calibration gap differs from zero.

3 Methods

This study combines (1) a large controlled knowledge-extraction experiment and (2) a benchmark against human reviews from Glassdoor. We foreground *what we measure* and how each measure maps to the hypotheses. The full collection protocol (design, models, prompts, effort) is in Appendix B.

3.1 Design

We use a complete factorial controlled experiment crossing **100 companies** $\times$ **61 open-weight models** $\times$ **2 system roles** $\times$ **5 temperatures** $\times$ **15 prompts**, yielding 915,000 typed responses across three modalities (binary, ordinal, free text). Open-weight models served locally give full control over the system prompt and sampling temperature and support a controlled, consistently documented design (Appendix B.3). Responses were elicited in a single collection window using the standardized model tags, prompts, system roles, and temperature schedule documented in Appendix B. Companies are the Fortune 100 (Appendix B.2). Prompts operationalize employer brand equity (Appendix B.4).

3.2 Measures

Symbolic vs. instrumental employer-image language (H1a). Employer brand equity comprises instrumental attributes (pay, benefits, work conditions) and symbolic associations (e.g., prestige, sincerity, innovativeness) (Lievens and Highhouse, 2003). We score each free-text response with two closed-vocabulary measures. The first is the expert-validated employer-image dictionary of Theurer et al. (2022), a 557-word lexicon distinguishing six symbolic dimensions (sincerity, innovativeness, competence, prestige, ruggedness, cheerfulness) from four instrumental ones (pay, benefits, advancement, teamwork). The second is a broader register check with a theory-driven Linguistic Inquiry and Word Count (LIWC-22) category map (Boyd et al., 2022). Its symbolic side comprises affect, social, affiliation, moral, insight, certitude, we, they, family, friend, polite, and prosocial. Its instrumental side comprises work, achieve, money, power, reward, time, cause, number, and quantity. Under each, we summarize the balance as a smoothed log-ratio of symbolic to instrumental content, winsorized at the 0.1% tails. Higher values indicate more identity/affect-laden language and lower values more utilitarian language. We model this balance with a linear mixed model using crossed random intercepts for company and model and fixed effects for model family, question type, system role, and temperature ($N = 304,998$, Table 1 and Figure 1). The complete free-text design contains 305,000 responses, of which two, containing no analyzable lexical content, were absent from the LIWC export. Treating both measures as indicators of linguistic register rather than direct measures of perception, we report their convergence following computer-aided text analysis validation logic (McKenny et al., 2018; Short et al., 2010). This mirrors prior work linking employer-review content to symbolic/instrumental framing (Höllig et al., 2024).

Binary and ordinal evaluations (H1b, H2, H3a, H3b). Binary prompts force a Yes/No attribute judgment. Ordinal prompts force a 1–10 rating. These are our measures of *how favorably* a model evaluates an employer and map directly onto the size, framing, and correspondence tests. Alongside the free-text questions, each model answered a battery of numeric questions in two formats. The first format is five binary Yes/No items (e.g., “Does [COMPANY] have a strong and positive workplace culture?”). The second is five ordinal 1–10 ratings (e.g., “Rate [COMPANY]’s compensation and benefits package...”). These are reported as the “Yes-rate” and the mean ordinal rating. Because these prompts are

Table 1. H1a: Linear mixed model of symbolic-versus-instrumental language (LIWC log-ratio)

Fixed effect or test	b or test statistic	SE	df	p
Intercept	-0.638	0.039	47.4	< .001
Question: culture/work environment	+0.685	0.002	304,832	< .001
Question: praise vs. criticize	+0.071	0.002	304,832	< .001
Question: attractiveness factors	-0.082	0.002	304,832	< .001
Question: growth & development	-0.111	0.002	304,832	< .001
System role: neutral	-0.025	0.001	304,832	< .001
Temperature	-0.013	0.002	304,832	< .001
LIWC family omnibus, model-level df	F(21, 39) = 1.68			.080
Theurer family omnibus, model-level df	F(21, 39) = 1.89			.042
LIWC family EMM range (log points)	0.37			

Note. N = 304,998 free-text responses, with crossed random intercepts for company (100) and model (61) and the reputation prompt as the reference question. Coefficient rows and the estimated marginal mean (EMM) range use the LIWC register measure. The separate Theurer employer-image model is included to show the two model-level omnibus tests together. Full coefficients are reported in Appendix Table A5 and asymptotic approximations in Appendix Table A1.

uniformly neutral in their user wording, the only framing factor that varies across them is the system role, the hidden persona instruction a job seeker does not see or control directly. We tested two personas by changing a single word in “You are a [helpful/neutral] job finder assistant.”

Model family and model size (H1a, H1b). We parse the model family (e.g., llama3) and tag-reported total parameter count in billions from each model identifier. Size is our measure of model capability/scale, though for mixture-of-experts models, total and active parameter counts may differ.

System role (H2) and prompt-topic variation (descriptive). The *system role* is the system-level persona (provider-set in commercial products, experimenter-set here) manipulated across all prompts. The two conditions are, verbatim, “You are a helpful job finder assistant.” and “You are a neutral job finder assistant.” (Appendix B.5). *User-prompt valence* (positive/neutral/skeptical) varies across the free-text prompts only. The numeric prompts are uniformly neutral by design. We therefore estimate the system-role effect on numeric outcomes and examine prompt-topic and wording differences in language descriptively.

Temperature. Sampling temperature follows a **prespecified five-level sweep** (0, 0.3, 0.5, 0.7, 1.0). We enter it as a numeric covariate and use the sweep as a standardized perturbation for measuring rating robustness (the within-condition dispersion across the five draws) because each condition is drawn once at each temperature. The resulting dispersion is not test-retest reliability at a fixed setting, and a user querying at temperature 0 would observe less variation than our dispersion measure implies.

Human benchmark (H3a, H3b). We benchmark LLM ratings against company-level Glassdoor overall (“place to work”) ratings as external human signal. The review corpus was collected from Glassdoor employer pages via a commercial service in two phases. The first was a July-2024 census of the reviews then available for 88 study firms. The second was a June-2026 follow-up collection of roughly 1,000 recent English reviews for each of the remaining 12 firms, mostly newer Fortune 100 entrants. The combined, deduplicated benchmark contains **1,213,399 unique reviews** across all 100 study companies, with review dates spanning 2008 through June 2026 (62% dated 2020 or later). Sub-dimension ratings stored as zero (the reviewer left the dimension blank) are treated as missing. Because review norms drift over time, we also report a benchmark-vintage sensitivity analysis that recomputes the comparison within review-date windows.

The LLM-side counterpart is the mean of the innovation and employer-reputation ordinal prompts, the two items closest in scope to an overall employer evaluation (the full prompt-to-dimension mapping is given in Appendix B.4). Neither item directly asks for an overall “place to work” rating, so we refer to their mean as the *overall-rating proxy* and interpret its Glassdoor correspondence as construct correspondence rather than validation of an identical measure. The reputation item explicitly invokes public perception and employer rankings, which may mechanically increase its correspondence with a public-review benchmark. We therefore report the two component items separately as a sensitivity check. The remaining ordinal prompts map directly to Glassdoor’s culture, compensation, and career-opportunities sub-dimensions and are benchmarked at the dimension level (Section 4.5). An LLM rating x on the 1–10 scale is mapped to Glassdoor’s 1–5 scale using endpoint anchoring, $x' = 1 + 4(x - 1)/9$, so 1 maps to 1 and 10 maps to 5. Spearman correlations and rank order are invariant to any monotone mapping, and Pearson correlations are invariant to this positive affine mapping. Mean gaps, standard deviation (SD) ratios, and rounded-star comparisons use the endpoint-anchored scale and therefore depend on this mapping. Where an analysis requires per-model company-level means, models lacking sufficient valid responses for the relevant items are excluded (in practice only gemma:2b, with 42.4% invalid responses, per Appendix B.6), so the overall-rating benchmark, per-model alignment, and cross-model convergence analyses rest on 60 models.

3.3 Analysis

Although the corpus contains 915,000 responses, those responses are generated by 61 models. Responses from the same model are correlated draws from one “informant,” so any effect of a model-level attribute (family, parameter count) must be tested against model-to-model variation, not response-to-response. We therefore pair every row-level estimation model with inference that respects the LLM model as the clustering unit, and we accompany each test with a *design-based* check computed at the model level (one observation per model, $n = 61$, or 60 for rating-based checks).

H1a. Linear mixed model of the log-ratio of symbolic to instrumental language on model family, question type, system role, and temperature with crossed random intercepts for company and model. Family contrasts use Bonferroni-adjusted estimated marginal means. The omnibus family test is additionally run at the model level (one mean log-ratio per model, one-way ANOVA across the 22 families), so its degrees of freedom reflect models, not responses.

H1b. Natural cubic spline ($df = 4$) in parameter count for the binary outcome (logistic regression) and the ordinal outcome (cumulative-link model), with controls for company, family, role, and temperature. Binary-model inference uses cluster-robust standard errors with the model as the clustering unit, including a cluster-robust joint Wald test of the spline. *Temperature-sweep robustness* is within-condition rating dispersion across the temperature sweep, summarized by size and re-tested at the model level. Robustness checks include a spline degrees-of-freedom sensitivity sweep, threshold-wise proportional-odds checks (with cluster-robust spline tests at each cut point), and a complete-condition robustness re-analysis that screens out conditions with missing temperature draws.

H2. The system-role effect is read from the binary and ordinal models with cluster-robust inference, and from a design-based paired test of per-model role differences.

H3a (Correspondence). A review-count-weighted regression of Glassdoor on LLM company means, Pearson and Spearman correlations, a mixed model of company $\times$ model LLM means on the Glassdoor mean with crossed company and model random intercepts, and a per-model alignment analysis that correlates each model’s company ratings with Glassdoor’s and relates that alignment to model size (with a leave-one-family-out sensitivity check). We report the company-level mean gap and its two-sided 95% confidence interval (CI) as a calibration estimate (the company, $n = 100$, is the unit of analysis), not as an equivalence test.

H3b (Departure). We report the ratio of company-level standard deviations as a direct measure of range compression, decomposed into individual-model and consensus-averaging components. We add a benchmark-vintage sensitivity analysis that recomputes the company-level comparison within review-date windows (2008–2015, 2016–2019, 2020–2026), with a yearly decomposition, a balanced-panel check, and an ordering-recency regression. We also report a dimension-level benchmark of the culture, compensation, and career-opportunities items. The mean calibration gap is examined with a two-sided test and confidence interval rather than a directional test.

Tests are organized by the prespecified hypotheses and accompanied by effect sizes and design-based checks. Pairwise family contrasts use a Bonferroni adjustment. Other reported p-values are not multiplicity-adjusted and should be interpreted with the pattern of estimates rather than alone.

Predicted values from the binary and ordinal models are reported as average marginal predictions (averaged over the observed covariate distribution) rather than predictions for a single reference cell. The binary curve (Appendix Figure A2a) includes delta-method 95% confidence intervals based on the model-clustered covariance. The ordinal curve (Appendix Figure A2b) is descriptive and is shown without an uncertainty interval. Figure 2 uses 20,000 within-bucket bootstrap resamples of models. Panel A recomputes the condition-weighted mean in each resample using each selected model’s number of estimable temperature-sweep cells. Panel B uses the unweighted mean across models, and no interval is computed for the one-model >70B bucket.

4 Results

We analyze 915,000 prompt-response observations from a controlled design that crosses 100 large employers, 61 open-weight models (22 families, 0.5B–120B parameters), two system roles, five sampling temperatures, and 15 prompts across three response modalities (binary judgments, 1–10 ratings, and free text). We report each hypothesis in turn, pairing summary tables with the corresponding figures. Throughout, row-level models provide adjusted estimates, while inference for model-level attributes (family, size) rests on model-clustered standard errors and design-based checks computed at the model level. Full coefficient tables are provided in Appendix Tables A5–A8.

4.1 Prompt-Topic Differences Exceed Model-Family Differences in Employer-Description Language

Evidence for model-family language signatures is modest and measure-dependent once we account for the 61 models being part of 22 model families. Each model answered five free-text questions about each company, namely its reputation as a place to work, its culture and work environment, how it supports employees’ growth and development, the factors that make it an attractive employer, and what employees praise versus criticize. Language differs far more across prompt topics than across model identities. Relative to the reputation prompt, the culture/work-environment prompt is markedly more symbolic ($b = +0.685$, $p < .001$). The contrasts are estimated within models, but remain descriptive because topic and wording vary together. The mixed model accounts for model and company intercepts, although it does not add a company-by-model random effect. Reporting the symbolic and instrumental indices separately shows that the topic effect moves both sides of the ratio. The culture prompt elicits the highest symbolic index of the five prompts (3.77, versus 2.41–3.31) and the lowest instrumental index (3.27, versus 4.14–5.88). A dimension-level decomposition shows the culture prompt elicits markedly more affiliation (+4.8 percentage points [pp]), social (+4.4 pp), and positive-tone (+2.1 pp) language, and less achievement (–2.3 pp), power (–2.9 pp), and work (–3.4 pp) language than the other prompts. That is, describing “culture” pulls models toward identity/belonging language and away from instrumental, task-oriented terms. The system role and temperature also shift the language slightly (neutral role $b = -0.025$, temperature $b = -0.013$, both $p < .001$, and the role effect replicates in a design-based check as a per-model paired difference of -0.025 , $t(60) =$

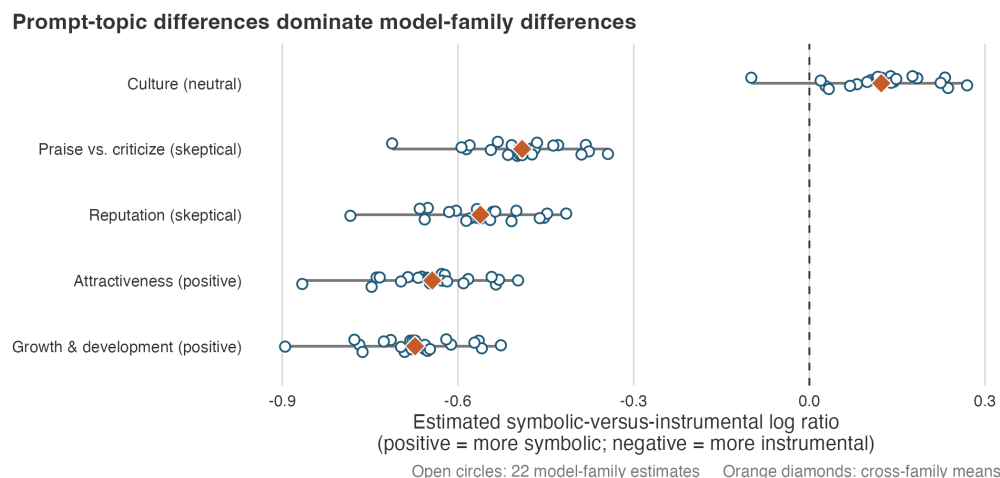


Fig. 1. Prompt-topic and model-family variation in symbolic-versus-instrumental language (LIWC register measure). Open circles show family-specific estimated marginal means for 22 model families. Diamonds show cross-family means and horizontal lines show the family range. Positive values indicate more symbolic language and negative values more instrumental language. Because prompt topic and wording vary together, the between-topic contrasts are descriptive. Prompt-topic differences are much larger than within-topic family differences.

-7.24, $p < .001$). A heatmap of the 30 companies with the greatest dispersion (Appendix Figure A1) illustrates how these signatures vary across model families.

The two excerpts below answer the same question (“How would you describe [COMPANY]’s work environment and company culture?”), yet illustrate the opposite sides of the scale. One model answers in the language of belonging and value, and the other with more of a fact sheet. A response at the symbolic extreme (log-ratio +1.56) reads, “Coca-Cola fosters a collaborative, inclusive environment that values innovation, teamwork, and a global perspective. Employees enjoy a supportive culture emphasizing sustainability, community engagement, and personal growth...” A response at the instrumental extreme (-2.23) reads, “Sysco is one of the largest food service distribution companies in North America, providing more than 1 million individual products to over 500,000 restaurant and foodservice customers. With a sales force of approximately 7,000 drivers...” These excerpts show belonging-and-identity language versus facts-and-figures language about the same kind of question. They are unverified model outputs reproduced to illustrate linguistic register. We do not treat their factual claims as established.

The employer-image dictionary developed by Theurer et al. (2022) supports the claim at the model level that model families are associated with differences in the language of the portrayal. The family omnibus is $F(21, 39) = 1.89$, $p = .042$ (asymptotic $\chi^2(21) = 39.6$, $p = .008$, Appendix Table A1).

H1a receives qualified support. Under model-level inference the family effect is significant under the expert-validated employer-image dictionary ($F(21, 39) = 1.89$, $p = .042$) and marginal under the broader LIWC register map ($F(21, 39) = 1.68$, $p = .080$), and the two measures order families differently (family-level $r = -.20$), so which family leans “symbolic” depends on the operationalization, while prompt topic is associated with much larger language differences than model identity under either measure (Figure 1 shows the LIWC register measure). Answers to different prompt topics use sharply different rhetoric, whereas model identity leaves a more modest signature even holding the company constant.

Table 2. H1b/H2: Size, system role, and temperature effects on numeric evaluations

Metric	Size spline (joint, 4 df)	System role: neutral	Temperature
Binary b (OR)	—	-0.33 (0.72)	-0.10 (0.91)
Binary cluster-robust test	$\chi^2(4) = 10.1, p = .039$	$z = -7.60, p < .001$	$z = -1.44, p = .15$
Ordinal b (OR)	—	-0.34 (0.71)	-0.11 (0.90)
Design check — Yes-rate	$F(4, 56) = 1.11, p = .36$	$t(60) = -7.53, p < .001$	$t(60) = -1.53, p = .13$
Design check — rating	$F(4, 55) = 1.35, p = .26$	$t(59) = -4.70, p < .001$	$t(59) = -2.03, p = .047$

Note. The binary model is a logistic regression ($N = 302,933$) and the ordinal model a cumulative-link model ($N = 296,577$). Both control for company and model family. Cluster-robust tests treat the model (61 clusters) as the unit of clustering. Design-based checks are computed from one observation per model (paired role differences and spline F-tests on per-model means). These checks use $n = 61$, except the rating-based checks, which rest on 60 models because gemma:2b yields too few valid ratings. Full coefficient tables, including naive standard errors, are in Appendix Tables A6 and A7. For the ordinal model, the size spline is summarized only by a naive in-sample likelihood-ratio test ($LR \chi^2(4) = 5,083$). This statistic is not cluster-robust and, unlike the design-based checks, should not be read as an inferential test of size.

4.2 Model Size Is Associated With Temperature-Sweep Robustness More Than Rating Level

We examined two size associations in numeric judgments, one with the level of evaluations and one with their robustness across the temperature sweep. The level analysis uses natural cubic splines ($df = 4$) for parameter count with controls for company, model family, system role, and temperature (Table 2).

Level. For binary “Yes” judgments (logistic regression, $N = 302,933$), the average marginal predicted probability of “Yes” dips from about .80 at the smallest models to roughly .64 near 5B parameters before recovering to about .81 for the largest models (Appendix Figure A2a). A model-clustered joint Wald test is modest, $\chi^2(4) = 10.1, p = .039$, and the design-based check (one mean Yes-rate per model) does not reject a flat relationship, $F(4, 56) = 1.11, p = .36$. A degrees-of-freedom sensitivity sweep confirms the dip shape is not an artifact of the chosen smoothing (a linear fit is the worst specification by the Akaike information criterion). For 1–10 ratings (ordinal cumulative-link model, $N = 296,577$), adding the size spline improves in-sample fit ($LR \chi^2(4) = 5,083$), but the average marginal expected rating is nearly flat, from 6.7 at the smallest models to a peak of 7.3 near 18B and 7.2 at 120B (Appendix Figure A2b), and threshold-wise cluster-robust spline tests are weak ($p = .041, .12$, and $.30$ at the $\geq 6, \geq 7$, and ≥ 8 cut points). We conclude that the association between model size and evaluation level is weak. Parameter count is not associated with a meaningfully different portrayal level in repeated sampling.

Temperature-sweep robustness. The clearest size association is with the **consistency of portrayal**. Because every company $\times$ model $\times$ role $\times$ prompt condition is sampled once at each of five temperatures (0–1), we measure how much a model’s rating moves under standardized perturbation. Within-condition rating dispersion falls steeply with size (Table 3). Mean SD declines from 0.84 ($\leq 3B$) to 0.30 ($> 20\text{--}70B$), the share of perfectly stable conditions rises from 23% to about 50%, and size correlates negatively with dispersion overall (Spearman $\rho = -.28$, Figure 2A). Because invalid responses concentrate in the smallest models (Appendix B.6), we re-ran the analysis restricted to conditions with all five temperature draws (97.6% of conditions). The gradient is essentially unchanged (mean SD $0.80 \rightarrow 0.30, \rho = -.26$), so incomplete cells are unlikely to generate it. Nor does any single model family account for it. At the model level (one mean within-condition SD per model), size correlates with dispersion at $\rho = -.42$, and recomputing that correlation leaving out each of the 22 families in turn keeps it between $-.38$ and $-.49$. The decline reaches 0.30 in the $> 20\text{--}70B$ bucket and then plateaus at the same value for the sole $> 70B$ model. In short, larger models do not rate employers higher. Their ratings are more robust to sampling perturbation. Descriptive temperature patterns are reported in Appendix Table A2 and Appendix Figure A3.

Table 3. H1b: Rating robustness across the temperature sweep, by model size

Model size	Mean within-condition SD	Mean SD (complete conditions)	Perfectly stable (%)	Conditions	Models
≤3B	0.84	0.80	22.7	14,988	15
>3–8B	0.53	0.49	36.5	16,737	17
>8–20B	0.49	0.49	33.5	12,000	12
>20–70B	0.30	0.30	50.3	15,000	15
>70B	0.30	0.30	46.6	1,000	1

Note. Within-condition SD is the standard deviation of the 1–10 rating across the five sampling temperatures within each company × model × role × prompt cell. “Complete conditions” restricts to cells with all five temperature draws (97.6% of cells), guarding against differential attrition. Spearman correlation between model size and within-condition SD = $-.28$ ($-.26$ in complete conditions). These summaries contain 60 rather than 61 models because gemma:2b did not yield sufficient valid ratings for model-level estimates.

H1b is supported for its primary claim that portrayal robustness across the temperature sweep is associated with model size across specifications. The secondary size-level association is weak and not robust to model-level inference. Smaller models give noisier evaluations under the designed perturbation. Greater model size is associated with temperature-sweep robustness, not a fundamentally different portrayal level.

4.3 A One-Word System-Role Change Lowers Employer Ratings

System-role framing shifts numeric evaluations. Switching the system prompt from “helpful” to “neutral” lowers positivity across response types. In marginal-effect terms, per model the neutral role lowers the Yes-rate by 5.4 percentage points (design-based paired $t(60) = -7.53$, $p < .001$) and the mean 1–10 rating by 0.16 points ($t(59) = -4.70$, $p < .001$). The 0.16-point shift on the 1–10 scale corresponds to about 0.07 stars under the endpoint-anchored 1–5 mapping. This is a modest level shift, but one that moves the entire signature in the same direction from a single hidden word change. The regression estimates show that the odds of a “Yes” decrease (odds ratio [OR] = 0.72, model-clustered $z = -7.60$, $p < .001$) and the odds of a higher ordinal rating also decrease (OR = 0.71, Table 2), and the size curves retain their shape while shifting downward under the neutral role.

H2 is supported. The system-role manipulation reliably shifts the valence of numeric employer evaluations, with consistent effects at both the response level and the model level. Within the two roles tested, an imposed system prompt measurably tilts how favorable an LLM’s employer portrayal is.

4.4 Correspondence With Human Consensus

When the company-level overall-rating proxy is benchmarked against 1.2 million reviews through June 2026, the mean LLM–Glassdoor calibration gap is +0.06 on the 1–5 scale (95% CI [+0.01, +0.11]), with 66% of companies rated above their Glassdoor mean (exact sign test $p = .002$), as shown in Figure 3B and Table 4. The small interval indicates close mean calibration in practical terms, but it is not a prespecified equivalence test. Rank-order correspondence is assessed separately. These are company-level analyses ($n = 100$).

The two measures are clearly positively associated. A review-count-weighted regression of Glassdoor on LLM means yields a positive slope ($b = 0.95$, $p < .001$, Pearson $r = .52$, Spearman $\rho = .48$, Figure 3A, Appendix Table A3), and the association holds within models. Regressing company × model LLM means on the Glassdoor mean with crossed company and model random intercepts gives a slope of 0.32 ($t(98) = 6.02$, $p < .001$), positive, though well below 1. Because slope depends on both correlation and relative dispersion, this estimate is consistent with the compression result rather than an independent test of it.

Table 4. H3a/H3b: Correspondence and departure of the LLM overall-rating proxy versus Glassdoor (1–5 scale)

Metric	Value
Matched companies	100
Reviews in benchmark	1,213,399
Mean gap (LLM proxy minus Glassdoor)	+0.06
Two-sided test of gap = 0	$t(99) = 2.55, p = .012$
95% CI for mean gap	[+0.01, +0.11]
Companies rated above Glassdoor mean (%)	66.0
Pearson r	.52
Spearman ρ	.48
SD ratio (LLM / Glassdoor)	0.62

Note. Gap is the company-level LLM overall-rating proxy minus the Glassdoor overall mean over all matched reviews (2008 through June 2026). The proxy averages the innovation and employer-reputation items. The mean gap is evaluated with a two-sided test and confidence interval. No equivalence margin was prespecified. The SD ratio compares company-level standard deviations and measures range compression directly.

The correspondence does not depend on averaging the two proxy items. Analyzed separately, the innovation item has a +0.07 mean gap, Pearson $r = .45$, Spearman $\rho = .42$, and SD ratio 0.68. The reputation item has a +0.06 mean gap, Pearson $r = .52$, Spearman $\rho = .50$, and SD ratio 0.66. The two items correlate at $r = .74$ across companies. This sensitivity check supports the proxy’s broad conclusion while preserving the caveat that neither item is identical to Glassdoor’s overall measure and that the reputation wording invokes public perception and rankings.

Scale is associated with alignment. Disaggregating the benchmark by model reveals a size association that complements H1b. The correlation between a model’s company ratings and Glassdoor’s rises steadily with parameter count, from a mean r of .20 among models of at most 3B parameters to .47 at >20–70B and .56 for the largest model (Figure 2B, Spearman correlation between log size and per-model alignment, $\rho = .64, p < .001$). The .56 endpoint is descriptive of one model in the >70B bucket. Leaving each of the 22 families out in turn keeps the correlation between .61 and .69. Per-model mean difference on the overall-rating proxy, by contrast, is unrelated to size ($\rho = .06, p = .66$). Larger models do not rate employers more favorably overall, but their rankings align more closely with the social benchmark. Compression also eases with scale as the smallest models’ company-rating spread is about three-quarters of Glassdoor’s, while models above 20B approach parity. Together with H1b, this bears on the access-stratification concern raised in the Introduction. In this open-weight sample, smaller models were less robust across the temperature sweep and less aligned with the human review consensus. If lower-cost or freely available tools disproportionately rely on smaller models, their users may therefore receive less robust and less aligned employer evaluations. The present study does not test how commercial access tiers map to model size.

H3a is supported for rank-order correspondence of the overall-rating proxy ($\rho = .48$, Pearson $r = .52$, weighted slope 0.95). The separate mean calibration estimate is small (+0.06 stars, 95% CI [+0.01, +0.11]) but is not an equivalence test. The correspondence is item-robust and is stronger among larger models (Figure 2B). Mean per-model r rises from .20 to .47 across the multi-model buckets, while the sole >70B model reaches .56.

4.5 Systematic Departure Relative to Human Reviews

Mean calibration gap. Against the full corpus, the +0.06 mean gap is statistically distinguishable from zero under a two-sided test ($t(99) = 2.55, p = .012$, 95% CI [+0.01, +0.11]) but small in practical terms. As the benchmark-vintage analysis below shows, its direction and magnitude depend on the review window rather than constituting stable across-the-board positivity.

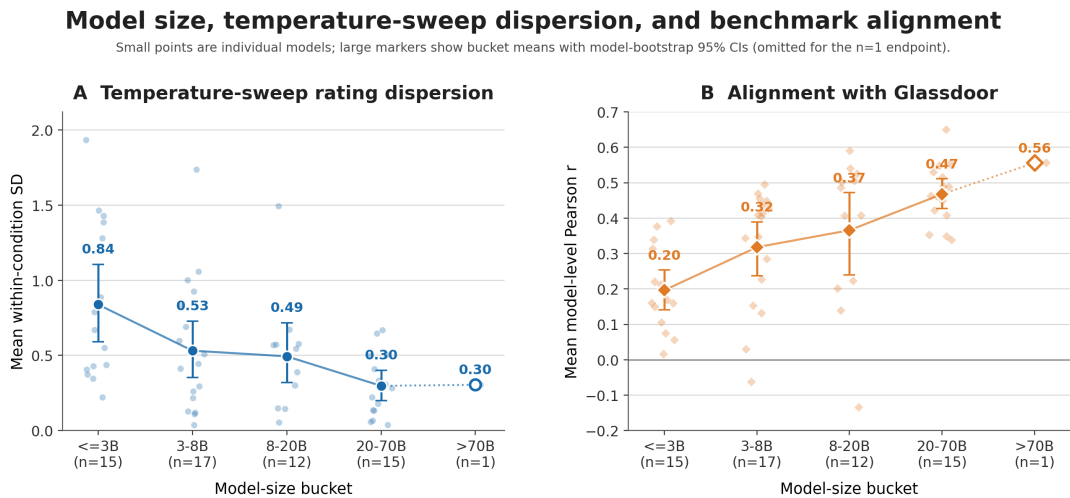


Fig. 2. Model-size associations with temperature-sweep dispersion and external alignment. Small points show individual models. Large markers show bucket means with 95% percentile intervals from 20,000 within-bucket bootstrap resamples of models. Panel A recomputes the condition-weighted mean in each resample using each selected model’s number of estimable temperature-sweep cells. Panel B uses the unweighted mean of per-model Pearson correlations with Glassdoor company means. The buckets are $\leq 3B$, $>3-8B$, $>8-20B$, $>20-70B$, and $>70B$, containing 15, 17, 12, 15, and 1 model, respectively. Bucket summaries are descriptive. Spearman correlations with model size, reported in the text, provide the inferential size checks. These summaries contain 60 rather than 61 models because gemma:2b did not yield sufficient valid ratings for model-level estimates. The $>70B$ estimate is descriptive of one model and therefore has no interval.

Range compression. The most consistent way the portrayals depart from the human benchmark is that they flatten differences between employers. The cross-model consensus has a company-level SD **0.62 times** that of Glassdoor means. At the individual-model level, the closer analogue to a user’s single-model query, the median SD ratio is 0.85 (interquartile range 0.56–1.06), and 65% of models compress the range. Compression attenuates with scale, with models above 20B approaching the benchmark’s spread. By contrast, averaging across all 60 models compresses the range further. In discrete-star terms, among the 31 firms Glassdoor places at 3 stars, the consensus rounds up to 4 stars for 81% of them, while none of the 69 firms at 4 stars is rounded up to 5 stars. The rank-order correlation ($\rho = .48$) is invariant to any monotone rescaling of the LLM scale. Descriptively, the difference is largest for lower-rated firms. Regressing the LLM–Glassdoor difference on the Glassdoor mean gives a slope of -0.68 (Figure 3B).

Benchmark-vintage sensitivity. Recomputing the comparison within review-date windows (Table 5) shows that the aggregate mean gap depends on the benchmark’s age and cannot be attributed to the models alone. Against 2008–2015 reviews the same LLM ratings appear higher by $+0.31$ stars (86% of firms above the benchmark), against 2016–2019 reviews by $+0.10$, and against 2020–2026 reviews they sit slightly *below* consensus (-0.05 , with 39% of firms above the benchmark). Meanwhile, Glassdoor’s own average company rating rose from 3.39 to 3.75 across these windows. A finer yearly decomposition (Figure 3C) shows the gap falls from roughly $+0.5$ against 2008–2010 reviews to an annual average of about $+0.11$ across 2016–2019 reviews and about -0.08 against 2020–2022 reviews, and is essentially zero against 2024 reviews (-0.01), the most recent year in the yearly analysis. Because the July-2024 census did not cover a complete calendar year and the added firms enter mainly in 2026, the yearly endpoints are partial and composition varies. But the gradient is not a composition artifact. In the balanced panel of 86 firms with at least 30 reviews in every window,

LLM portrayals correspond with—but systematically reshape—human employer reviews

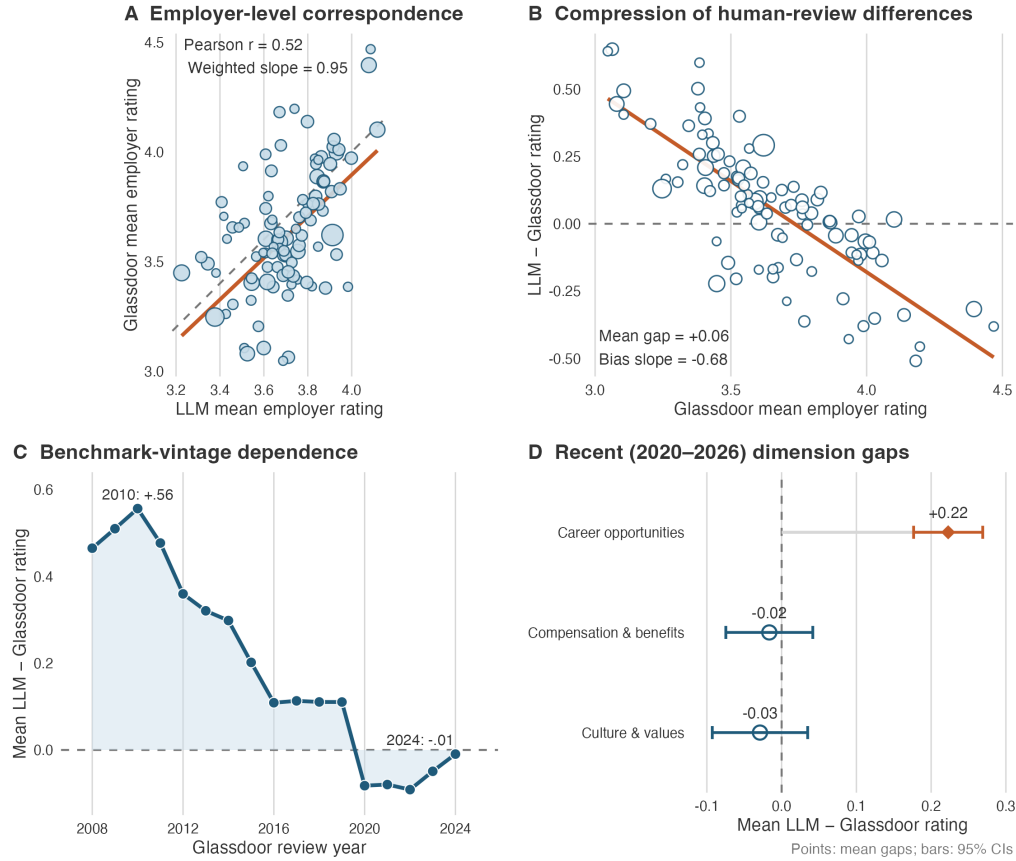


Fig. 3. Correspondence and systematic departure relative to human employer reviews. Panel A compares the company-level LLM overall-rating proxy and Glassdoor means ($n = 100$), with point area reflecting review count. The dashed line marks equality, and the solid line is the review-count-weighted fit. Panel B plots rating gaps against Glassdoor means, showing range compression. Panel C recomputes the mean gap using each review year's benchmark. Panel D compares dimension-specific gaps with 2020–2026 reviews. Career opportunities remain elevated while compensation and culture gaps are not detectably different from zero.

the gaps for 2008–2015, 2016–2019, and 2020–2026 are +0.30, +0.11, and -0.09. Across the same three windows, the shares above the benchmark are 86.0%, 69.8%, and 32.6%, and Pearson correlations are .49, .40, and .58. The cross-firm ordering tells the same story. LLM company means correlate more strongly with the 2022–2024 consensus ($r = .54$) than with the 2016–2019 consensus ($r = .40$), and when both windows enter the same regression, the recent window absorbs all the unique predictive weight ($b = 0.36$, $p < .001$, versus $b = -0.01$, ns, for the older window, with 86 firms having at least 30 reviews in each). Against the 2020–2026 benchmark, LLM levels are close to the benchmark mean. The design establishes benchmark-vintage sensitivity, but it does not observe each model's training corpus and therefore cannot establish lag, training era, or that reviews from any particular period entered training or caused the pattern. It does establish that an audit of an LLM calibration gap against human ratings must hold the human benchmark's vintage fixed before interpreting that gap.

Table 5. H3b: Benchmark-vintage sensitivity (the same LLM ratings compared with reviews from different periods)

Review window	Companies	Reviews	Glassdoor mean	Mean difference	Higher than Glassdoor (%)	Pearson r
2008–2015	87	195,555	3.39	+0.31	86.2	.48
2016–2019	88	266,165	3.60	+0.10	69.3	.38
2020–2026	100	751,679	3.75	-0.05	39.0	.50
Full window (2008–2026)	100	1,213,399	3.64	+0.06	66.0	.52

Note. Each row recomputes the company-level comparison using only reviews dated within the window. The LLM ratings are identical across rows. Companies appear in a window only if they have reviews dated within it (the 12 firms added in the June-2026 follow-up contribute mainly to the most recent window). Glassdoor’s average company rating rose by roughly 0.4 stars over the covered period, so the aggregate mean calibration gap shrinks monotonically (and reverses sign) as the benchmark becomes more recent.

The headline correspondence is not driven by the two-phase collection. Restricting the full-window analysis to the 88 firms in the July-2024 census yields a +0.05 mean gap, Pearson $r = .54$, and SD ratio 0.67, similar to the all-firm estimates.

Career-opportunities ratings remain elevated in the dimension-level comparison. In addition to the overall-rating proxy, the protocol measures three dimensions that correspond to Glassdoor sub-ratings, namely career opportunities, compensation and benefits, and culture and values. We compare each LLM dimension with its corresponding Glassdoor sub-rating to determine whether the benchmark-vintage pattern also holds within specific dimensions. It holds for two of the three. Against recent reviews the career-opportunities gap is the one that remains (Table 6). Compensation and culture behave like the overall-rating proxy. Their full-corpus gaps are small (+0.05 each, ns), and their gaps are not detectably different from zero against the 2020–2026 benchmark (-0.02 and -0.03). However, the career-opportunities sub-dimension is elevated by +0.35 stars against the full corpus ($t(99) = 16.1$, $p < .001$, with 94 of 100 firms and 52 of 60 models above the benchmark on average) and still by +0.22 against 2020–2026 reviews, as Figure 3D shows ($t(99) = 9.6$, $p < .001$, 87% of firms). The elevation in portrayal is on a sub-dimension where employees are most critical. Career opportunities is the lowest-rated of the three dimensions on Glassdoor itself (3.45, versus 3.54 for culture and 3.69 for compensation), and compression is sharper at the dimension level as well (SD ratios of 0.44 to 0.49, versus 0.62 for the overall-rating proxy). An exploratory keyword analysis suggests that models typically justify favorable career-opportunities ratings by citing formal development programs rather than employees’ lived experiences. With one fixed item per dimension, the design cannot fully separate model-side displacement from this item–criterion mismatch or from dimension-specific reviewer selection (the optional career sub-rating may draw reviewers for whom advancement was salient), though the absence of displacement on the parallel compensation and culture items constrains these alternatives. In short, the observed positive displacement is local to the growth-prospects dimension, which job seekers may act on and which the largest, otherwise best-aligned models elevate most.

H3b is supported as the departure is systematic, taking the form of range compression, benchmark-vintage sensitivity, and dimension-specific displacement. Compression is robust, the apparent level difference depends on benchmark vintage, and career opportunities remain displaced upward. The specifically *positivity-shaped* form of departure, however, appears only against older review vintages and on the career-opportunities dimension, not as the across-the-board rosiness the sycophancy literature predicts.

4.6 Cross-Model Convergence and Company-Level Divergence

Two further patterns fall outside the hypothesis tests but bear directly on what AI-mediated employer information looks like in practice.

Table 6. H3b: Dimension-level calibration of LLM sub-dimension ratings versus Glassdoor sub-ratings

Dimension	Mean difference, full corpus	Mean difference vs. 2020–2026 reviews	Higher than Glassdoor, 2020–2026 (%)	Pearson r	SD ratio
Career opportunities	+0.35 (t(99) = 16.1)	+0.22 (t(99) = 9.6)	87.0	.58	0.48
Compensation and benefits	+0.05 (t(99) = 1.7)	-0.02 (t(99) = -0.6)	43.0	.48	0.49
Culture and values	+0.05 (t(99) = 1.8)	-0.03 (t(99) = -0.9)	46.0	.53	0.44
Overall-rating proxy	+0.06 (t(99) = 2.55)	-0.05 (t(99) = -1.8)	39.0	.52	0.62

Note. Each LLM dimension rating is the across-model consensus for that company, rescaled to Glassdoor's 1–5 scale. Glassdoor sub-ratings are optional reviewer fields, and blank (zero-stored) entries are treated as missing. The overall-rating proxy averages the innovation and employer-reputation items. Mean difference is the company-level LLM mean minus the Glassdoor mean ($n = 100$ firms). In the two 2020–2026 columns, firms enter with at least 30 valid sub-ratings. Pearson r and the SD ratio are computed against the full corpus. The career-opportunities elevation is the only positive gap that remains in the 2020–2026 benchmark. Against the full corpus, it is essentially unchanged among the 88 census-collected firms (+0.34) and is largest for the consensus of models above 20B parameters (+0.43).

Different models, one answer. If each model offered an independent reading of the evidence, consulting several would amount to collecting second opinions. The data show the opposite. Across the 60 models, the average correlation between two models' company-rating profiles (pairwise $r = .45$) exceeds the average correlation between a model and Glassdoor ($r = .34$), and the agreement rises steeply with scale. Mean pairwise correlations climb from .18 among models of at most 3B parameters to .80 among the 16 models above 20B. This convergence is not explained simply by common correspondence with the Glassdoor benchmark. Partialing the Glassdoor means out of every pairwise comparison leaves the agreement among the largest models nearly intact (partial $r = .75$). Large models share a view of employers that departs from the benchmark in similar ways. This also cannot be explained by shared model family. Above 20B, pairs of models from *different* families agree nearly as strongly ($r = .80$ across 117 pairs) as pairs from the same family ($r = .88$ across 3 pairs). The 60-model consensus reaches $r = .52$, exceeding 52 of the 60 individual models, while eight individual models exceed the consensus. This pattern is consistent with correlated departures that averaging cannot remove. Among the largest models, therefore, asking another model is largely asking the same informant again. Whether proprietary, retrieval-augmented assistants behave likewise is now an open question.

The same employer, many answers. Convergence among the largest models coexists with substantial heterogeneity across the full model market a job seeker actually faces. For the typical firm, the 60 models' company ratings span 2.9 stars on the 1–5 scale (median cross-model SD = 0.43), and by the discrete star rating an average of 23% of models sit on the minority side of the 4-star threshold. For 36% of firms, at least a quarter of models disagree with the majority (Appendix Table A4). This divergence is not concentrated where one might expect. It is unrelated to a firm's Glassdoor level ($r = .08$, ns) and its review volume ($r = -.09$, ns), and it does not differ by sector ($F(7, 92) = 0.75$, ns). The *direction* of the mean consensus–benchmark difference, however, varies by sector in the full-window benchmark ($F(7, 92) = 2.82$, $p = .010$). Energy firms are portrayed 0.20 stars *below* the full-window Glassdoor benchmark, whereas consumer-staples (+0.17) and health-care (+0.16) firms are portrayed above it. We return to the employer-side implications in the Discussion.

Takeaway. LLMs provide a moderately aligned *directional* signal whose correspondence depends on benchmark vintage. As a channel, they compress information and exhibit correlated departures from the benchmark. They mute differences between employers (pulling weak employers up toward the middle), align more closely with recent than older review windows, and rate career opportunities above the 2020–2026 Glassdoor sub-rating. Because large models converge on the same answer, consulting several is not a second opinion in the classic sense.

5 Discussion

5.1 Core Findings

The paper's central finding is what we call *calibration with compression*. Calibration refers to the average level of the ratings. Across companies, the LLM ratings differ only slightly from the full-window Glassdoor benchmark. The LLM and Glassdoor ratings also show moderate agreement in how they order companies, meaning that firms rated more highly on Glassdoor generally receive higher LLM ratings as well. Compression refers to the amount of differentiation between companies. The company ratings produced by the cross-model consensus have only 62% as much spread (standard deviation across companies) as the Glassdoor ratings. This pattern is less pronounced, but still present, for individual models, with the median model having 85% as much spread as Glassdoor. Thus, the models preserve part of the relative ordering of employers while making the companies appear more similar than they do in human reviews. This compression occurs especially because lower-rated employers are pulled upward toward the middle of the rating distribution.

Model size helps explain some, but not all, of these patterns. Parameter count has little association with the average favorability of a model's ratings. Larger models do not systematically rate employers higher or lower than smaller models. They do, however, produce ratings that vary less across the standardized temperature sweep and correspond more closely with the Glassdoor ordering. Because model size was not experimentally manipulated and is associated with other model characteristics, these findings should be interpreted as associations rather than causal effects of scale.

Larger models also converge with one another. Among models above 20B parameters, company-rating profiles correlate at $r = .80$. Agreement among several large models therefore does not necessarily provide several independent assessments of an employer. It may instead reflect multiple models producing similar portrayals. For job seekers, this means that consulting another large model may provide less additional information than the appearance of a second opinion suggests.

The comparison with Glassdoor also depends on which review period is used as the benchmark. Across the full set of 1.2 million reviews, the cross-model consensus is only 0.06 stars above the Glassdoor company means and correlates with them at $r = .52$. This close average correspondence coexists with substantial compression. The consensus covers less than two-thirds of the company-level spread found in Glassdoor, and most companies with a 3-star Glassdoor rating are rounded up to 4 stars by the LLM consensus.

The average difference changes when the same LLM ratings are compared with reviews from different periods. Against reviews from 2008–2015, the LLM ratings are 0.31 stars higher on average. Against reviews dated 2020–2026, the average difference is close to zero. The result therefore shows benchmark-vintage sensitivity, meaning that an LLM can appear more or less positive depending on the period of human reviews used for comparison. It does not establish that the models are temporally lagged, identify what information entered their training data, or explain why their ratings correspond more closely with some review periods than with others.

The dimension-level results identify one important exception. The LLM career-opportunities item remains 0.22 stars above the corresponding Glassdoor sub-rating even when the comparison is restricted to reviews dated 2020–2026. The compensation and culture gaps, by contrast, are not detectably different from zero in that recent benchmark. The career-opportunities difference should not be attributed entirely to model behavior, however. Each dimension is measured with a single LLM prompt, which may not capture exactly the same construct as the corresponding Glassdoor item. Glassdoor sub-ratings are also optional, so the reviewers who rate career opportunities may differ systematically

from those who leave that field blank. The observed gap may therefore reflect both model behavior and differences in measurement and reviewer selection.

The experiment also shows that system-role wording can shift numeric evaluations. Changing the system prompt from “helpful” to “neutral” produces less favorable binary and ordinal responses. Because the design compares only these two roles, it establishes a difference between them but does not identify which one departs from an unprompted baseline. It also does not establish how the hidden system prompts used by commercial chatbots are configured.

The free-text results require a different interpretation. Responses to different prompt topics use substantially different kinds of language. Questions about culture elicit more affiliative, identity-oriented language and less instrumental language than the other four free-text prompts, which ask about reputation, growth and development, employer attractiveness, and what employees praise and criticize. However, topic and wording vary together in the prompt battery. These differences are therefore descriptive, in that they show how the responses differ across the prompts used in this study but do not isolate a causal effect of topic or wording alone.

Model families also display symbolic-versus-instrumental language signatures, although these differences are modest and depend on the dictionary used to measure them. Model scale is more consistently associated with temperature-sweep robustness and correspondence with Glassdoor. Taken together, these results show that model family, model size, system-role wording, and prompt topic relate to different aspects of an employer portrayal. They should not be treated as interchangeable explanations for the same outcome.

Methodologically, the study provides a documented approach for auditing employer portrayals across many language models. The pipeline standardizes the companies, prompts, system roles, and temperature settings used for each model. It validates the resulting binary, ordinal, and free-text responses, and it uses statistical inference that accounts for responses being clustered within models. The full collection protocol appears in Appendix B. These documented design and analysis choices support methodological replication of the audit.

Core takeaway. The models audited here should not be treated as sole arbiters of comparative employer quality. Their average ratings may appear well calibrated against Glassdoor, but that average agreement conceals meaningful differences in how they portray individual employers. The models compress distinctions between companies, their ratings depend partly on model scale and hidden system-role wording, and large models often repeat similar assessments rather than provide independent perspectives.

LLMs are therefore better positioned as tools for gathering, organizing, and generating questions about employer information than as final authorities on where someone should work. Job seekers can use them to structure a search, but consequential comparisons among employers should be checked against current reviews, role-specific conversations, and other direct sources.

5.2 Theoretical Implications for Employer Branding

Our findings position LLMs as algorithmic channels whose conversational outputs serve users as a distinct fourth-person informant, not as an independent source of underlying social evidence. Rather than passively reflecting a firm’s reputation or aspects of its employer value proposition, this intermediary transforms them along at least four axes. The first is compression, with company-level spread muted to under two-thirds of the human benchmark’s and weak employers pulled toward the middle. The second is benchmark-vintage sensitivity, in which levels correspond more closely with recent than older review consensus, without identifying lag or revealing which evidence entered training. The third is context sensitivity, in which the experimentally manipulated system role changes evaluation level, while descriptive language differs across prompt topics. The fourth is homogenization, with models, especially the largest,

converging on one shared answer, so that apparent variety across providers does not deliver variety of perspective. This has direct implications for how awareness and associations are formed and maintained, extending existing literature to an AI-mediated environment.

For SIP, the results give empirical content to the fourth-person informant we theorized. SIP assumes that informants vary, and that decision makers can partly calibrate for a source’s slant through repeated interaction and social context. The fourth-person informant breaks both assumptions in measurable ways. Its slant is systematic but invisible, its identity is unstable across providers, and its consistency varies with attributes that no lay user can observe. The theory’s classic moderators of cue influence are credibility and consistency, exactly the properties our audit quantifies, which suggests a research program, namely treating model identity, scale, and provider-set role framing as informant characteristics within SIP rather than as technical parameters. Within this open-weight sample, observed convergence makes large-model answers resemble the same fourth person, so the canonical SIP remedy of consulting additional informants can fail when all of the informants are large models. These four transformation axes map directly onto signal properties receivers use. Compression erodes *distinctiveness* and *diagnosticity*, benchmark-vintage sensitivity makes *recency* uncertain, provider-set role framing reflects covert *source control over credibility*, and homogenization manufactures *perceived consensus*.

Employer brand signals have traditionally been discounted by receivers who understand the sender’s incentives (Celani and Singh, 2011). Review platforms emerged precisely as a less sender-controlled channel (Dineen et al., 2019). An LLM that compresses the rating range re-introduces a departure that favors weaker senders, but one that originates in the channel rather than the sender, and that receivers may have little basis to discount if the tool is perceived as neutral. The vintage result shows that correspondence changes with the review window, but the cross-sectional benchmark comparison does not establish whether, when, or how quickly changes in employer conditions enter the AI channel.

LLMs also re-weight which brand characteristics are made salient as a function of prompt topic and specificity, which may affect how brand associations form. This introduces a query-contingent pathway for brand construction, although the present prompt-topic contrasts do not isolate wording from topic.

In AI-mediated search, information distinctiveness competes with a smoothing force where the models have to compress complex representations of the world. This can dilute differentiation of brand messaging. The implication is a shift from “what is said” to “what is machine-readable and easily surfaced” by an AI system.

5.3 What Do LLM Employer Portrayals Reflect?

Correlation with Glassdoor does not imply that models evaluate and portray employers accurately. It implies, at most, that models have absorbed the public sentiment record well enough to reproduce its central tendencies. Three results are consistent with shared provenance, though they do not identify it. First, model ratings correspond more closely with recent than older review windows, establishing benchmark-vintage sensitivity without identifying its cause. Because the design does not observe model training corpora, it cannot locate a training era or infer that 2024 reviews caused this correspondence. Second, large models converge on shared *departures* from the benchmark (pairwise partial $r = .75$ after removing the Glassdoor signal). Independent reasoners would not depart from the benchmark in the same ways. Third, the 60-model consensus correlates more strongly with the benchmark than 52 individual models do but less strongly than the remaining eight. This is consistent with shared, rather than independent, departures that averaging cannot remove. If Glassdoor-derived text contributes to training corpora, our correspondence estimate could partly reflect memorization, and “alignment with human consensus” should be read as *correspondence with the benchmark’s*

source material, not as independent validation of employer quality. One finding cuts against the simplest version of this story. The absolute difference between the cross-model consensus and the human consensus for a given firm is unrelated to how many Glassdoor reviews that firm has ($r = -.12$, ns). If alignment were just “more Glassdoor text in training,” the most-reviewed firms should be matched most closely, and they are not. Even so, an audit like this can document how a model’s portrayals compare to a benchmark, but not why the model produces them.

5.4 High-Information Employers and Open-Weight Models

Two scope conditions bound our claims. First, the Fortune 100 constitute a most-likely case for correspondence. They are heavily documented on the public web, on review platforms, and in business journalism. For smaller, younger, or less-visible employers, the information environment thins, and the same compression-toward-the-middle dynamic plausibly fills the gap with genericized or fabricated content. Our correspondence estimates should therefore be read as an upper bound on correspondence, not extrapolated down-market, though it is notable that even within this high-information set, the absolute consensus–benchmark difference is unrelated to a firm’s review volume. This nonassociation is consistent with a plateau once firms are sufficiently documented, but the cross-sectional design cannot establish that mechanism.

Second, our claims characterize open-weight models under controlled conditions. This design choice makes the system prompt and temperature controllable and the procedure consistently documented. A methodological replication should use the listed model tags, prompts, system roles, temperature schedule, validation rules, and analysis specifications, then evaluate the same design-level estimands. As in other stochastic-generation studies, token-level outputs may vary across software, hardware, and model builds. Proprietary LLM systems differ in alignment, scale, and increasingly in retrieval augmentation, which could alter benchmark-vintage sensitivity or compression. That is an empirical question, not a presumption in either direction. Claims about “the systems most job seekers use” therefore require the proprietary extension we outline in Section 5.8. Our contribution is the controlled baseline and the audit protocol itself.

5.5 Guidance for Job Seekers

LLMs are a useful *starting point* for employer research but a poor *final arbiter*. We translate each result into concrete advice.

Don’t mistake agreement for corroboration (exploratory convergence). Individual models differ in how they characterize the same employer, and not always along family lines, yet the largest models converge on the same ranking (pairwise $r = .80$), agreeing with each other more than with Glassdoor. Asking two or three models is still worthwhile, but treat consistent answers across large models as one opinion repeated, not several confirming, and treat divergence as a signal that the question is genuinely contested.

Prefer larger models for robustness, while treating alignment as correspondence, and re-ask (H1b, H3a). Smaller models give evaluations that are less robust across the standardized temperature sweep and correspond less closely to the Glassdoor ordering. This is benchmark correspondence, not evidence that larger models assess employers more accurately. Two further caveats apply. First, consumer products often obscure which model (and what size) is answering, so model identity is itself a question worth asking. Second, temperature-sweep robustness is not fixed-setting test–retest reliability. Varied re-asking is only the closest user analogue to our designed perturbation. If an available model is known to be small, ask the same question more than once. If answers vary significantly, discount them.

System roles can shift answers (H2). Our two-role contrast produced more favorable answers under “helpful” than “neutral.” Without an unprompted baseline, it does not identify which role departs from default behavior or how

commercial chatbots are configured. Counteract such framing by asking explicitly for balance (“what would a critic say?”, “list concerns, not just strengths”).

Mind the compression and the clock (H3b). The models in our audit showed moderate correspondence with the review benchmark, but the median individual model covered only 85% of the benchmark’s spread and the cross-model consensus only 62%. Most 3-star firms come back sounding like 4-star firms. Use the LLM for generating questions, but check against direct sources. The benchmark-vintage result makes timeliness an open question but does not establish lag. Treat the career-opportunities item with particular caution. It averages +0.22 stars above Glassdoor sub-ratings dated 2020–2026, although item wording and reviewer selection may explain part of the gap.

Use LLMs to structure, not to decide. They are strong for orienting, summarizing, generating questions, and triangulating sources, but weak as a single normative signal for “is this a good employer?” For final judgments, weigh current employee reviews, role-specific conversations, and direct recruiter and hiring-manager information.

5.6 Implications for Employers in Monitoring AI-Generated Portrayals

A firm’s AI portrayal is now a reputational asset the firm neither authors nor can directly inspect, and our divergence analysis shows that firms differ in how much that portrayal depends on which model a candidate happens to consult. For the typical firm, ratings span 2.9 stars across this broad set of open-weight models, and roughly a quarter of models disagree with the majority rating at the 4-star threshold (Appendix Table A4). High-divergence firms are portrayed materially inconsistently across models, while low-divergence firms receive nearly the same answer everywhere. Whether that inconsistency translates into differing candidate impressions is not tested here. Notably, divergence is unrelated to a firm’s Glassdoor level or review volume. A strong human reputation does not purchase a consistent AI portrayal.

Two systematic displacements deserve employer attention. First, the *direction* of the consensus–benchmark difference is sector-patterned. Energy firms are portrayed 0.20 stars below the full-window Glassdoor benchmark, while consumer-staples and health-care firms sit above it. Second, the benchmark-vintage pattern shows that the apparent calibration gap depends on the comparison period. The present cross-sectional audit cannot establish how quickly improvement or deterioration would be reflected.

Just as firms learned to manage Glassdoor by responding to reviews, they will need strategies for the AI layer. The practical agenda that follows from our results has three parts. The first is periodic cross-model portrayal audits (the protocol here is reusable). The second is specific attention to the career-opportunities dimension (the one persistent elevation, and one that may set up expectation violations for new hires). The third is monitoring portrayal divergence as an early-warning indicator of a contested or thinly documented reputation. For research, this raises a question we flag for future work. How can organizations shape what LLMs convey about them? Candidate levers include the volume and machine-readability of authoritative employer content, third-party review presence, and structured data that models can ingest and surface. The shift is from “what is said” to “what is machine-readable and easily surfaced.” Firm-level moderators of divergence (visibility, sector, documentation depth) and intervention studies with portrayal change as the outcome are natural next steps for employer branding research.

5.7 Limitations

(1) We audit portrayals, not perceptions. Effects on trust, employer image, and application intentions are downstream links of the conceptual chain that remain untested here. (2) We study open-weight models served locally. Widely used proprietary systems (GPT-4/5-class, Claude, Gemini) may behave differently, though open-weight coverage is

what makes the protocol controllable and procedurally documentable. The documented protocol supports replication across open-weight model environments. (3) Glassdoor is a convenience human benchmark with its own self-selection biases, and correlation with it may partly reflect training-data overlap with the benchmark’s source material. Each sub-dimension benchmark, moreover, rests on one fixed LLM item matched to an optional reviewer field, so dimension-level gaps (most notably career opportunities) cannot be fully separated from item-wording and dimension-specific selection effects. The overall-rating proxy likewise averages two items rather than directly asking Glassdoor’s overall “place to work” question, and one component explicitly mentions public perception and employer rankings. The Glassdoor corpus was assembled in two phases (a July-2024 census for 88 firms and a June-2026 recent-review follow-up for 12). The census-only sensitivity reduces but does not eliminate the resulting composition concern. (4) The sample is the Fortune 100, a high-information setting. Correspondence estimates are best read as an upper bound and should not be extrapolated to less-documented employers. (5) Responses are in English and reflect a single collection window, so the study does not estimate changes across collection dates or support causal claims about temporal lag. (6) The design is observational at the model level (we did not intervene on training), and with 61 models, size and family are partially confounded. No family spans the full size range, so size-curve extrapolations across families should be read with care (the leave-one-family-out checks bound, but cannot eliminate, this concern, and parameter count proxies a bundle of correlated model properties, such as training data, recency, and instruction tuning, rather than scale alone). We do not observe the models’ training corpora, so the benchmark-vintage result cannot identify training-data provenance or a causal training era. (7) Invalid-response attrition is concentrated among the smallest models. The dispersion gradient persists when analysis is restricted to complete temperature sweeps, but available-response estimates can still be affected by nonrandom formatting failures.

5.8 Future Work

Human-perception studies, testing the chain’s downstream links. Field and lab studies should place job seekers in front of AI-generated portrayals (versus review-platform and employer sources) and measure trust, source credibility, employer image, and application intentions. Two of the properties documented here are directly manipulable. Compression can be varied by restoring or further muting the spread between companies, and apparent consensus by presenting one model’s answer rather than several convergent ones. Such designs would convert the portrayal properties documented here into testable predictions about receiver-side effects, and would establish whether receivers discount a portrayal once its machine origin is disclosed.

Cognitive offloading. Whether reliance on LLM summaries reflects genuine cognitive offloading, and how it changes the depth and diversity of information job seekers gather, remains untested. Process measures such as the number and type of sources consulted, time to decision, and willingness to seek disconfirming evidence would separate an LLM that orients a search continuing elsewhere from one that ends it. Convergence sharpens the question. If agreement among large models is read as corroboration, offloading would be reinforced by a signal carrying less independent information than it appears to.

Employer influence. Experiments on whether and how firms can change LLM portrayals of themselves would establish how contestable those portrayals are. Intervention studies could manipulate the levers sketched above (authoritative employer content and its machine-readability, third-party review presence, and structured data), with portrayal change as the measured outcome. The same designs would show whether the AI layer can eventually be managed as review platforms were, or whether it resists firm-side influence strongly enough that periodic auditing is the only practical remedy.

Proprietary and longitudinal coverage. Extending the protocol to commercial and retrieval-augmented systems would test whether compression, convergence, and benchmark-vintage sensitivity survive the alignment and inference pipelines job seekers actually encounter. Repeating the collection in waves would convert our single-window, cross-sectional vintage result into a direct estimate of drift, and would establish how quickly, if at all, improvement or deterioration in employer conditions reaches the AI channel.

6 Conclusion

Job seekers are increasingly asking AI, “Is this a good place to work?” and getting a single, confident answer. The central result is calibration with compression. The overall-rating proxy has a small mean gap and moderate rank-order correspondence with Glassdoor. The cross-model consensus mutes differences between employers more than the median individual model. Model size is observationally associated with temperature-sweep robustness and benchmark correspondence, while the experimentally manipulated system role shifts numeric valence. Compression, benchmark-vintage sensitivity, context sensitivity, and convergence make these portrayals ill-suited as a sole source. The career-opportunities item, a dimension job seekers may weigh heavily, remains elevated relative to the 2020–2026 Glassdoor benchmark, and because large models give similar answers, consulting several of them is not a second opinion in the classic sense. Used as one voice among several, an assistant for orienting and structuring rather than a final arbiter, LLMs can help job seekers. Used alone, they transform the employer information on which people act, not by being wildly wrong, but by sounding calibrated while muting the differences that may matter.

7 Ethical Considerations

Ethics and data handling. The Glassdoor benchmark consists of reviews posted publicly by their authors and was collected via a commercial service from public employer pages. We use review data exclusively in aggregate (company-level and company-by-window means). No reviewer identities are analyzed or reported, no attempt is made to re-identify reviewers, and review text is not quoted. Records were deduplicated by review identifier. The LLM response corpus contains only machine-generated text. The WildChat analysis is a secondary analysis of an existing research dataset containing human-authored chat turns. The dataset providers report affirmative opt-in for the original collection (Zhao et al., 2024). The WildChat-4.8M dataset card reports deidentification with Microsoft Presidio, custom regular expressions, and manual adjustments (Allen Institute for AI, 2025). We analyzed only first-user-turn text, did not use location, request-header, or hashed-IP fields, made no attempt to link records or re-identify users, and report only aggregate rates and a small number of deidentified, non-sensitive examples. Thus, this study did not recruit or interact with participants, but it does analyze previously collected human-authored text.

References

- Allen Institute for AI. (2025). *WildChat-4.8M* [Data set]. Hugging Face. <https://huggingface.co/datasets/allenai/WildChat-4.8M>
- Ambler, T., & Barrow, S. (1996). The employer brand. *Journal of Brand Management*, 4(3), 185–206. <https://doi.org/10.1057/bm.1996.42>
- Askill, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T. B., Clark, J., . . . Kaplan, J. (2021). A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*. <https://doi.org/10.48550/arXiv.2112.00861>
- Bangerter, A., Roulin, N., & König, C. J. (2012). Personnel selection as a signaling game. *Journal of Applied Psychology*, 97(4), 719–738. <https://doi.org/10.1037/a0026078>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2022). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. <https://doi.org/10.48550/arXiv.2108.07258>

- Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. University of Texas at Austin. <https://www.liwc.app>
- Cable, D. M., & Turban, D. B. (2001). Establishing the dimensions, sources and value of job seekers' employer knowledge during recruitment. In G. R. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 20, pp. 115–163). Emerald Group Publishing. [https://doi.org/10.1016/S0742-7301\(01\)20002-4](https://doi.org/10.1016/S0742-7301(01)20002-4)
- Celani, A., & Singh, P. (2011). Signaling theory and applicant attraction outcomes. *Personnel Review*, 40(2), 222–238. <https://doi.org/10.1108/00483481111106093>
- Cheng, X., Zhang, X., Cohen, J., & Mou, J. (2022). Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms. *Information Processing & Management*, 59(3), 102940. <https://doi.org/10.1016/j.ipm.2022.102940>
- Coher, R. T., Brown, D. J., Keeping, L. M., & Levy, P. E. (2004). Recruitment on the net: How do organizational web site characteristics influence applicant attraction? *Journal of Management*, 30(5), 623–646. <https://doi.org/10.1016/j.jm.2004.03.001>
- Collins, C. J., & Kanar, A. M. (2013). Employer brand equity and recruitment research. In K. Y. T. Yu & D. M. Cable (Eds.), *The Oxford handbook of recruitment* (pp. 284–297). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199756094.013.0016>
- Collins, C. J., & Martinez-Moreno, J. E. (2022). Recruitment brand equity for unknown employers: Examining the effects of recruitment message claim verifiability and credibility on job pursuit intentions. *Human Resource Management*, 61(5), 585–597. <https://doi.org/10.1002/hrm.22089>
- Collins, C. J., & Stevens, C. K. (2002). The relationship between early recruitment-related activities and the application decisions of new labor-market entrants: A brand equity approach to recruitment. *Journal of Applied Psychology*, 87(6), 1121–1133. <https://doi.org/10.1037/0021-9010.87.6.1121>
- Connolly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management*, 37(1), 39–67. <https://doi.org/10.1177/0149206310388419>
- Dineen, B. R., Van Hove, G., Lievens, F., & Rosokha, L. M. (2019). Third party employment branding: What are its signaling dimensions, mechanisms, and sources? In M. R. Buckley, A. R. Wheeler, J. E. Baur, & J. R. B. Halbesleben (Eds.), *Research in personnel and human resources management* (Vol. 37, pp. 173–226). Emerald Publishing. <https://doi.org/10.1108/S0742-730120190000037006>
- Evertz, L., Kollitz, R., & Süß, S. (2021). Electronic word-of-mouth via employer review sites: The effects on organizational attraction. *The International Journal of Human Resource Management*, 32(16), 3428–3457. <https://doi.org/10.1080/09585192.2019.1640268>
- Glassdoor. (2021, July 1). *40+ stats for companies to keep in mind for 2021*. <https://www.glassdoor.com/blog/40-stats-for-companies-to-keep-in-mind-for-2021/>
- Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2024). Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. <https://openreview.net/forum?id=kGteeZ18lr>
- Hayati, S. A., Lee, M., Rajagopal, D., & Kang, D. (2024). How far can we extract diverse perspectives from large language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 5336–5366). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.306>
- Höllig, C. E., Ademi, E., & Tumasjan, A. (2025). Employer responsiveness to online reviews: A signal of caring about employees. *Human Resource Management*, 64(5), 1481–1502. <https://doi.org/10.1002/hrm.22316>
- Höllig, C. E., Tumasjan, A., & Lievens, F. (2024). What drives employers' favorability ratings on employer review platforms? The role of symbolic, personal, and emotional content. *International Journal of Selection and Assessment*, 32(4), 579–593. <https://doi.org/10.1111/ijasa.12478>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*. <https://doi.org/10.48550/arXiv.2001.08361>
- Lievens, F., & Highhouse, S. (2003). The relation of instrumental and symbolic attributes to a company's attractiveness as an employer. *Personnel Psychology*, 56(1), 75–102. <https://doi.org/10.1111/j.1744-6570.2003.tb00144.x>
- Lievens, F., & Slaughter, J. E. (2016). Employer image and employer branding: What we know and what we need to know. *Annual Review of Organizational Psychology and Organizational Behavior*, 3(1), 407–440. <https://doi.org/10.1146/annurev-orgpsych-041015-062501>
- Lievens, F., Srivastava, M., & Saini, G. K. (2023). Third-party employment branding: Current status and future directions. In C. King & E. Murillo (Eds.), *A research agenda for brand management in a new era of consumerism* (pp. 239–256). Edward Elgar Publishing. <https://doi.org/10.4337/9781803925516.00022>
- Lievens, F., Van Hove, G., & Anseel, F. (2007). Organizational identity and employer image: Towards a unifying framework. *British Journal of Management*, 18(S1), S45–S59. <https://doi.org/10.1111/j.1467-8551.2007.00525.x>
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 195. <https://doi.org/10.1145/3560815>
- Love, L. F., & Singh, P. (2011). Workplace branding: Leveraging human resources management practices for competitive advantage through “Best Employer” surveys. *Journal of Business and Psychology*, 26(2), 175–181. <https://doi.org/10.1007/s10869-011-9226-5>
- Malmqvist, L. (2025). Sycophancy in large language models: Causes and mitigations. In K. Arai (Ed.), *Intelligent computing: Proceedings of the 2025 Computing Conference* (Vol. 4, pp. 61–74). Springer. https://doi.org/10.1007/978-3-031-92611-2_5
- Massenon, R., Gambo, I., Khan, J. A., Agbonkhese, C., & Alwadain, A. (2025). “My AI is lying to me”: User-reported LLM hallucinations in AI mobile apps reviews. *Scientific Reports*, 15(1), 30397. <https://doi.org/10.1038/s41598-025-15416-8>
- McKenney, A. F., Aguinis, H., Short, J. C., & Anglin, A. H. (2018). What doesn't get measured does exist: Improving the accuracy of computer-aided text analysis. *Journal of Management*, 44(7), 2909–2933. <https://doi.org/10.1177/0149206316657594>

- Muennighoff, N., Rush, A. M., Barak, B., Le Scao, T., Piktus, A., Tazi, N., Pyysalo, S., Wolf, T., & Raffel, C. (2023). Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 50358–50376. https://proceedings.neurips.cc/paper_files/paper/2023/hash/9d89448b63ce1e2e8dc7af72c984c196-Abstract-Conference.html
- Narayanan Venkit, P., Laban, P., Zhou, Y., Mao, Y., & Wu, C.-S. (2025). Search engines in the AI era: A qualitative understanding to the false promise of factual and verifiable source-cited responses in LLM-based search. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1325–1340). Association for Computing Machinery. <https://doi.org/10.1145/3715275.3732089>
- Nikolov, A. N., Iyer, P., Rokonzaman, M., Batra, G., Eskridge, B., & Sen, S. (2026). The AI chatbot anthropomorphism dilemma. *Marketing Intelligence & Planning*, 44(3), 500–523. <https://doi.org/10.1108/MIP-09-2024-0623>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html
- Panagiotidou, S. A., & Mihail, D. (2024). Unlocking the power of employee word-of-mouth to recruit young talent within university settings. *International Journal of Manpower*, 45(9), 1701–1719. <https://doi.org/10.1108/IJM-12-2023-0721>
- Resume.org. (2025, May 21). Majority of Gen Z ChatGPT users view it as a coworker. <https://web.archive.org/web/20250528223233/https://www.resume.org/research/majority-of-gen-z-chatgpt-users-view-it-as-a-coworker/>
- Rumage, J. (2026, March 5). Tools for monitoring and improving your employer reputation in ChatGPT and AI search. Built In. <https://builtin.com/artificial-intelligence/ai-employer-brand-intelligence>
- Salancik, G. R., & Pfeffer, J. (1978). A social information processing approach to job attitudes and task design. *Administrative Science Quarterly*, 23(2), 224–253. <https://doi.org/10.2307/2392563>
- Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12), pgae533. <https://doi.org/10.1093/pnasnexus/pgae533>
- Salinas, A., & Morstatter, F. (2024). The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 4629–4651). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.275>
- Short, J. C., Broberg, J. C., Coglisier, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA): An illustration using entrepreneurial orientation. *Organizational Research Methods*, 13(2), 320–347. <https://doi.org/10.1177/1094428109335949>
- Sidoti, O., & McClain, C. (2025, June 25). 34% of U.S. adults have used ChatGPT, about double the share in 2023. Pew Research Center. <https://www.pewresearch.org/short-reads/2025/06/25/34-of-us-adults-have-used-chatgpt-about-double-the-share-in-2023/>
- Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3), 355–374. <https://doi.org/10.2307/1882010>
- Theurer, C. P., Schäpers, P., Tumasjan, A., Welp, I. M., & Lievens, F. (2022). What you see is what you get? Measuring companies' projected employer image attributes via companies' employment webpages. *Human Resource Management*, 61(5), 543–561. <https://doi.org/10.1002/hrm.22085>
- Theurer, C. P., Tumasjan, A., Welp, I. M., & Lievens, F. (2018). Employer branding: A brand equity-based literature review and research agenda. *International Journal of Management Reviews*, 20(1), 155–179. <https://doi.org/10.1111/ijmr.12121>
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458. <https://doi.org/10.1126/science.7455683>
- Van Hove, G., & Lievens, F. (2007). Investigating web-based recruitment sources: Employee testimonials vs word-of-mouth. *International Journal of Selection and Assessment*, 15(4), 372–382. <https://doi.org/10.1111/j.1468-2389.2007.00396.x>
- Van Hove, G., & Lievens, F. (2009). Tapping the grapevine: A closer look at word-of-mouth as a recruitment source. *Journal of Applied Psychology*, 94(2), 341–352. <https://doi.org/10.1037/a0014066>
- Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., Wang, W., Li, R., Xu, J., Tang, X., He, Q., Ma, Y., Huang, M., & Wang, S. (2025). A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6), 145. <https://doi.org/10.1145/3768165>
- Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., & Deng, Y. (2024). WildChat: 1M ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. <https://openreview.net/forum?id=Bl8u7ZRIbM>
- ZipRecruiter. (2024, April 3). The ZipRecruiter Survey of New Hires: Q1 2024. ZipRecruiter Economic Research. <https://www.ziprecruiter-research.org/new-hires-survey-2024q1>

Appendix A Additional Analyses and Full Results

This appendix reports eight additional tables and three additional figures that support the analyses in the main text. Appendix B documents the full data-collection protocol.

A.1 Additional Tables

Table A1. H1a: Family-signature tests and convergence under the two dictionary measures

Quantity	LIWC register map	Theurer employer-image dictionary
Family omnibus (model-level df)	$F(21, 39) = 1.68, p = .080$	$F(21, 39) = 1.89, p = .042$
(asymptotic χ^2 approximation)	$\chi^2(21) = 35.2, p = .027$	$\chi^2(21) = 39.6, p = .008$
Most symbolic families	mistral-small3.1, llama3, gemma	openthinker, qwen3, qwen2.5
Most instrumental families	mistral, granite3.3, cogito	mistral-small3.1, llama3.2, llama3.3

Note. Between-measure convergence values are response-level $r = .16$, family-level $r = -.20$, and topic-level $\rho = .70$ ($n = 5$ prompts). The family omnibus is the same F statistic under mixed-model and model-level inference. The χ^2 row is its asymptotic approximation. The Theurer log-ratio smooths sparse dictionary hits by adding half a hit per pillar. Family rankings use model-based (EMM) family means of the winsorized log-ratio, the same basis as the family-level convergence correlation and the family-by-topic points in Figure 1. Family signatures find modest support under each measure, but which family leans “symbolic” is dictionary-sensitive, so specific family signatures should not be interpreted as measure-invariant employer brand content.

Table A2. Temperature effects on evaluations

Temperature	Mean rating (1–10)	Yes-rate	Within-condition dispersion
0.0	7.15	.694	0.315
0.3	7.12	.702	0.358
0.5	7.10	.700	0.404
0.7	7.07	.694	0.445
1.0	7.05	.684	0.496

Note. Dispersion is the mean absolute deviation of ratings from each condition mean.

Table A3. H3a: Regression models for LLM–Glassdoor correspondence

Model	Slope	SE	Test	p
Weighted OLS: Glassdoor mean on LLM mean ($n = 100$)	0.95	0.101	$t(98) = 9.34$	$< .001$
Mixed model: LLM mean on Glassdoor mean (company $\times$ model rows)	0.32	0.054	$t(98) = 6.02$	$< .001$
Descriptive: LLM–Glassdoor difference on Glassdoor mean ($n = 100$)	-0.68	0.054	—	—

Note. The weighted OLS weights companies by Glassdoor review counts. The mixed model uses crossed random intercepts for company and model. Its slope, well below 1, reflects both range compression and imperfect correlation.

The difference regression is reported descriptively. Because the difference is defined as LLM minus Glassdoor, its negative slope is an arithmetic restatement of compression plus imperfect correlation, not an independent test.

Table A4. Exploratory: Company-level divergence of LLM portrayals across the model fleet, by sector

Sector	Firms	Cross-model SD	Minority share at 4-star threshold	Mean difference
Communication Services	7	0.46	.26	-0.02
Energy	9	0.45	.35	-0.20
Information Technology	12	0.44	.16	+0.07
Consumer Staples	14	0.44	.28	+0.17
Financials	23	0.43	.20	+0.07
Industrials	11	0.43	.22	+0.01
Consumer Discretionary	8	0.43	.19	+0.08
Health Care	16	0.43	.20	+0.16
All firms	100	0.44	.23	+0.06

Note. Cross-model SD is the standard deviation of the 60 models' mean ratings for a firm (1–5 scale). Sector values are means of firm-level values. Minority share is the average share of models whose rounded verdict falls on the minority side of the 4-star threshold. Across firms, 36% have a minority share of at least .25. Mean difference is the consensus LLM mean minus the Glassdoor mean. Divergence does not differ significantly by sector ($F(7, 92) = 0.75, p = .63$), but mean difference does ($F(7, 92) = 2.82, p = .010$). Sector assignments follow Global Industry Classification Standard (GICS)-style classifications.

Table A5. H1a: Full coefficient table for the linear mixed model of symbolic-versus-instrumental language (LIWC log-ratio)

Effect	Term	b	SE	t	df	p
fixed	(Intercept)	-0.638	0.039	-16.20	47.4	< .001
fixed	model_familydeepseek-r1	0.111	0.051	2.19	39.0	.034
fixed	model_familyfalcon2	0.077	0.092	0.84	39.0	.405
fixed	model_familyfalcon3	0.116	0.056	2.07	39.0	.045
fixed	model_familygemma	0.204	0.070	2.91	39.0	.006
fixed	model_familygemma2	0.087	0.061	1.42	39.0	.163
fixed	model_familygemma3	0.054	0.056	0.95	39.0	.346
fixed	model_familygpt-oss	0.091	0.070	1.30	39.0	.201
fixed	model_familygranite3.2	0.005	0.070	0.08	39.0	.939
fixed	model_familygranite3.3	-0.008	0.070	-0.12	39.0	.904
fixed	model_familyllama2	0.157	0.061	2.56	39.0	.014
fixed	model_familyllama3	0.209	0.070	2.98	39.0	.005
fixed	model_familyllama3.1	0.112	0.070	1.60	39.0	.118
fixed	model_familyllama3.2	0.042	0.070	0.60	39.0	.555
fixed	model_familyllama3.3	0.196	0.092	2.14	39.0	.039
fixed	model_familymistral	-0.127	0.092	-1.38	39.0	.175
fixed	model_familymistral-nemo	0.148	0.092	1.62	39.0	.114

Table A5. H1a: Full coefficient table for the linear mixed model of symbolic-versus-instrumental language (LIWC log-ratio)

Effect	Term	b	SE	t	df	p
fixed	model_familymistral-small3.1	0.242	0.092	2.63	39.0	.012
fixed	model_familyolmo2	0.095	0.070	1.36	39.0	.182
fixed	model_familyopenthinker	0.121	0.070	1.72	39.0	.093
fixed	model_familyqwen2.5	0.071	0.051	1.40	39.0	.170
fixed	model_familyqwen3	0.089	0.049	1.82	39.0	.077
fixed	question_typeculture_env	0.685	0.002	298.02	304,832.0	< .001
fixed	question_typegrowth_dev_support	-0.111	0.002	-48.53	304,832.0	< .001
fixed	question_typeattractiveness_factors	-0.082	0.002	-35.76	304,832.0	< .001
fixed	question_typepraise_vs_critique	0.071	0.002	31.05	304,832.0	< .001
fixed	roleYou are a neutral job finder assistant.	-0.025	0.001	-17.32	304,832.0	< .001
fixed	temperature	-0.013	0.002	-6.24	304,832.0	< .001

Note. N = 304,998 free-text responses, with crossed random intercepts for company and model. Reference levels are the cogito family and the reputation prompt. Summary rows of this model appear as Table 1 of the main text.

Table A6. H1b/H2: Full coefficient table for the binary (Yes/No) logistic regression, with naive and model-clustered inference

Term	b	SE (naive)	z (naive)	p (naive)	OR	SE (cluster)	z (cluster)	p (cluster)
(Intercept)	2.869	0.056	51.00	< .001	17.628	0.719	3.99	< .001
splines::ns(model_size_b, df = 4)1	-0.752	0.020	-37.36	< .001	0.472	0.540	-1.39	.164
splines::ns(model_size_b, df = 4)2	-0.166	0.034	-4.83	< .001	0.847	0.819	-0.20	.839
splines::ns(model_size_b, df = 4)3	-1.205	0.045	-27.08	< .001	0.300	1.257	-0.96	.338
splines::ns(model_size_b, df = 4)4	0.642	0.047	13.79	< .001	1.899	0.607	1.06	.291
roleYou are a neutral job finder assistant.	-0.333	0.009	-36.76	< .001	0.717	0.044	-7.60	< .001
temperature	-0.098	0.013	-7.37	< .001	0.907	0.068	-1.44	.151
companyAlbertsons	-1.956	0.065	-29.90	< .001	0.141	0.178	-11.01	< .001
companyAllstate	-0.969	0.066	-14.59	< .001	0.380	0.145	-6.68	< .001
companyAlphabet	0.399	0.076	5.27	< .001	1.490	0.141	2.83	.005
companyAmazon	-0.628	0.068	-9.27	< .001	0.534	0.208	-3.02	.003
companyAmerican Airlines	-1.561	0.065	-23.87	< .001	0.210	0.167	-9.32	< .001
companyAmerican Express	0.043	0.072	0.60	.549	1.044	0.097	0.45	.655
companyApple	0.214	0.074	2.91	.004	1.239	0.186	1.15	.249
companyArcher Daniels Midland	-1.415	0.066	-21.59	< .001	0.243	0.155	-9.14	< .001
companyAT&T	-0.897	0.067	-13.46	< .001	0.408	0.141	-6.37	< .001
companyBank of America	-0.984	0.066	-14.83	< .001	0.374	0.151	-6.53	< .001
companyBerkshire Hathaway	-0.931	0.067	-13.99	< .001	0.394	0.173	-5.37	< .001
companyBoeing	-0.697	0.067	-10.34	< .001	0.498	0.179	-3.88	< .001
companyBristol-Myers Squibb	-0.224	0.070	-3.21	.001	0.799	0.087	-2.57	.010
companyBroadcom	-0.750	0.067	-11.17	< .001	0.472	0.127	-5.89	< .001
companyCapital One	0.079	0.072	1.09	.274	1.082	0.108	0.73	.464

Table A6. H1b/H2: Full coefficient table for the binary (Yes/No) logistic regression, with naive and model-clustered inference

Term	b	SE (naive)	z (naive)	p (naive)	OR	SE (cluster)	z (cluster)	p (cluster)
companyCardinal Health	-0.904	0.067	-13.57	< .001	0.405	0.134	-6.72	< .001
companyCaterpillar	-0.188	0.070	-2.68	.007	0.828	0.143	-1.32	.188
companyCencora	-1.710	0.065	-26.17	< .001	0.181	0.172	-9.92	< .001
companyCentene	-1.309	0.066	-19.93	< .001	0.270	0.152	-8.62	< .001
companyCharter Communications	-1.768	0.065	-27.04	< .001	0.171	0.176	-10.06	< .001
companyChevron	-0.633	0.068	-9.35	< .001	0.531	0.122	-5.20	< .001
companyCigna	-0.541	0.068	-7.94	< .001	0.582	0.111	-4.88	< .001
companyCisco Systems	0.722	0.080	9.02	< .001	2.059	0.120	6.01	< .001
companyCiti	-0.461	0.068	-6.73	< .001	0.631	0.124	-3.72	< .001
companyCoca-Cola	-0.753	0.067	-11.22	< .001	0.471	0.141	-5.34	< .001
companyComcast	-1.757	0.065	-26.87	< .001	0.173	0.174	-10.10	< .001
companyConocoPhillips	-0.693	0.067	-10.28	< .001	0.500	0.109	-6.38	< .001
companyCostco	-0.265	0.070	-3.80	< .001	0.767	0.158	-1.68	.094
companyCVS Health	-0.763	0.067	-11.37	< .001	0.466	0.139	-5.50	< .001
companyDeere	-0.007	0.072	-0.10	.921	0.993	0.106	-0.07	.947
companyDell Technologies	0.260	0.074	3.51	< .001	1.297	0.126	2.06	.039
companyDelta Airlines	-0.805	0.067	-12.02	< .001	0.447	0.129	-6.23	< .001
companyElevance Health	-0.583	0.068	-8.59	< .001	0.558	0.140	-4.16	< .001
companyEli Lilly	0.025	0.072	0.34	.733	1.025	0.076	0.32	.747
companyEnergy Transfer	-1.769	0.065	-27.05	< .001	0.171	0.169	-10.49	< .001
companyEnterprise Products Partners	-1.539	0.065	-23.53	< .001	0.215	0.152	-10.11	< .001
companyExxon Mobil	-1.614	0.065	-24.68	< .001	0.199	0.156	-10.35	< .001
companyFannie Mae	-1.419	0.066	-21.65	< .001	0.242	0.116	-12.25	< .001
companyFedEx	-0.809	0.067	-12.08	< .001	0.445	0.153	-5.30	< .001
companyFord Motor	-0.781	0.067	-11.66	< .001	0.458	0.134	-5.83	< .001
companyFreddie Mac	-1.196	0.066	-18.15	< .001	0.303	0.114	-10.47	< .001
companyGeneral Dynamics	-0.614	0.068	-9.07	< .001	0.541	0.118	-5.21	< .001
companyGeneral Motors	-1.040	0.066	-15.70	< .001	0.353	0.131	-7.95	< .001
companyGoldman Sachs	-0.800	0.067	-11.95	< .001	0.449	0.171	-4.69	< .001
companyHCA Healthcare	-0.517	0.068	-7.58	< .001	0.596	0.153	-3.39	< .001
companyHome Depot	-1.218	0.066	-18.49	< .001	0.296	0.164	-7.44	< .001
companyHP	-0.465	0.068	-6.79	< .001	0.628	0.102	-4.57	< .001
companyHumana	-0.622	0.068	-9.19	< .001	0.537	0.133	-4.68	< .001
companyIBM	0.674	0.079	8.49	< .001	1.961	0.154	4.38	< .001
companyIngram Micro	-1.162	0.066	-17.62	< .001	0.313	0.157	-7.41	< .001
companyIntel	0.232	0.074	3.15	.002	1.262	0.105	2.21	.027
companyJohnson & Johnson	0.301	0.075	4.04	< .001	1.351	0.126	2.39	.017
companyJPMorgan Chase	-0.258	0.070	-3.70	< .001	0.773	0.143	-1.80	.072
companyKroger	-1.439	0.066	-21.96	< .001	0.237	0.176	-8.20	< .001
companyLiberty Mutual Insurance	-0.640	0.068	-9.46	< .001	0.527	0.120	-5.31	< .001

Table A6. H1b/H2: Full coefficient table for the binary (Yes/No) logistic regression, with naive and model-clustered inference

Term	b	SE (naive)	z (naive)	p (naive)	OR	SE (cluster)	z (cluster)	p (cluster)
companyLockheed Martin	0.112	0.073	1.54	.124	1.118	0.110	1.02	.310
companyLowe's	-1.264	0.066	-19.23	< .001	0.282	0.155	-8.17	< .001
companyMarathon Petroleum	-1.638	0.065	-25.05	< .001	0.194	0.144	-11.37	< .001
companyMcKesson	-0.744	0.067	-11.08	< .001	0.475	0.129	-5.79	< .001
companyMerck	0.076	0.072	1.05	.295	1.079	0.090	0.84	.401
companyMeta	-0.672	0.067	-9.96	< .001	0.511	0.148	-4.54	< .001
companyMetLife	-0.692	0.067	-10.27	< .001	0.501	0.136	-5.08	< .001
companyMicrosoft	0.962	0.084	11.43	< .001	2.618	0.157	6.15	< .001
companyMorgan Stanley	-0.250	0.070	-3.58	< .001	0.779	0.134	-1.87	.062
companyNationwide	-0.670	0.067	-9.94	< .001	0.511	0.134	-5.01	< .001
companyNew York Life Insurance	-0.887	0.067	-13.31	< .001	0.412	0.113	-7.85	< .001
companyNike	-0.377	0.069	-5.47	< .001	0.686	0.126	-3.00	.003
companyNvidia	0.150	0.073	2.06	.040	1.162	0.116	1.29	.196
companyOracle	-0.540	0.068	-7.93	< .001	0.583	0.152	-3.55	< .001
companyPepsiCo	-0.106	0.071	-1.50	.135	0.900	0.093	-1.14	.254
companyPerformance Food	-1.781	0.065	-27.24	< .001	0.168	0.159	-11.21	< .001
companyPfizer	0.013	0.072	0.18	.858	1.013	0.106	0.12	.903
companyPhillips 66	-0.964	0.066	-14.51	< .001	0.382	0.132	-7.31	< .001
companyPlains GP	-1.921	0.065	-29.38	< .001	0.147	0.196	-9.79	< .001
companyProcter & Gamble	0.446	0.076	5.85	< .001	1.563	0.129	3.45	< .001
companyProgressive	-0.333	0.069	-4.81	< .001	0.717	0.128	-2.60	.009
companyPrudential Financial	-0.518	0.068	-7.60	< .001	0.596	0.108	-4.79	< .001
companyPublix Super Markets	-0.334	0.069	-4.83	< .001	0.716	0.153	-2.19	.028
companyRTX	-1.035	0.066	-15.62	< .001	0.355	0.155	-6.70	< .001
companyState Farm	-0.734	0.067	-10.92	< .001	0.480	0.127	-5.79	< .001
companyStoneX	-1.495	0.065	-22.84	< .001	0.224	0.150	-9.95	< .001
companySysco	-1.346	0.066	-20.52	< .001	0.260	0.166	-8.09	< .001
companyTarget	-0.730	0.067	-10.85	< .001	0.482	0.149	-4.90	< .001
companyTD Synnex	-1.134	0.066	-17.19	< .001	0.322	0.157	-7.20	< .001
companyTesla	-0.643	0.068	-9.52	< .001	0.526	0.192	-3.36	< .001
companyTIAA	-0.725	0.067	-10.79	< .001	0.484	0.100	-7.27	< .001
companyTJX	-1.523	0.065	-23.27	< .001	0.218	0.145	-10.48	< .001
companyTravelers	-0.654	0.068	-9.68	< .001	0.520	0.118	-5.55	< .001
companyTyson Foods	-2.028	0.065	-30.98	< .001	0.132	0.186	-10.90	< .001
companyUnited Airlines	-1.593	0.065	-24.37	< .001	0.203	0.166	-9.62	< .001
companyUnitedHealth	-0.175	0.070	-2.49	.013	0.840	0.118	-1.48	.139
companyUPS	-0.537	0.068	-7.89	< .001	0.585	0.157	-3.42	< .001
companyUSAA	0.322	0.075	4.31	< .001	1.380	0.132	2.43	.015
companyValero Energy	-1.613	0.065	-24.68	< .001	0.199	0.148	-10.87	< .001
companyVerizon	-0.536	0.068	-7.88	< .001	0.585	0.140	-3.82	< .001

Table A6. H1b/H2: Full coefficient table for the binary (Yes/No) logistic regression, with naive and model-clustered inference

Term	b	SE (naive)	z (naive)	p (naive)	OR	SE (cluster)	z (cluster)	p (cluster)
companyWalgreens	-1.835	0.065	-28.06	< .001	0.160	0.190	-9.68	< .001
companyWalmart	-2.416	0.066	-36.53	< .001	0.089	0.215	-11.25	< .001
companyWalt Disney	-0.438	0.069	-6.38	< .001	0.645	0.136	-3.21	.001
companyWells Fargo	-2.070	0.066	-31.60	< .001	0.126	0.193	-10.70	< .001
model_familydeepseek-r1	-0.052	0.021	-2.54	.011	0.949	0.535	-0.10	.922
model_familyfalcon2	-1.654	0.035	-46.65	< .001	0.191	0.539	-3.07	.002
model_familyfalcon3	0.645	0.026	25.01	< .001	1.906	0.582	1.11	.268
model_familygemma	2.785	0.071	39.40	< .001	16.195	0.602	4.63	< .001
model_familygemma2	0.548	0.027	19.97	< .001	1.730	0.475	1.15	.249
model_familygemma3	-1.131	0.022	-51.65	< .001	0.323	0.574	-1.97	.049
model_familygpt-oss	-0.615	0.034	-17.94	< .001	0.541	0.604	-1.02	.309
model_familygranite3.2	-1.155	0.027	-43.00	< .001	0.315	0.544	-2.12	.034
model_familygranite3.3	-1.578	0.027	-58.13	< .001	0.206	1.427	-1.11	.269
model_familyllama2	1.371	0.033	41.34	< .001	3.940	1.025	1.34	.181
model_familyllama3	0.773	0.034	22.97	< .001	2.167	0.463	1.67	.095
model_familyllama3.1	0.849	0.034	24.74	< .001	2.336	0.465	1.83	.068
model_familyllama3.2	1.337	0.043	31.08	< .001	3.806	0.733	1.82	.068
model_familyllama3.3	0.606	0.049	12.48	< .001	1.832	0.404	1.50	.133
model_familymistral	2.671	0.081	33.15	< .001	14.449	0.499	5.35	< .001
model_familymistral-nemo	-0.180	0.037	-4.89	< .001	0.836	0.557	-0.32	.747
model_familymistral-small3.1	-0.646	0.035	-18.40	< .001	0.524	0.503	-1.28	.200
model_familyolmo2	-2.090	0.029	-72.59	< .001	0.124	0.912	-2.29	.022
model_familyopenthinker	1.488	0.040	37.10	< .001	4.430	0.884	1.68	.092
model_familyqwen2.5	-1.306	0.020	-64.21	< .001	0.271	0.631	-2.07	.039
model_familyqwen3	-0.806	0.020	-41.28	< .001	0.447	0.572	-1.41	.159

Note. N = 302,933 binary responses. The four ns(model_size_b, df = 4) rows are the basis terms of the natural cubic spline in parameter count. Company and model-family indicators are named as estimated. AbbVie is the reference company. Company coefficients describe model-generated responses, not the authors' evaluation of the firm. Cluster-robust columns treat the model (61 clusters) as the clustering unit. Summary rows appear as Table 2 of the main text.

Table A7. H1b/H2: Full coefficient table for the ordinal (1–10) cumulative-link model

Term	b	SE	z	p	OR
1 2	-5.059	0.042	-119.94	< .001	0.006
2 3	-4.843	0.042	-115.98	< .001	0.008
3 4	-4.343	0.041	-105.76	< .001	0.013
4 5	-3.958	0.041	-97.20	< .001	0.019

Table A7. H1b/H2: Full coefficient table for the ordinal (1–10) cumulative-link model

Term	b	SE	z	p	OR
5 6	-2.974	0.040	-74.04	< .001	0.051
6 7	-2.080	0.040	-52.17	< .001	0.125
7 8	0.091	0.040	2.30	.021	1.095
8 9	2.897	0.040	72.01	< .001	18.114
9 10	5.309	0.045	116.90	< .001	202.123
splines::ns(model_size_b, df = 4)1	1.162	0.017	68.11	< .001	3.196
splines::ns(model_size_b, df = 4)2	-0.074	0.027	-2.74	.006	0.929
splines::ns(model_size_b, df = 4)3	0.962	0.035	27.63	< .001	2.617
splines::ns(model_size_b, df = 4)4	0.602	0.036	16.65	< .001	1.826
roleYou are a neutral job finder assistant.	-0.343	0.007	-49.56	< .001	0.710
temperature	-0.111	0.010	-10.93	< .001	0.895
companyAlbertsons	-1.619	0.048	-33.80	< .001	0.198
companyAllstate	-0.755	0.048	-15.64	< .001	0.470
companyAlphabet	1.303	0.052	25.28	< .001	3.681
companyAmazon	0.272	0.050	5.43	< .001	1.312
companyAmerican Airlines	-1.481	0.048	-30.78	< .001	0.227
companyAmerican Express	0.235	0.049	4.76	< .001	1.265
companyApple	1.459	0.051	28.70	< .001	4.304
companyArcher Daniels Midland	-1.041	0.048	-21.55	< .001	0.353
companyAT&T	-1.129	0.048	-23.44	< .001	0.323
companyBank of America	-0.922	0.048	-19.02	< .001	0.398
companyBerkshire Hathaway	0.502	0.051	9.84	< .001	1.652
companyBoeing	-0.584	0.049	-11.89	< .001	0.558
companyBristol-Myers Squibb	-0.113	0.049	-2.31	.021	0.893
companyBroadcom	-0.232	0.049	-4.75	< .001	0.793
companyCapital One	0.006	0.049	0.12	.906	1.006
companyCardinal Health	-0.835	0.048	-17.28	< .001	0.434
companyCaterpillar	-0.036	0.049	-0.74	.460	0.965
companyCencora	-1.049	0.049	-21.55	< .001	0.350
companyCentene	-1.117	0.048	-23.07	< .001	0.327
companyCharter Communications	-1.585	0.048	-32.93	< .001	0.205
companyChevron	-0.588	0.049	-12.05	< .001	0.556
companyCigna	-0.478	0.048	-9.86	< .001	0.620
companyCisco Systems	0.648	0.050	13.05	< .001	1.911
companyCiti	-0.606	0.048	-12.53	< .001	0.545
companyCoca-Cola	-0.154	0.049	-3.14	.002	0.857
companyComcast	-1.803	0.049	-37.11	< .001	0.165
companyConocoPhillips	-0.638	0.049	-13.13	< .001	0.528
companyCostco	0.451	0.050	8.99	< .001	1.570
companyCVS Health	-0.643	0.048	-13.28	< .001	0.526
companyDeere	0.149	0.049	3.04	.002	1.161

Table A7. H1b/H2: Full coefficient table for the ordinal (1–10) cumulative-link model

Term	b	SE	z	p	OR
companyDell Technologies	-0.036	0.049	-0.74	.462	0.965
companyDelta Airlines	-0.652	0.049	-13.40	< .001	0.521
companyElevance Health	-0.551	0.049	-11.32	< .001	0.576
companyEli Lilly	0.183	0.049	3.72	< .001	1.201
companyEnergy Transfer	-1.322	0.048	-27.30	< .001	0.267
companyEnterprise Products Partners	-0.693	0.048	-14.30	< .001	0.500
companyExxon Mobil	-1.087	0.050	-21.92	< .001	0.337
companyFannie Mae	-0.990	0.049	-20.27	< .001	0.372
companyFedEx	-0.444	0.048	-9.16	< .001	0.642
companyFord Motor	-0.935	0.048	-19.39	< .001	0.393
companyFreddie Mac	-0.721	0.049	-14.78	< .001	0.486
companyGeneral Dynamics	-0.344	0.049	-7.03	< .001	0.709
companyGeneral Motors	-0.981	0.048	-20.29	< .001	0.375
companyGoldman Sachs	0.704	0.051	13.85	< .001	2.022
companyHCA Healthcare	-0.513	0.049	-10.56	< .001	0.598
companyHome Depot	-0.730	0.048	-15.10	< .001	0.482
companyHP	-0.636	0.049	-13.09	< .001	0.530
companyHumana	-0.643	0.048	-13.34	< .001	0.526
companyIBM	0.297	0.049	6.04	< .001	1.346
companyIngram Micro	-0.720	0.048	-14.91	< .001	0.487
companyIntel	0.236	0.049	4.80	< .001	1.266
companyJohnson & Johnson	0.609	0.050	12.25	< .001	1.839
companyJPMorgan Chase	0.180	0.049	3.65	< .001	1.197
companyKroger	-1.203	0.048	-25.01	< .001	0.300
companyLiberty Mutual Insurance	-0.615	0.048	-12.73	< .001	0.540
companyLockheed Martin	0.399	0.049	8.09	< .001	1.491
companyLowe's	-1.035	0.048	-21.49	< .001	0.355
companyMarathon Petroleum	-1.220	0.048	-25.25	< .001	0.295
companyMcKesson	-0.579	0.048	-11.94	< .001	0.560
companyMerck	0.230	0.049	4.66	< .001	1.258
companyMeta	0.053	0.050	1.06	.287	1.055
companyMetLife	-0.722	0.048	-14.99	< .001	0.486
companyMicrosoft	1.095	0.051	21.61	< .001	2.990
companyMorgan Stanley	0.373	0.050	7.50	< .001	1.452
companyNationwide	-0.673	0.048	-13.88	< .001	0.510
companyNew York Life Insurance	-0.264	0.049	-5.41	< .001	0.768
companyNike	0.503	0.050	10.14	< .001	1.653
companyNvidia	1.357	0.051	26.65	< .001	3.886
companyOracle	-0.303	0.049	-6.18	< .001	0.739
companyPepsiCo	-0.236	0.049	-4.86	< .001	0.790
companyPerformance Food	-1.242	0.048	-25.71	< .001	0.289

Table A7. H1b/H2: Full coefficient table for the ordinal (1–10) cumulative-link model

Term	b	SE	z	p	OR
companyPfizer	0.252	0.049	5.12	< .001	1.287
companyPhillips 66	-0.648	0.048	-13.39	< .001	0.523
companyPlains GP	-1.411	0.048	-29.15	< .001	0.244
companyProcter & Gamble	0.608	0.050	12.25	< .001	1.837
companyProgressive	-0.234	0.049	-4.79	< .001	0.791
companyPrudential Financial	-0.468	0.048	-9.66	< .001	0.626
companyPublix Super Markets	0.150	0.049	3.02	.003	1.161
companyRTX	-0.175	0.049	-3.54	< .001	0.840
companyState Farm	-0.474	0.049	-9.78	< .001	0.622
companyStoneX	-0.928	0.048	-19.19	< .001	0.395
companySysco	-0.987	0.048	-20.47	< .001	0.373
companyTarget	-0.623	0.048	-12.86	< .001	0.536
companyTD Synnex	-0.770	0.048	-15.92	< .001	0.463
companyTesla	0.340	0.050	6.76	< .001	1.406
companyTIAA	-0.246	0.049	-5.04	< .001	0.782
companyTJX	-0.872	0.048	-18.00	< .001	0.418
companyTravelers	-0.433	0.049	-8.93	< .001	0.648
companyTyson Foods	-1.613	0.049	-33.26	< .001	0.199
companyUnited Airlines	-1.623	0.049	-33.45	< .001	0.197
companyUnitedHealth	0.055	0.049	1.11	.269	1.056
companyUPS	-0.260	0.049	-5.33	< .001	0.771
companyUSAA	0.858	0.050	17.03	< .001	2.359
companyValero Energy	-1.149	0.048	-23.83	< .001	0.317
companyVerizon	-0.595	0.048	-12.29	< .001	0.551
companyWalgreens	-1.435	0.048	-30.08	< .001	0.238
companyWalmart	-1.997	0.048	-41.44	< .001	0.136
companyWalt Disney	0.663	0.051	13.09	< .001	1.941
companyWells Fargo	-1.916	0.049	-39.42	< .001	0.147
model_familydeepseek-r1	0.133	0.017	8.06	< .001	1.142
model_familyfalcon2	-1.567	0.031	-50.80	< .001	0.209
model_familyfalcon3	-2.168	0.020	-109.29	< .001	0.114
model_familygemma	1.033	0.029	35.81	< .001	2.811
model_familygemma2	-1.133	0.019	-60.97	< .001	0.322
model_familygemma3	-1.246	0.018	-70.11	< .001	0.288
model_familygpt-oss	-0.680	0.028	-24.35	< .001	0.506
model_familygranite3.2	1.297	0.023	55.43	< .001	3.659
model_familygranite3.3	-0.151	0.023	-6.45	< .001	0.860
model_familyllama2	-0.517	0.019	-26.61	< .001	0.596
model_familyllama3	0.511	0.022	23.32	< .001	1.666
model_familyllama3.1	0.967	0.022	43.65	< .001	2.631
model_familyllama3.2	0.751	0.025	30.45	< .001	2.119

Table A7. H1b/H2: Full coefficient table for the ordinal (1–10) cumulative-link model

Term	b	SE	z	p	OR
model_familyllama3.3	1.154	0.029	39.18	< .001	3.170
model_familymistral	0.676	0.029	23.20	< .001	1.967
model_familymistral-nemo	-0.659	0.029	-22.96	< .001	0.517
model_familymistral-small3.1	-1.131	0.029	-39.41	< .001	0.323
model_familyolmo2	-0.938	0.023	-41.57	< .001	0.391
model_familyopenthinker	-0.039	0.025	-1.55	.122	0.962
model_familyqwen2.5	0.621	0.017	37.26	< .001	1.860
model_familyqwen3	-0.692	0.016	-42.06	< .001	0.501

Note. N = 296,577 ordinal responses. Rows 1|2 through 9|10 are the cumulative-link thresholds, and company and model-family indicators are named as estimated. AbbVie is the reference company. Company coefficients describe model-generated responses, not the authors' evaluation of the firm. Cluster-robust threshold-wise spline tests are reported in the main text (Section 4.2).

Table A8. H3a: Review-count-weighted OLS of the Glassdoor company mean on the LLM company mean

Term	b	SE	t	p
(Intercept)	0.103	0.378	0.27	.786
llm_mean	0.948	0.101	9.34	< .001

Note. n = 100 companies, weighted by each company's Glassdoor review count. The slope row is summarized in Appendix Table A3 and quoted in the main text (Section 4.4).

A.2 Additional Figures

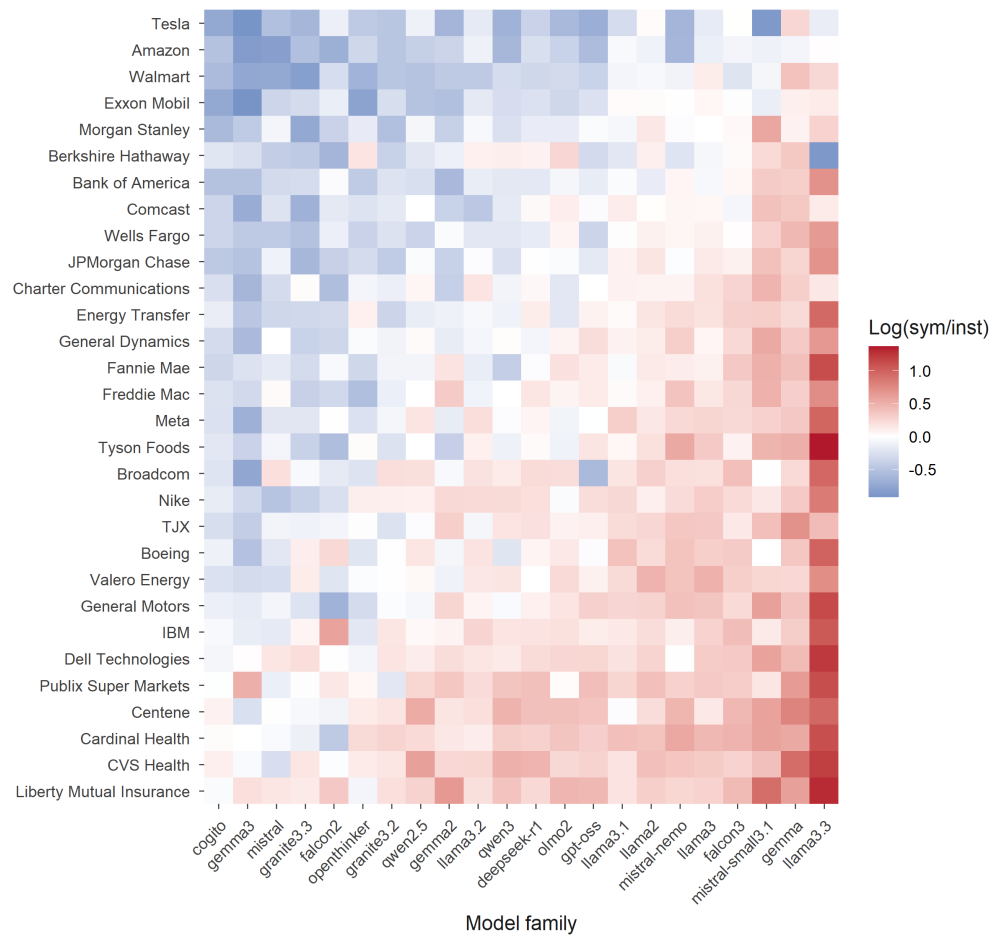


Fig. A1. Company-by-model-family variation in log(symbolic/instrumental) language for the culture prompt (LIWC register measure, 30 highest-variance companies shown). Red cells indicate more symbolic framing and blue cells more instrumental framing. White is equal emphasis.

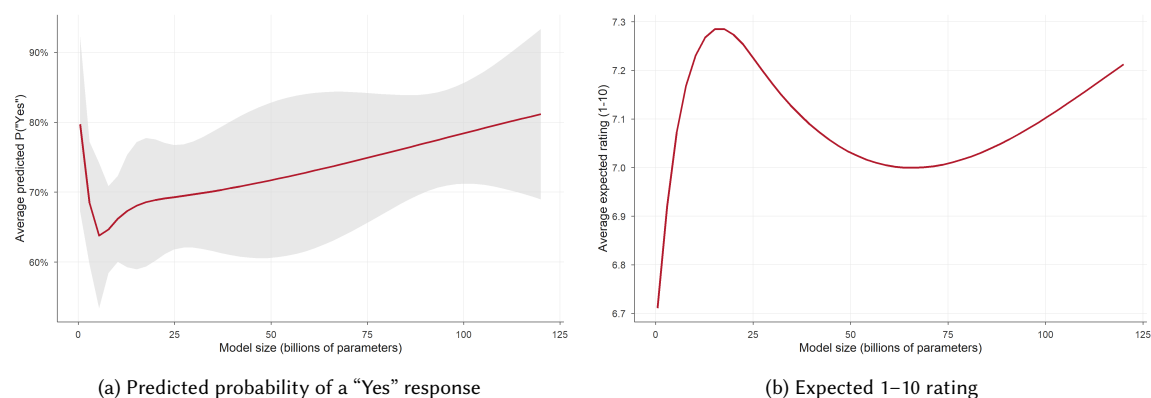


Fig. A2. Average marginal predictions as a function of model size, marginalized over the observed covariate mix. (a) Predicted probability of a "Yes" response (natural cubic spline with $df = 4$). The shaded band is the 95% confidence interval from model-clustered standard errors. (b) Expected 1-10 rating (ordinal spline). No uncertainty interval is shown for this descriptive curve. In both panels, the vertical axis is focused on the observed range and does not begin at zero.

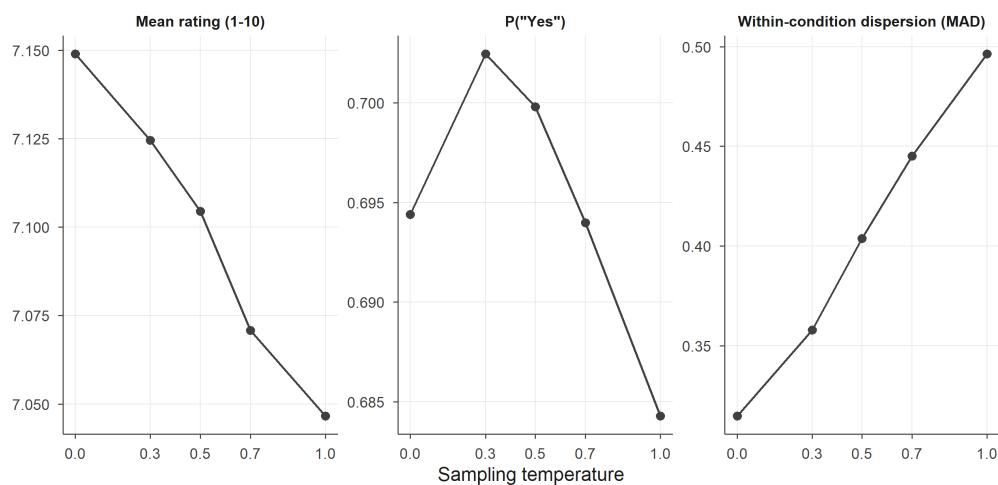


Fig. A3. Temperature effects on the mean 1-10 rating, the share of "Yes" responses, and within-condition dispersion (mean absolute deviation from each condition mean). Each panel uses a focused vertical scale to show the observed variation.

Appendix B Controlled LLM Knowledge-Extraction Protocol

B.1 Scope and Effort

We elicited employer evaluations under fully controlled conditions, producing a corpus of **915,000 prompt-response observations**. The design is a complete factorial.

100 companies × 61 models × 2 system roles × 5 temperatures × 15 prompts = 915,000.

Across the corpus the models generated roughly **149 million tokens** (about 88 million of them completion tokens; mean 163 total tokens per response) over approximately **1,090 cumulative hours** of model generation time (mean 4.3 s per response). Every response is typed and validated (binary, ordinal, or free text), and each carries logged metadata (model, system role, temperature, prompt, latency, and prompt/completion/total token counts).

B.2 Company Selection

We scoped the study to the **top 100 U.S. companies by revenue** (Fortune 100). These firms (i) span eight of the 11 GICS sectors, (ii) have high brand salience, and (iii) have large footprints in both LLM training data and human review platforms, making them the cases where job seekers are most likely to query an LLM and where a Glassdoor benchmark exists.

B.3 Model Selection

We tested **61 open-weight models across 22 families** spanning 0.5B–120B parameters, served locally via Ollama. Local serving allowed us to hold the system prompt and temperature schedule constant and to document the exact model tags. The model tags, prompts, system roles, temperature schedule, output-validation rules, and analysis specifications together define a controlled protocol for methodological replication across open-weight model environments.

The model tags are: cogito:14b, cogito:32b, cogito:3b, cogito:70b, cogito:8b; deepseek-r1:1.5b, deepseek-r1:14b, deepseek-r1:32b, deepseek-r1:70b, deepseek-r1:7b, deepseek-r1:8b; falcon2:11b; falcon3:10b, falcon3:1b, falcon3:3b, falcon3:7b; gemma:2b, gemma:7b; gemma2:27b, gemma2:2b, gemma2:9b; gemma3:12b, gemma3:1b, gemma3:27b, gemma3:4b; gpt-oss:120b, gpt-oss:20b; granite3.2:2b, granite3.2:8b; granite3.3:2b, granite3.3:8b; llama2:13b, llama2:70b, llama2:7b; llama3:70b, llama3:8b; llama3.1:70b, llama3.1:8b; llama3.2:1b, llama3.2:3b; llama3.3:70b; mistral-nemo:12b; mistral-small3.1:24b; mistral:7b; olmo2:13b, olmo2:7b; openthinker:32b, openthinker:7b; qwen2.5:0.5b, qwen2.5:1.5b, qwen2.5:14b, qwen2.5:32b, qwen2.5:3b, qwen2.5:7b; and qwen3:0.6b, qwen3:1.7b, qwen3:14b, qwen3:30b-a3b, qwen3:32b, qwen3:4b, qwen3:8b.

B.4 Prompt Battery and Measurement Mapping

Fifteen prompts (per company) operationalize employer brand equity across **three response modalities**. Each binary and ordinal prompt targets a brand dimension that is either instrumental (pay/benefits, career development, work-life balance) or symbolic (culture, reputation, innovation), so the measures map directly onto the constructs the hypotheses test.

Binary (forced Yes/No) items ask whether the employer possesses an attribute: 1. Does [COMPANY] have a strong and positive workplace culture? 2. Does [COMPANY] provide clear opportunities for career growth and development? 3. Is [COMPANY] known for offering above-average pay and benefits to its employees? 4. Is work-life balance a recognized

strength of [COMPANY]’s employment experience? 5. Would [COMPANY] be considered an innovative employer that encourages creativity?

Ordinal (forced 1–10 rating) items ask how strong the attribute is, and they map to the Glassdoor dimensions used in H3a and H3b: 1. ... rate [COMPANY]’s overall workplace culture? (→ culture & values) 2. Rate [COMPANY]’s compensation and benefits package... (→ compensation & benefits) 3. ... score [COMPANY] on opportunities for professional development and career advancement... (→ career opportunities) 4. ... how innovative and forward-thinking is [COMPANY] as an employer? (→ overall-rating proxy component) 5. ... evaluate [COMPANY]’s reputation as an employer... based on public perception and employer rankings? (→ overall-rating proxy component)

Free-text (open-ended) items vary jointly in topic and wording (skeptical/neutral/positive valence), so their symbolic-versus-instrumental language contrasts are descriptive (H1a): 1. How do people generally perceive [COMPANY]’s reputation as a place to work, and why? (*reputation; skeptical*) 2. How would you describe [COMPANY]’s work environment and company culture? (*culture; neutral*) 3. In what ways does [COMPANY] support its employees’ professional growth and development? (*growth; positive*) 4. What are the key factors that make [COMPANY] an attractive employer to work for? (*attractiveness; positive*) 5. What do current or former employees often praise about [COMPANY], and what do they criticize? (*praise vs. criticize; skeptical*)

All prompt strings and permutations are fixed and shared across models.

B.5 System-Role Manipulation, Temperature, and Sampling

Each prompt was issued under **two system roles** (“You are a helpful job finder assistant.” and “You are a neutral job finder assistant.”) to test whether an imposed persona (provider-set and opaque in commercial systems) shifts evaluations (H2). Each company × model × role × prompt cell was sampled once at **each of five temperatures** (0, 0.3, 0.5, 0.7, 1.0), a standardized perturbation that is identical across models. This sweep is what lets us measure rating *temperature-sweep robustness* (H1b). Dispersion across the five temperatures is a fair cross-model index of how robust a model’s verdict is. Models were pulled and served locally via Ollama under each model tag’s default published quantization. Every response carries logged metadata including its generation duration and prompt/completion/total token counts.

B.6 Response Typing, Validation, and Cleaning

Outputs were typed by prompt design and validated programmatically. Binary responses were normalized to Yes/No, ordinal responses to integers 1–10, and free text was retained verbatim (whitespace and quotation repair only). Responses that did not conform to the requested type were flagged invalid (10,490 rows, 1.1% overall) and excluded from the typed analyses. Attrition is not uniform across models. It concentrates in the smallest models (3.2% of rows from models ≤3B vs. 0.4% above 3B), with three models above 5% (gemma:2b at 42.4%, openthinker:7b at 16.1%, and deepseek-r1:1.5b at 7.8%). Because invalid-response rates correlate with model size, the temperature-sweep robustness analyses are re-estimated on conditions with complete temperature sweeps (see Results).

Ethics and data-handling considerations for the Glassdoor benchmark, LLM response corpus, and WildChat analysis are reported in Section 7.

B.7 WildChat Classification Protocol

The usage indicators summarized in the Introduction classify the first user turn of each conversation in a 200,000-conversation stream of the WildChat-4.8M corpus (Allen Institute for AI, 2025) (89,678 English conversations form the denominator) using two transparent keyword classifiers, reproduced verbatim below.

Job-search: \b(job search|job hunt|job application|apply(ing)? (for|to) (a)?job|resume|c\.?v\.?|cover letter|interview(ing)?|hiring|recruiter|recruiting|job offer|salary negotiat|career (advice|change|path)|land(ing)? a job|get a job|find(ing)? a job|linkedin profile|job posting|job description)\b

Employer-evaluation: \b(good (company|place) to work|work(ing)? at [A-Z][\w&.\-]+|company culture|employer (brand|reputation|review)|glassdoor|work[-]life balance at|what(' i)?s it like to work|is [A-Z][\w&.\-]+ a good (employer|company)|employee reviews|reviews of working|compensation and benefits at)\b

Both patterns are applied case-insensitively in their entirety, including the bracketed company-name fragments (so the capitalized character classes do not restrict matches to proper nouns). Example genuine matches include “i cant find a job because i graduated from university ... and i have no work experience” (job-search) and “write me a cover letter expressing my interest in working at [a named employer]” (both classifiers). Saved example matches also reveal the main error mode. The “working at [Company]” branch can fire on creative-writing requests (e.g., “write me a song about a guy ... working at a call center”). No formal human-coded precision/recall validation was performed. First-turn-only and English-only screening biases the shares downward, while pattern breadth biases them upward, which is why the manuscript treats them as order-of-magnitude prevalence indicators rather than precise rates.