The Skill Content of Job Postings: Evidence from Internal Hiring Decisions

October 2026

Abstract

Job posting text has become an important source of information for research on technological change and human capital, yet its informational content has rarely been tested against the hiring decisions postings are written to support. We link posting text to 1,034 internal application events in the IT division of a major financial institution. We define documented skill fit as the similarity between an applicant's current-role posting and the target vacancy posting, then test whether it is associated with selection among applicants to the same vacancy. Models include vacancy fixed effects, applicant random intercepts, tenure in role, and whether the application is for a higher job grade. A one-standard-deviation difference in documented skill fit is associated with 2.14 times higher odds of selection (95% CI [1.10, 4.17]). The result establishes that richly specified technical postings contain candidate-level signal relevant to selection, while not claiming that posting text captures an employee's complete skill portfolio or that the association is causal. The resulting skill space also allows firms to examine how skill-aligned opportunities are distributed and which close alternatives employees do not pursue.

Acknowledgment: AI writing tools assisted with editing and manuscript preparation. Research design, analysis, findings, and conclusions are the work of the authors, who take full responsibility for all content and style.

1. Introduction

Job vacancy postings have emerged as a valuable source of information in labor economics. They have helped answer questions about how technological change reshapes skill demand (Deming and Kahn, 2018), how employers upgrade what they seek from workers in downturns (Hershbein and Kahn, 2018), and where automation and AI exposure concentrate across occupations (Acemoglu et al., 2022; Eloundou et al., 2024); in each case the key variable was built from posting text aggregated across firms, occupations, or over time.[1] There has been little systematic examination of what skill content the posting text carries, or at what resolution (how finely it distinguishes between specific roles). In this paper we measure the skill content of professional-occupation postings directly: how finely it resolves, and what it makes visible about the workforce.

Cross-firm skill benchmarks, workforce analytics, and research on technological change use posting text as a record of employer demand. Those aggregate inferences depend on the information contained in individual postings, yet prior work has rarely linked that text to the hiring decisions it was written to support. Such a linkage provides a criterion-based calibration: does similarity between the documented requirements of an applicant's current role and a target role distinguish candidates whom the firm evaluates differently? Answering that question requires the origin-role posting, the target vacancy, the competing applicant pool, and the selection outcome. These records are generally unavailable in external job-board data but can be linked inside a firm.

We test this at the IT division of a major U.S. financial services institution. The division employs a large professional workforce across richly specified technical roles, with steady opportunity for internal movement as new systems are built and older ones retired. The firm's HRIS tracks all new job postings; it allows us to link each internal applicant to the job description of the role they currently hold. Consider a software developer working on trading infrastructure applying for a machine learning engineer role: their current role posting specifies the systems and responsibilities the firm documented when that role was created; the target posting specifies what the new role demands. The origin posting is the firm's own description of what their current role requires. Across 1,034 application events over four years, we can trace each candidate's origin posting, observe the full pool that competed for the same vacancy, and follow the outcome for each, a combination that exists only inside the firm: external labor market data records placements but never the pool a placement was drawn from, nor what the selected candidate was already doing.

Skill fit is the cosine similarity between two job postings measured in a skill space. For each application, we compare the posting for the candidate's current role with the posting

[1] The systematic collection of job postings at scale originated in efforts to match individual workers to open roles across firms — an objective not quite met. The data assembled for that purpose has since become the foundation of an industry in general workforce intelligence services and catalyzed this research program.

for the target vacancy. The measure therefore captures documented role-to-role alignment, not the employee's complete latent skill set. We compare applicants to the same vacancy, so vacancy fixed effects hold constant factors shared within a requisition. Because the data contain screened internal applicants, the estimand is deliberately narrow: whether documented skill fit distinguishes candidates within the applicant pool the firm chose to evaluate. Applicant random intercepts model residual heterogeneity in baseline selection propensity.

Among people applying to the identical opening, a one-standard-deviation difference in skill fit is associated with 2.14 times higher odds of selection. The encoder places every posting in a common space, so the score can also be calculated for role pairs that never appeared in the same competition. In this setting, selected applicants generally score higher than nonselected applicants both within vacancies and across application events. The result is consistent with technical postings containing distinctions more specific than job titles alone, such as programming languages, platforms, systems, and domain methods. It does not establish that posting text captures every consideration used in hiring.

The empirical setting is one large financial institution whose technical postings are richly specified. Competitive technical labor markets provide a reason to expect other firms to describe requirements concretely. The study should therefore be read as a within-firm calibration under conditions that may be reproducible elsewhere. Firms with linked postings, application records, and hiring outcomes can conduct the same assessment in their own settings.

We apply the validated skill fit across the firm's full workforce. For 14.8% of applicants, no open role anywhere in the firm resolves as a close skill fit to the position they currently hold. Among those who were turned down and had a comparable vacancy open at the same moment, 54.9% had a closely matching alternative they did not apply to. Reading posting text through a skill space anchored to actual hiring decisions is what brings both patterns into view. Job titles carry no trace of either, and neither do stated preferences.

A technical literature on job recommendation and worker–job matching builds skill-fit measures from posting text and resume content in a similar way (Bian et al., 2020; Qin et al., 2020; Zhu et al., 2018; Yu, Zhang, and Yu, 2024), but the aim is algorithmic: rank candidates for a role, or rank roles for a candidate, and surface who to screen in. The candidate side of that match is typically a self-reported resume, unverified by any employer. They measure success by clicks, callbacks, or interview advancement in a large pool of applicants, not by whether a specific person was hired from a closed competing pool. Actual selection data of that kind does not appear in this literature; external platform and job-board data records placements, not the pools they were drawn from. Guan et al. (2024) find that a single job description taken alone suffers from an information deficit for

reliable recall and compensate by aggregating neighboring postings, a fix that works for recommendation but leaves the prior question open: what does a single posting actually contain? Kokkodis and Ipeirotis (2023) come closest on outcome structure: applicants competing for a single opening with an observed hire. But the predictive signal there is reputation, ratings, bid, prior performance, not the text of the posting. Their contribution is a technical innovation in building a recommender; ours focuses on the skill information in professional job postings, and how finely that information resolves who fits a role.

The posting-text literature uses what employers write as a measure of skill demand. We provide criterion-related evidence for that practice by linking documented role similarity to a consequential outcome: selection among applicants to the same vacancy. The analysis shows that, in this setting, posting-derived similarity corresponds to distinctions the firm makes during selection. It does not establish that postings are complete records of employee capabilities or that skill fit alone explains the hiring decision.

An active strand of literature studies internal worker movement within a firm (Baker, Gibbs, and Holmstrom, 1994a,b; Bidwell, 2011; Bidwell and Keller, 2024; Huitfeldt et al., 2023). These papers ask who advances to which roles, what frictions govern that movement, and whether workers reach positions that match what they bring. Much of this work rests on single-firm studies (Minni, 2026; Bidwell, 2011), as firms are the natural locus of the relevant data. While our focus is on measuring the richness of job postings and secondarily on how firms can exploit that richness for analytics, we share some common elements with this literature in management and economics. Our observation about plausible information frictions shaping internal movement illustrates the rich possibilities open to management and organizational economics once this strand leverages the detailed textual descriptions of employee roles, descriptions that pre-trained language models now make tractable.

This paper makes three contributions. First, it provides a direct within-vacancy test of whether posting-derived role similarity is associated with selection among screened internal applicants. A one-standard-deviation difference in skill fit predicts 2.14 times higher odds of selection among candidates competing for the same vacancy. Second, it shows how firms can construct and evaluate a common skill space from linked postings and selection records already stored in an HRIS. Validation against hiring outcomes provides criterion-related evidence, although it does not remove questions about missing postings, text construction, LLM-generated supervision, or unobserved applicant characteristics. Third, applying the measure to the firm's opportunity landscape reveals two descriptive patterns consistent with information frictions: 14.8 percent of applicants have no close documented skill match among contemporaneous vacancies, and 54.9 percent of unsuccessful applicants with a comparable vacancy available had a closely skill-matched alternative they did not apply to. These patterns motivate further research but do not

identify why employees overlook or avoid particular openings.

The remainder of the paper is organized as follows. Section 2 describes the data and the hiring records. Section 3 builds the skill-fit measure from posting text. Section 4 tests whether skill fit predicts which candidate the firm selects within a competition. Section 5 applies the validated measure across the workforce. Section 6 discusses the scope of the result and its implications.

2. Research Setting

2.1. Organizational Context

The data come from one unit of the Information Technology division of a major U.S.-based financial services institution. The division supports trading platforms, wealth management systems, wholesale payments, risk systems, and trade finance — functions built on different stacks, maintained by different specialists. Application developers here write Java for global markets or COBOL for mainframe risk systems. QA engineers test FX options in one team and wholesale credit in another. Hadoop engineers move pipelines off Teradata while the mainframe team they are pulling from is still staffed. Postings specify the domain, the stack, and the platforms by name. A role summary names three to five specific technologies before it reaches the qualifications section.

Each role sits within one technical specialty. Application programmers build trading platforms in a named language. Data analysts work risk systems through particular database tools. Quality assurance specialists test against fixed frameworks. Individual contributor technical roles here span different technology platforms, but the posting for each is a skill specification, not a general description: required competencies and responsibilities are richly specified in them.

Internal mobility runs through a central job board. Each posting carries a role overview, the responsibilities, and the qualifications, mandatory and preferred. Appendix A reproduces one. Employees apply without prior managerial approval. Recruiting units screen the applicants through human review and run structured interviews; no automated text-matching tool was used in this initial screen. External candidates enter only after a set window, which holds a clear advantage for internal applicants. The data capture every applicant who clears the initial screen; for each vacancy, the HRIS records the full pool and the selection outcome. That pool is already filtered before the text-derived measure is asked to separate anyone.

2.2. Data Structure and Coverage

The analysis draws on the firm's Human Resource Information System (HRIS). The system records every job requisition, application, and selection outcome. Each record holds the

TABLE 1. Job Postings and Internal Mobility Patterns (2018–2021)

Year	Total Positions	Internal Applications (%)	Internal Fills (%)
2018	1,326	40.3	8.3
2019	1,370	50.1	17.3
2020	1,606	75.7	42.5
2021	2,240	62.5	33.9
Total	6,542	58.6	27.3

Note: Data from HRIS covering individual contributor positions posted 2018–2021, supplemented by 1,638 pre-2018 postings. Internal Applications (%): positions receiving at least one internal application. Internal Fills (%): positions ultimately filled by an internal candidate.

TABLE 2. Structure of an Application Record

Date	Req ID	Role	App ID	Emp ID	Internal	Selected
2021-03	R21-456	Hadoop Dev	7K89	E123	Yes	Yes
2021-03	R21-456	Hadoop Dev	7K90	E124	Yes	No
2021-03	R21-456	Hadoop Dev	7K91	—	No	No
2021-04	R21-457	Data Analyst	7K92	—	No	Yes
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2021-05	R21-459	Sr Dev	7L95	E127	Yes	Yes
2021-05	R21-459	Sr Dev	7L96	—	No	No

Note: Illustrative entries; real identifiers are anonymized. Employee ID appears for internal applicants and for selected external candidates. The Internal column records employee status at the time of application.

full posting text, the application, and the hiring decision. For an internal applicant, the HRIS links the application to the posting of the role the applicant currently holds.

The data cover 6,542 positions across the 2018–2021 window, summarized in Table 1. Of these, 3,836 (58.6%) drew at least one internal application. Internal candidates filled 1,789 (27.3%). A further 1,638 postings predate 2018. These supply the current role for applicants whose positions fall before the main window. The share of positions drawing internal applications rose from 40.3% in 2018 to 62.5% in 2021.

Table 2 lays out the elements the HRIS records for each application event. It holds the application timing, the position specification, and the outcome. Each posting names the languages, the systems, and the domain knowledge the role requires. For an internal applicant, the employee identifier links the application to the origin posting, the description of the role the applicant currently holds. For each application event, then, the data hold two employer-authored documents and an outcome: the role the applicant left, the role the applicant sought, and whether the firm said yes.

Of the internal applications recorded across the 3,836 positions, many applicants held

TABLE 3. Sample of Linked Application Records

Requisition	Current Position	Target Position	Date	Selected
REQ2021-456	Data Engineer	Hadoop Dev.	2021-03	Yes
REQ2021-456	ML Engineer	Hadoop Dev.	2021-03	No
REQ2021-458	Data Analyst	ML Engineer	2021-04	Yes
REQ2021-459	Software Engineer	Sr. Developer	2021-05	Yes
⋮	⋮	⋮	⋮	⋮

Note: Illustrative entries from the 1,034 application events for which both origin and target postings are available.

positions posted before our observation window. For the skill-fit measure to be computed, both the target vacancy posting and the applicant's current-role posting must be present in the HRIS. Where the current-role posting is not available — roles filled through channels that archived no job description — no origin text can be recovered. The 1,034 application events are those for which both postings are present and the measure can be constructed; Table 3 shows a snapshot.

2.3. Measurement Design

Each application event pairs two employer-authored documents. The origin posting describes the role the applicant currently holds. The target posting describes the vacancy the applicant is pursuing. The origin posting is the firm's own specification of what that role requires. No resume, no self-report, no test score enters either side. The firm wrote both postings to one standard and one vocabulary. The origin and target descriptions are directly comparable, which is what lets the test ask whether the text resolves finely enough to separate specific candidates.

The test needs three things together: the origin posting, the target posting, and the selection outcome — a combination that external labor market data do not hold. Within one division, the posting template — the language, format, and vocabulary used to describe roles — stays more or less fixed. The distance between the representations of two postings then more accurately captures the difference in their skill content. If the postings were aggregated across firms, differences in how firms write their postings would fold into that distance, and the measure would pick up writing style as much as skill content. Working within one division holds writing style more or less fixed and leaves the distance to read skill content.

Of the 3,836 positions that drew at least one internal application, 1,789 were filled internally (46.6%); the remaining majority went to the external market despite receiving internal candidates. External search is slower and costlier than an internal fill, and the

firm bore that cost in most of these vacancies. When no internal candidate is selected and the firm goes to the external market instead, it conveys that no applicant cleared the bar, the acceptable minimum fit. Both endpoints of this comparison are role descriptions. The measure asks whether the demands of the origin role carry into the target role; it requires no estimate of what any individual knows, and offers none. Individuals do differ from the roles they occupy, and they differ more the longer they hold them. That gap lies between a person and a posting, and a distance computed between two postings does not run through it.

Section 4 shows that: an applicant random effect for what the individual brings to the assessment, and a test of whether skill fit's predictive content erodes with tenure in role.

3. Measuring Skill Fit from Posting Text

The HRIS structure described in the preceding section lets us link each internal applicant to the job posting used to hire them into their current role, which specifies what that role requires. Reading that posting as the firm's specification of what the applicant's current role requires opens the possibility of assessing skill fit against any vacancy. Evaluating any specific pair (does the skill in the employee's origin posting overlap the skills listed in the new role?) requires a comparative reading of the two texts together, not a rating of either one alone. A language model can be prompted to produce that comparison directly, generating a score for the pair without first needing a labeled dataset built for this exact task. This capability comes from pre-training on a large, varied text corpus, further tuned across many different tasks.

Direct evaluation produces a score for each pair; it leaves no geometry behind. A skill space fixes one: it places every employee and every vacancy in a common embedding, so the full set of positions in a firm can be assessed against each other. In this space, two postings with similar skill sets, where transfer between the roles would be relatively easy, sit near each other; two that do not sit far apart. Here, we use E5 (Wang et al., 2022), a general-purpose sentence embedding model, as a starting point: this paper asks whether posting text carries the signal, not which encoder extracts it best (Appendix F repeats the raw-cosine and skill-space tests with an alternative encoder, all-mpnet-base-v2). Every role in the firm gets a vector in this space, and any pair can then be compared by reading their distance.

E5 belongs to a class of sentence encoders trained on large text corpora. The pre-trained model converts a block of text to a dense vector representation. Where a posting lands in this space reflects the overall profile of the role it describes, not just the terms it contains. Two postings from the same organizational context sit close together in that space partly because they share vocabulary and framing (the language of one financial services IT division, say), and not only because the skills they require transfer. What the

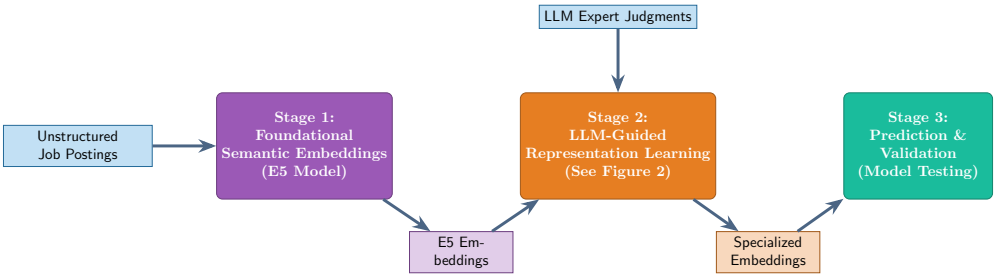


FIGURE 1. Three stages of skill-fit space construction. Stage 1: E5-large-v2 encodes each job posting as a 1,024-dimensional vector. Stage 2: GPT-4.1-mini assesses a stratified sample of 7,500 posting pairs on two score types: technical skill transferability and overall transition viability. Stage 3: A projection network trained on each score type maps E5 embeddings into a 128-dimensional space, where cosine similarity is trained to track assessed skill fit between postings.

measurement needs is a space organized around functional skill overlap: whether what one role requires carries into another, independent of the surrounding vocabulary. The projection reorients E5's richness toward that organizing principle. The question is what to fine-tune against, and the most direct source is a language model's own comparison of the two texts.

Selection outcomes could in principle supervise the projection directly, but the firm's hiring signal covers only the role pairs it happened to evaluate, leaving most role relationships in the firm without an observed signal. Building a skill space that generalizes requires graded comparisons across many role pairs, including ones the firm never put into direct competition. The LLM provides those at scale.

We use generative LLMs in two distinct roles, worth separating from the criterion that stands outside both. GPT-4.1-mini served as the label source: it assessed 7,500 posting pairs on two score types during training and produced the labels the projection networks learned from. GPT-4.1 appears in Section 4 in a second role, as a direct-scoring check: given the same application pairs, it produces direct pairwise judgments independent of the projection, to test whether the hiring-predictive signal resides in the posting text itself. The projection is not judged against GPT scores in either role, but against actual hiring decisions: whether skill-fit in the projected space distinguishes the candidates the firm selected from those it did not.

We preprocessed all job posting texts to remove administrative boilerplate (location details, compensation information, and EEO statements) using a smaller language model; we explicitly preserved skill-relevant content in this step. The construction proceeded in three stages, shown in Figure 1. We drew a stratified sample of 7,500 posting pairs to cover the full range of role relationships, from closely related positions to distant ones. GPT-4.1-mini assessed each pair twice: once on technical skill transferability and once on overall

transition viability, producing two sets of training labels on a 1–10 scale. The first prompt asks the model to judge how well the skills listed in the origin posting would transfer for successful adaptation to the target role, following the demands-abilities framework in person-job fit research (Kristof-Brown, Zimmerman, and Johnson, 2005). The second asks for a more holistic judgment, weighing those same skills alongside seniority and function. We trained a projection network on each score type to map E5 embeddings into a 128-dimensional space, where cosine similarity is trained to track the assessed skill-fit (Hadsell, Chopra, and LeCun, 2006; Schroff, Kalenichenko, and Philbin, 2015). Full construction details, including sampling procedure, assessment protocol, network architecture, and training objective, are in Appendix B.

E5 Cosine is the cosine similarity between raw E5-large-v2 embeddings, without fine-tuning, a baseline reflecting full textual content as organized by the encoder's training corpus rather than skill transferability specifically. **Technical Skill Fit** is cosine similarity in the projection trained on technical skill transferability scores, a reading of posting text focused on hard skill overlap, covering tools, software, programming languages, and domain methods, trained to place postings closer together when their technical requirements transfer. Both are computed from posting text alone, using each employee's origin posting as the origin endpoint of the role-to-role distance. A projection trained on overall transition viability was also estimated, but technical skill transferability proved the stronger predictor of hiring outcomes; full results are in Appendix B. GPT-4.1 direct scoring appears in Section 4 as a separate probe of whether the hiring-predictive signal resides in the posting text, regardless of how it was extracted.

4. Skill Fit and Selection

Across the 1,034 application events, text-derived skill fit varies widely: the interquartile range spans 0.244 to 0.744. The typical selected applicant (median 0.672) sits above the typical applicant who was not (median 0.442) on a measure built from two postings: each applicant's current role and the vacancy they sought.

4.1. Comparing Representations

One skill space has to work across the full range of the firm's technical roles. AUC asks, across all 1,034 application events, whether the higher-scoring applicant was the one selected — including events from entirely different vacancies and hiring managers. The projection trained on skill transferability reaches 0.645; the raw E5 encoder, reading the same postings, reaches 0.596. The bootstrap confidence interval on the difference excludes zero (Table 4).

Table 4 compares three internal benchmarks applied to the same 1,034 application

TABLE 4. Discrimination by Measure: AUC with Bootstrap Confidence Intervals

Measure	AUC	95% CI	Δ vs. E5
E5 Cosine (baseline)	0.596	[0.558, 0.631]	—
Technical Skill Fit	0.645	[0.608, 0.682]	+0.049**
GPT-4.1 Direct	0.625	[0.593, 0.660]	+0.029
N		1,034	

Note: AUC = area under the ROC curve. 95% CIs from 1,000 bootstrap resamples. Δ vs. E5 = difference in AUC relative to E5 Cosine baseline. ** difference significant at the 5% level.

events. Raw E5 cosine similarity provides an unprojected semantic baseline. Technical Skill Fit tests whether the LLM-guided projection adds discrimination. GPT-4.1 Direct tests whether a model reading the two postings without the projection recovers similar signal. Technical Skill Fit reaches an AUC of 0.645 compared with 0.596 for raw E5; the bootstrap confidence interval for that difference excludes zero. GPT-4.1 Direct reaches 0.625, but its difference from E5 is not statistically distinguishable from zero. These comparisons show that the result is not confined to one scoring procedure. They do not establish state-of-the-art predictive performance or exhaust the available lexical, encoder-based, and supervised matching baselines. The goal is to test the informational content of posting text rather than to present a production hiring classifier.

4.2. The Hiring Threshold

Of the 1,034 application events, 634 concern vacancies for which the firm did not select an internal applicant and instead hired externally. These outcomes indicate that no observed internal applicant was selected, but they do not reveal the firm's complete reason for going outside. Candidates may have fallen below a hiring threshold, or the decision may have reflected considerations not observed in the data. The remaining 82 nonselected applications concern vacancies filled by another internal candidate and are excluded from the comparison that follows.

The comparison therefore contrasts 318 selected internal applications with 634 applications to vacancies filled externally. Because external recruiting is generally slower and more expensive than internal movement, hiring externally is consistent with the firm finding the internal pool insufficient on one or more dimensions. It is not, however, a pure or externally validated measure of candidate quality.

For both groups, the score rests on two postings: the origin posting, which documents what the applicant's current role requires, and the target posting, which documents the vacancy. No applicant-specific résumé, assessment, or interview information enters the measure. The 318 selected applicants average 0.572, while the 634 applicants to externally

TABLE 5. Skill Fit by Selection Outcome

Outcome	*N*	Mean Skill Fit	SD
Selected internally	318	0.572	0.296
No internal candidate chosen	634	0.411	0.310
Difference		0.161	

Note: Skill fit: cosine similarity in the fine-tuned skill space (Technical Skill Fit). "No internal candidate chosen": vacancies the firm filled from the external market despite receiving internal applications.

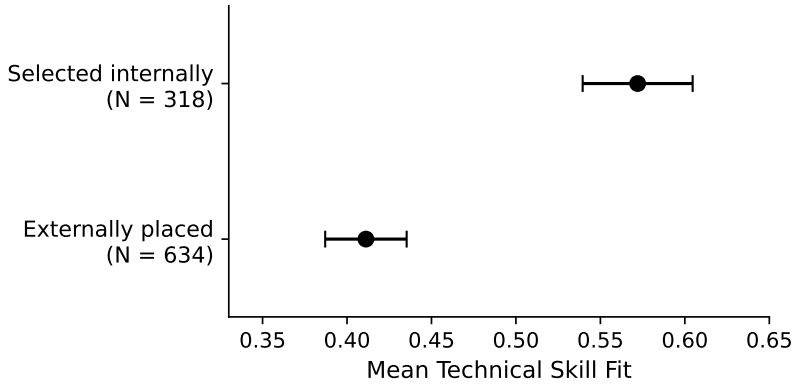


FIGURE 2. Skill Fit: Selected vs. No Internal Candidate Chosen. Mean Technical Skill Fit for applicants selected internally ($N = 318$) against applicants whose positions the firm filled from the external labor market ($N = 634$). Error bars are 95% confidence intervals.

filled vacancies average 0.411, a gap of 0.161, or more than half a standard deviation ($p < .001$). Thus, a distance formed without observing the applicant's selection materials is associated with the firm's eventual decision. The comparison provides criterion-related evidence for documented role similarity, while leaving open the influence of interviews, prior performance, networks, timing, and other unobserved considerations.

A competing account attributes the gap to role composition rather than candidate quality: selected applicants may systematically seek out roles that sit closer to their current positions — a manager backfilling a routine opening, a specialized vacancy that almost always goes internal. If self-selection into proximity is the mechanism, any proximity measure would reproduce the same gap; it would read where people choose to apply, not what the text says about them. The two accounts cannot be separated until two candidates are placed against the same posting.

4.3. Within-Competition Evidence

The within-competition design estimates the relationship between posting-derived skill distance and selection while holding the target vacancy constant. Applicants to the same requisition face the same posted role and hiring context. What varies is the distance between each applicant's current-role posting and the target posting. Ideally, we would like to include candidate fixed effects. However, data about a candidate applying to multiple positions is sparse, making it difficult to use the fixed effects. We include candidate-specific random effects to account for unobserved candidate characteristics.

The documented skill fit varies across an employee's applications because the target posting changes. Applicant random intercepts allow the baseline selection propensity to vary between employees without consuming this role-pair variation. This specification is useful for modeling clustered outcomes,

The within-competition design requires at least two internal applicants competing for the same vacancy. Of the 1,034 application events, 339 meet that condition, across 154 vacancies that drew more than one internal candidate. Of those 154, the firm filled 65 internally; for the remaining 89, it went to the external market despite having multiple internal candidates in the pool. Identification comes from the internally filled vacancies: settings where the firm chose among internal rivals, with everything about the role held fixed and only applicants' skill profiles varying. The variation contracts with the sample: the standard deviation of skill fit within a competition is 0.189, against 0.313 across the full sample. This specification deliberately restricts variation to the within-competition setting, and the tradeoff is a smaller sample. This is the environment in which the text is asked to separate one applicant from another. Competitions that end in internal selection are fewer than those drawing multiple applicants; the odds ratio is estimated with limited precision, but it clears the significance threshold under the within-competition constraint.

The specification accounts for vacancy-level heterogeneity, observed applicant characteristics, and repeated observations from the same applicant. For application i to vacancy j,

$$\text{logit}(p_{ij}) = \beta \cdot s^*_{ij} + \gamma^\top X_{ij} + \alpha_j + u_i, \tag{1}$$

where s^*_{ij} is standardized documented skill fit, α_j are vacancy fixed effects, $u_i \sim \mathcal{N}(0, \sigma^2_u)$ is an applicant random intercept, and X_{ij} contains tenure in the applicant's current role and an indicator for whether the target vacancy is one job-grade level higher. Vacancy fixed effects account for factors shared by applicants to the same requisition. The applicant random-intercept variance is 0.492, although the likelihood-ratio test does not reject a model without that variance component at conventional levels ($p = .19$). The coefficient on skill fit captures its conditional association with selection under the stated model

TABLE 6. Skill Fit and Selection: Mixed-Effects Logistic Regression

	Three Measures, Vacancy FE + Applicant RE		
	E5 Cosine	Technical Skill Fit	GPT-4.1 Direct
Skill Fit (OR)	1.633	2.140**	2.209**
95% CI	[0.966, 2.762]	[1.099, 4.169]	[1.169, 4.176]
p-value	0.067	0.025	0.015
Applying Up (OR)	0.740	0.729	0.574
Tenure in Role (OR)	1.000	1.000	1.000
Applicant RE Variance	0.346	0.492	0.398
N (applications)	339		
N (competitions)	154		

Note: Mixed-effects logistic regression. All models include vacancy fixed effects and applicant random intercepts. s^*_{ij} standardized to full-sample mean and SD (mean = 0.469, SD = 0.313). GPT-4.1 Direct: scored directly from posting text, without embedding or training on firm data. Control variables (Applying Up, Tenure in Role) not significant in any specification ($p > 0.5$).
** $p < 0.05$.

assumptions.

Among applicants applying to positions that received multiple internal applicants, a one-standard-deviation increase in documented skill fit is associated with 2.14 times higher odds of selection (95% CI [1.10, 4.17]; $p = .025$; Table 6). The estimate is conditional on vacancy fixed effects, applicant random intercepts, tenure in role, and whether the application is for a higher grade. It is descriptive: it compares applicants to the same vacancy and does not describe how a given applicant's odds would change if that applicant's skill fit were higher. Across those comparisons, the firm's selection decisions favoured higher skill fit, which is evidence that the measure tracks information reflected in those decisions. It does not imply that the current-role posting captures the applicant's complete abilities. The practical point is narrower: firms with linked origin postings, target postings, and outcomes can test whether their documented role requirements contain selection-relevant information and, once validated, can apply the skill space to subsequent applications.

The design asks whether posting text carries enough information about documented role alignment to predict selection within a competition, not why any particular candidate was chosen. GPT-4.1, reading the same posting pairs without fine-tuning on this firm's data or using the projection network, yields a similar odds ratio of 2.209 under the same specification. This direct-scoring comparison supports the conclusion that the signal is present in the text rather than created solely by the projection procedure. It does not rule out organizational confounding or validate every component of the LLM-generated training labels. The scoring protocol is documented in Appendix C, and model calibration is reported in Appendix D.

We further ask what detail and information provides the value in the skill fit mea-

sure. We rerun the identical specification (Equation 1), with skill fit computed from each posting's extracted list of named skills alone, cutting word count by roughly 90 percent (construction and examples in Appendix E). On the 317 of 339 applications with an extractable skill list, this narrower measure remains positive and substantial (OR = 1.728, p = .085), smaller than the 2.14 estimated on the full posting text. The named skill list alone already carries substantial signal: technical and professional skills are specific, nameable things, and naming them appears to be most of what a posting needs to do. But part of what makes a posting informative seems to sit outside that list as well, in the level at which a skill is expected and the responsibility that comes with it.

The within-competition test establishes that posting text carries enough signal about skill fit to predict who a firm selects from among candidates competing for the same role. A one-standard-deviation difference in skill fit predicts 2.14 times higher odds of selection within a competition, net of applicant characteristics that make candidates generally more or less successful in assessments.

5. Opportunity and Application

Posting text that resolves finely enough to separate candidates within a single competition is specific enough to do something else: locate every employee in the same skill space, and ask what that map shows. Read at the resolution of a specific hiring decision confirmed in Section 4, the text places each person and each vacancy at measurable distances from one another, and the structure of those distances becomes legible.

Two patterns surface from that structure. The first is that skill-aligned internal opportunity is far scarcer than the raw count of open vacancies suggests, and for a meaningful share of applicants it is absent altogether. The second is that employees apply to skill-distant vacancies while closer matches sit open and unpursued. Neither pattern is ordinarily visible without a validated skill-fit measure of this kind.

For a firm, the implication is that analytics of this kind require no new instrument. A firm with a record of its job postings already has the data to construct that skill space and read its internal labor market through it.

5.1. New Vacancies and Skill Fit

For each applicant, each current role is scored against every vacancy open during their application window ($\pm$15 days), and positions that fall within the range the firm's own selection decisions confirmed as close-fit are identified. A position counts as a "skill-aligned opportunity" when its skill-fit score exceeds 0.572, the mean skill-fit among successful applications in the sample (Table 5). The set of such positions is the applicant's opportunity landscape.

TABLE 7. Internal Opportunity Landscapes

Metric	Mean	Median
Total Market Opportunities	75.20	78.0
Skill-Aligned Opportunities	7.62	5.0
Proportion Skill-Aligned	0.098	0.064
Applicants with zero aligned opportunities	153 (14.8%)	

Note: $N = 1{,}034$ application events. Skill-aligned opportunities: vacancies open within a 30-day window with skill-fit ≥ 0.572 against the applicant's current role.

The posting-text distances separate a large field of vacancies from a small field of relevant ones. The average employee faces roughly 75 total vacancies in the window around their application (mean 75.20, median 78.0), but only 7 to 8 of these are skill-aligned (mean 7.62, median 5.0). Among the 1,034 applicants, 153 (14.8%) face zero skill-aligned opportunities, a group whose current skill profile leaves no close internal path open at the moment they apply.

These are employees applying to new roles from a position of constrained opportunity. Their probability of transitioning is low, yet the application itself signals latent interest in moving. Locating an employee in the skill space calibrated to the firm's selection patterns also locates the gap between their current posting and the adjacent vacancies they cannot yet reach. The magnitude of that gap describes what a development program would need to close: which competencies stand between the employee and the positions just outside their range. The gap is computable from the posting text of the employee's current role and the postings of the stream of new roles opening up in the firm, and it can identify employees who may be facing limited advancement opportunities within it.

5.2. Missed Opportunities and Information Frictions

The second pattern concerns whether employees apply to skill-distant vacancies when skill-aligned ones are available to them at the same time. Applications with low skill-fit (below 0.5) reveal the pattern. Many come from employees with few opportunities, the case examined above; but 42% come from employees who had at least one skill-aligned vacancy open and did not pursue it. When a closer match is available and a more distant one is chosen, information frictions, rather than the absence of a fit, become a plausible account of the choice.

Some of these choices may reflect genuine preferences for a particular career path. But employees also pass over positions that closely resemble the ones they applied to. Cosine similarity between the embeddings of each applicant's target job and every other position open in their 30-day window is then computed, applying thresholds derived from the firm's own posting database (95th percentile = 0.698 for "highly similar" positions). Among the 452 unsuccessful applicants for whom at least one other posting was open within the application window, enabling comparison, 54.9% had a highly similar alternative available that they did not pursue; the other 264 non-selected applicants had no comparable posting open in that window. A match that closely resembles a person's revealed choice, left unpursued, is harder to attribute to preference and points more directly to how employees discover internal opportunities.

For a firm, the distances that surface the missed match also identify who missed it. Where employees apply while overlooking closer openings, the skill space names the more suitable posting and makes it possible to surface that information to the employee.

That value grows with the number of employees and the volume of open positions.

An employee who adds skills to their existing set moves closer in this space to positions currently out of range. The distances make legible which specific skills would move them, and toward which vacancies: the gap between an employee's current competencies and the adjacent openings names the skills a targeted program would build. Read against internal demand, these distances show which of those skills the firm needs repeatedly, and which are more readily acquired given an employee's existing competencies. These records are sufficient to align the skills employees bring with what vacancies require and to design development that serves both.

5.3. Scope and Limits

Both patterns are descriptive. The 14.8% who face no aligned opportunity and the 54.9% who passed over a highly similar vacancy are features of the data, not established effects. Both figures depend on the 0.572 threshold, which is based on the mean fit among selected applicants rather than a structural criterion. The 54.9% figure, in particular, comes from unsuccessful applicants, for whom rational non-application is possible. An employee may have had reasons the posting text cannot see: location, timing, private information. The pattern points toward information frictions as a plausible account, but testing it requires data on how employees discover internal opportunities.

The patterns establish a narrower, firmer claim. The skill space read from posting text reveals structure in the internal labor market that the firm had no prior means to observe. Neither the concentration of genuine opportunity within a large field of open vacancies nor the non-application to close matches is legible without reading the text at the resolution the validation confirmed. Both patterns depend on the posting text being specified at this resolution. Where posting text is thinner, the distances collapse, and the patterns dissolve with them. How richly a firm must specify its postings before this structure appears is a research question in its own right.

6. Discussion

Among applicants applying to positions that received multiple internal applicants, a one-standard-deviation increase in documented skill fit is associated with 2.14 times higher odds of selection. Direct GPT-4.1 scoring produces a similar odds ratio of 2.209 under the same specification. Together, these results show that richly specified technical postings contain information associated with selection at the individual-applicant level.

Both the labor economics literature on technological change and skill demand and the empirical organizational economics literature are positioned to draw more deeply on job postings as an information source: the text is rich and, through HRIS systems

and commercial providers, now abundantly available. The first tracks how technological change reshapes role requirements by reading posting text across occupations and over time; pushing that inference below the occupation level requires knowing the text carries information at individual resolution, which linking individual postings to hiring outcomes establishes. The second studies careers and internal mobility with firm-level administrative data, and posting text is the natural source of conditioning information about what roles actually require; conditioning on it rather than job titles requires knowing it reflects what firms select for, and the linkage to hiring patterns provides that assurance. Even careful structural work on multidimensional skill sorting and mismatch has relied on occupation-level skill measures, defined at the level of broad occupation codes (Lise and Postel-Vinay, 2020); individual posting text anchored to hiring outcomes opens that work to a finer grain. This approach shows how to build that anchor: wherever the origin posting, the target posting, and the hiring outcome sit in one system, you can produce the calibration.

Pre-trained language models make fine-grained comparisons of professional postings feasible. Here, GPT-4.1-mini supplies supervisory scores for 7,500 posting pairs, while actual hiring decisions provide a separate downstream criterion. The association with hiring outcomes shows that the resulting representation contains selection-relevant signal. It does not by itself establish that the supervisory scores are unbiased, prompt-invariant, or equivalent to expert human ratings. Direct GPT-4.1 scoring provides a useful construction-independent comparison, but both models belong to the same model family. Human-rating validation, repeated scoring, alternative prompts, and alternative label models remain important robustness tests. The projection's practical advantage is that it creates reusable role vectors in a common geometry, allowing large sets of roles to be compared without a new pairwise LLM call for every comparison.

Those distances, applied to application patterns, surface observations consistent with frictions the internal mobility literature has theorized but cannot observe at this grain (Baker, Gibbs, and Holmstrom, 1994a; Bidwell, 2011; Huitfeldt et al., 2023). For 14.8 percent of applicants, no vacancy was open when they applied that matched closely with the role they hold: a pattern consistent with skill-gap frictions. Among applicants turned down who had a comparable vacancy open at the same moment, 54.9 percent had one closely resembling the role they pursued and did not apply. Job titles and aggregate data leave both patterns unobserved; they appear only when the text is read at individual resolution. Neither number pins down the mechanism behind the pattern it describes: any given passed-over match may reflect preference, sequencing, or information the text cannot see. Why workers pass close matches is now a researchable question, and the instrument opens it.

The evidence comes from one professional workforce at a large financial institution

with richly specified technical postings. It establishes that posting-derived similarity is informative under those conditions. It does not establish that the estimated magnitude generalizes to other firms, occupations, countries, or external hiring markets. Competitive technical labor markets provide a reason to expect other employers to describe requirements concretely, and commercial job-posting databases cover this segment extensively, but replication is required to determine the boundary conditions. Firms that can link applicants to prior-role postings and hiring outcomes can reproduce the calibration with their own records and contribute comparable evidence.

References

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo. 2022. "Artificial Intelligence and Jobs: Evidence from Online Vacancies." *Journal of Labor Economics* 40 (S1).

Baker, George, Michael Gibbs, and Bengt Holmstrom. 1994a. "The internal economics of the firm: Evidence from personnel data." *The Quarterly Journal of Economics* 109 (4): 881–919.

Baker, George, Michael Gibbs, and Bengt Holmstrom. 1994b. "The Wage Policy of a Firm." *The Quarterly Journal of Economics* 109 (4): 921–955.

Bian, Shuqing, Xu Chen, Wayne Xin Zhao, Kun Zhou, Yupeng Hou, Yang Song, Tao Zhang, and Ji-Rong Wen. 2020. "Learning to match jobs with resumes from sparse interaction data using multi-view co-teaching network." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management,*: 65–74.

Bidwell, Matthew. 2011. "Paying more to get less: The effects of external hiring versus internal mobility." *Administrative Science Quarterly* 56 (3): 369–407.

Bidwell, Matthew, and J. R. Keller. 2024. "Stepping Sideways to Step Up: Lateral Mobility and Career Advancement Inside Organizations." *Management Science*.

Deming, David J, and Lisa B Kahn. 2018. "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals." *Journal of Labor Economics* 36 (S1): S337–S369.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. "GPTs are GPTs: Labor market impact potential of LLMs." *Science* 384 (6702): 1306–1308.

Guan, Zhihao, Jia-Qi Yang, Yang Yang, Hengshu Zhu, Wenjie Li, and Hui Xiong. 2024. "JobFormer: Skill-Aware Job Recommendation with Semantic-Enhanced Transformer." *ACM Transactions on Knowledge Discovery from Data* 19 (1).

Hadsell, Raia, Sumit Chopra, and Yann LeCun. 2006. "Dimensionality reduction by learning an invariant mapping." In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2: 1735–1742, IEEE.

Hershbein, Brad, and Lisa B Kahn. 2018. "Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings." *American Economic Review* 108 (7): 1737–1772.

Huitfeldt, Ingrid, Andreas R Kostøl, Jan Nimczik, and Andrea Weber. 2023. "Internal labor markets: A worker flow approach." *Journal of Econometrics* 233 (2): 661–688.

Kokkodis, Marios, and Panagiotis G. Ipeirotis. 2023. "The Good, the Bad, and the Unhirable: Recommending Job Applicants in Online Labor Markets." *Management Science* 69 (11): 6969–6987.

Kristof-Brown, Amy L., Ryan D. Zimmerman, and Erin C. Johnson. 2005. "Consequences of Individuals' Fit at Work: A Meta-Analysis of Person-Job, Person-Organization, Person-Group, and Person-Supervisor Fit." *Personnel Psychology* 58 (2): 281–342.

Lise, Jeremy, and Fabien Postel-Vinay. 2020. "Multidimensional Skills, Sorting, and Human Capital Accumulation." *American Economic Review* 110 (8): 2328–2376.

Minni, Virginia. 2026. "Making the Invisible Hand Visible: Managers and the Allocation of Workers to Jobs." *The Quarterly Journal of Economics* 141 (3): 1871–1931.

Qin, Chuan, Hengshu Zhu, Tong Xu, Chen Zhu, Chao Ma, Enhong Chen, and Hui Xiong. 2020. "An enhanced neural network approach to person-job fit in talent recruitment." *ACM Transactions on Information Systems (TOIS)* 38 (2): 1–33.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),: 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. 2015. "FaceNet: A Unified Embedding for Face Recognition and Clustering." In *Proceedings of the IEEE conference on computer vision and pattern recognition*,: 815–823.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. "MPNet: Masked and Permuted Pre-training for Language Understanding." *Advances in Neural Information Processing Systems* 33.

Wang, Liang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. "Text embeddings by weakly-supervised contrastive pre-training." *arXiv preprint arXiv:2212.03533*.

Yu, Xiao, Jinzhong Zhang, and Zhou Yu. 2024. "ConFit: Improving Resume-Job Matching using Data Augmentation and Contrastive Learning." In *Proceedings of the 18th ACM Conference on Recommender Systems*.

Zhu, Chen, Hengshu Zhu, Hui Xiong, Chao Ma, Fang Xie, Pengliang Ding, and Pan Li. 2018. "Person-job fit: Adapting the right talent for the right job with joint representation learning." *ACM Transactions on Management Information Systems (TMIS)* 9 (3): 1–17.

Appendix A. Sample Job Posting

The following example illustrates the typical depth and structure of job postings in our dataset. Each posting contains detailed specifications across multiple dimensions including role overview, process context, responsibilities, and skill requirements.

Representative Job Posting from Dataset

Overview
We are one of the world's leading financial institutions ... and risk management products and services.

Process Overview
Build and evolve a consistent Authorized Data Source within Consumer & Small Business Bank (CSBB) ... both the strategic and tactical analytics needs of the Consumer Bank.

Job Description
Hadoop developer for multiple initiatives. Develop Big Data Strategy and Roadmap for the Enterprise. Experience in Capacity Planning, Cluster Designing and Deployment. Benchmark systems, analyze system bottlenecks, and propose solutions to eliminate them. Develop highly scalable and extensible Big Data platform, which enables collection, storage, modeling, and analysis of massive data sets from numerous channels.

Responsibilities
Assists the team with the design of the architect layer to ensure re-usable metrics and attributes within the reporting layer. Responsible for creating and maintaining necessary documentation (MDR) to ensure audit readiness. Prototype improvement ideas. Work effectively with the global team. Expected to play technical leadership as an individual contributor.

Mandatory Skills
Extensive knowledge of Hadoop stack and storage technologies HDFS, MapReduce, Yarn, HIVE, sqoop, Impala, spark, flume, kafka and oozie. Extensive Knowledge on Bigdata Enterprise architecture (Cloudera preferred). Experience in No SQL Technologies (Cassandra, Hbase).

Desired Skills
Experience in Real time streaming (Kafka). Experience with Big Data Analytics & Business Intelligence and Industry standard tools integrated with Hadoop ecosystem. (R, Python). Visual Analytics Tools knowledge (Tableau). Data Integration, Data Security on Hadoop ecosystem.

Note: Firm and location details have been redacted. This posting demonstrates the rich technical content typical of positions in our dataset, with detailed skill requirements that enable precise measurement of job characteristics.

Appendix B. Skill Space Construction

This appendix documents the construction of the skill-fit spaces introduced in Section 3. The construction proceeds in three stages: we generate labeled pairs by prompting an LLM to score job-pair skill transferability, train a projection network on those scores, and embed all job postings into the resulting spaces. A skill-fit space is a learned projection of E5 embeddings (1,024 $\rightarrow$ 128 dimensions), trained so that cosine similarity in the projected space tracks LLM-assessed skill transferability rather than full textual content. Subsections cover training pair sampling (§B.1), LLM assessment protocol (§B.2), projection network architecture (§B.3), and training objective and procedure (§B.4).

B.1. Training Pair Sampling

Training the projection network requires pairs of job postings with known skill-similarity labels; the 7,500 pairs described here provide that supervision. A random draw over-represents common job families and under-represents rare roles; stratified sampling covers the full range of skill distances. Jobs were clustered in the E5 embedding space using HDBSCAN (a density-based clustering method), which handles varying-density clusters and labels outlier postings as noise; both properties are required for the three pair types below. Within each pair-type stratum, pairs were drawn by random draw.[2]

Intra-cluster pairs provide similar-skill training signal; cross-cluster pairs provide dissimilar-skill signal; noise-involved pairs capture postings without stable cluster membership.

TABLE B.1. Training pair composition.

Pair Type	Definition	*N*	%
Intra-cluster	Both jobs from the same cluster	1,727	23.0
Cross-cluster	Jobs from two different clusters	3,801	50.7
Noise-involved	At least one job labeled as noise by HDBSCAN	1,972	26.3
Total		7,500	100.0

B.2. LLM Assessment Protocol

Two assessment types were collected for each pair. Technical skill transferability assesses alignment of hard skills only: tools, software, languages, and domain-specific methods. Overall transition viability provides a holistic judgment of the full transition, synthesizing

[2] 6,308 unique job postings contribute to the 7,500 pairs; mean pairs per posting is 2.38, maximum 14.

technical skills, seniority alignment, management responsibilities, and role fit. Two assessment types are collected in a single API call; training separate projection networks on each score type produces two distinct spaces. Table B.2 shows the scoring rubric for each type.

TABLE B.2. Scoring rubric for LLM assessments.

Score Range	Technical Skill Transferability	Overall Transition Viability
1–3	Few overlaps; significant relearning required.	Highly challenging or ill-advised.
4–6	Some core overlap; notable gaps or adaptation needed.	Feasible but with significant challenges or uncertainties.
7–9	Good key skill overlap; most skills present or easily transferable.	Viable and likely successful with reasonable effort.
10	Excellent alignment; near-perfect technical match.	Highly promising; very strong and natural fit.

Both scores were collected in a single API call for each pair. GPT-4.1-mini (`gpt-4.1-mini-2025-04-14`) was used to generate assessments for all 7,500 pairs; the prompt is shown below.

LLM Assessment Prompt Template

You are an expert HR analyst. Your task is to compare Job X (Source Role) and Job Y (Target Role) based **only** on their provided descriptions. Assess the transition compatibility from Job X to Job Y. Do not use external knowledge.

Provide your assessment as a single JSON object. The JSON must contain two top-level keys: "technical_skill_transferability" and "overall_transition_viability". Each key maps to an object with "score" (integer 1-10) and "justification" (1-2 concise sentences referencing JD specifics).

— JOB DESCRIPTIONS —

Job X (Source Role):

`{jd_text_A}`

Job Y (Target Role):

`{jd_text_B}`

— ASSESSMENT CRITERIA & SCORING GUIDELINES —

Assign scores from 1 (Very Poor) to 10 (Excellent). **Strictly adhere to the Scoring Guide for each criterion.**

1. technical_skill_transferability:

- **Definition**: Alignment of technical skills (tools, software, languages, methods). Consider direct overlaps and reasonable learnability from Job X's foundation.
- **Scoring Guide**:
  - 1-3: Few overlaps; significant relearning.
  - 4-6: Some core overlap; notable gaps or adaptation needed.
  - 7-9: Good key skill overlap; most skills present or easily transferable/learnable.
  - 10: Excellent alignment; Job X's technical skillset is a near-perfect match for Job Y.

2. overall_transition_viability:

- **Definition**: Holistic judgment of transition success (Job X to Job Y). Synthesize technical skills, seniority alignment, management responsibilities, growth potential, and role fit.
- **Scoring Guide**:
  - 1-3: Highly challenging or ill-advised.
  - 4-6: Feasible but with significant challenges/uncertainties.
  - 7-9: Viable and likely successful with reasonable effort.
  - 10: Highly promising, very strong and natural fit.

Ensure your output is ONLY the JSON object, starting with `{{` and ending with `}}`.

Justifications were collected but not used in model training; scores only served as supervision signals. Of the two projections, Technical Skill Fit outperforms Overall Transition Fit in predicting hiring outcomes (AUC 0.645 vs. 0.610); the technical reading is used throughout the main analysis.

The LLM assessments are supervisory signals rather than ground-truth measures of employee skill. The current study does not include expert human ratings of the 7,500 training pairs, repeated ratings under alternative prompts, or labels from a different model family. Downstream association with hiring outcomes provides criterion-related evidence for the learned representation, but it does not establish prompt invariance or eliminate systematic bias in the supervisory scores. These limitations constrain claims about construct validity and motivate direct reliability tests in future work.

Implementation parameters:

- Model: GPT-4.1-mini (`gpt-4.1-mini-2025-04-14`)
- Temperature: 0.05
- Output format: JSON
- Pairs assessed: 7,500
- Total assessments: 15,000 (2 per pair)

B.3. Projection Network Architecture

Two projection networks were trained separately, one on technical skill transferability scores and one on overall transition viability scores. Each network maps the 1,024-dimensional E5 embedding into a lower-dimensional space in which cosine similarity reflects skill transferability rather than general textual content. The architecture of each network is shown in Table B.3.

TABLE B.3. Projection network architecture.

Layer	Dimensions	Operations
Input	1,024	E5 embeddings
Hidden 1	512	LayerNorm $\rightarrow$ GELU $\rightarrow$ Dropout (0.2)
Hidden 2	256	LayerNorm $\rightarrow$ GELU $\rightarrow$ Dropout (0.1)
Output	128	LayerNorm $\rightarrow$ L2-normalized to unit length

L2 normalization at the output constrains all embeddings to the unit hypersphere, so cosine similarity is well-defined and bounded to $[-1, 1]$. Development set Pearson correlation measures how well the projected cosine similarities rank pairs relative to the LLM scores on the 15% held-out development set (see §B.4 for the data split); the dimensionality producing the highest correlation was selected. Results across three candidate dimensions are shown in Table B.4. 128 dimensions selected.

TABLE B.4. Output dimensionality selection.

Output Dimensions	Dev Set Pearson Correlation
64	0.813
128	0.840
256	0.836

B.4. Training Objective and Procedure

The loss operates on pairs of job pairs (Hadsell, Chopra, and LeCun, 2006). The network is trained to rank pairs by cosine similarity in the same order the LLM ranked them by skill transferability, with the required margin scaled by how different the LLM scores were. For a quadruplet (A, B, C, D) in which $\tau_{AB} > \tau_{CD}$, the training objective is:

$$L = \max\big(0,\ |\tau_{AB} - \tau_{CD}| \times \kappa - \big(\mathrm{sim}(z_A, z_B) - \mathrm{sim}(z_C, z_D)\big)\big) \tag{B.1}$$

where τ_{AB}, τ_{CD} are the LLM scores for job pairs (A, B) and (C, D); z_i is the projected 128-dimensional embedding of job i; κ is the margin scale parameter; and $\mathrm{sim}(\cdot, \cdot)$ is cosine similarity. Using pairs of pairs rather than single pairs preserves the interval-level information in the 1–10 scores: the model learns not just which pair is more skill-similar but by how much.

Figure B.1 illustrates the training step.

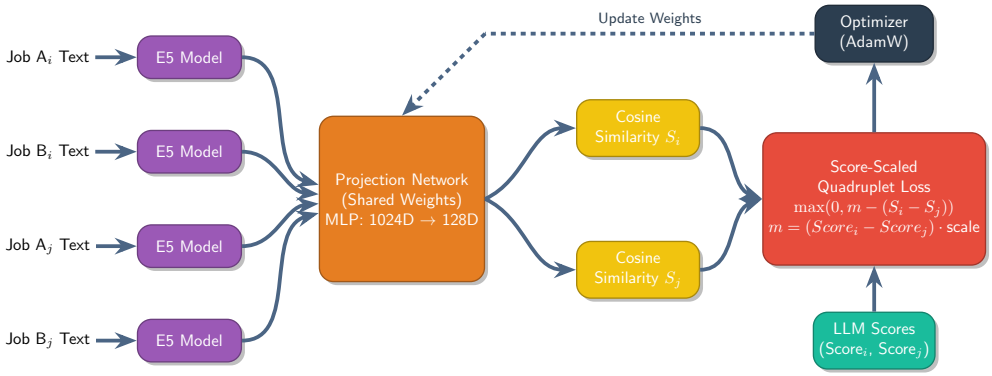


FIGURE B.1. Training step. Four job postings form two pairs. Each posting's E5 embedding passes through the projection network to produce a 128-dimensional representation; cosine similarity is computed for each pair; the loss is computed from the two pair-level similarities and used to update network weights.

Training parameters:

- Data split: 6,375 training pairs (85%), 1,125 development pairs (15%)
- Optimizer: AdamW
- Learning rate: 0.0001
- Weight decay: 0.0001
- Batch size: 256
- Epochs: 50
- Training time: ∼9.3 minutes (Tesla T4)

The margin scale κ was selected by grid search on the development set; results are

shown in Table B.5. $\kappa = 0.10$ selected.

TABLE B.5. Margin scale selection.

κ	Dev Set Pearson Correlation
0.05	0.821
0.10	0.840
0.20	0.831

Convergence:

- Best development Pearson correlation: 0.840 (epoch 46)
- Inference: $<$1 ms per posting on CPU
- Storage: $\sim$0.5 KB per posting

Appendix C. GPT-4.1 Direct Scoring Protocol

The GPT-4.1 Direct measure serves as an independent check on the main result: if the hiring-predictive signal arises from the construction process (the training labels, the projection network, the fine-tuning on this firm's postings) rather than from the posting text itself, a language model with no exposure to any of that construction should produce a weaker or null result. The purpose of this appendix is to document how GPT-4.1 Direct scores were generated.

Approach

GPT-4.1 (`gpt-4.1`) was used to score each of the 1,034 application pairs directly. For each pair, the model received the posting text for the applicant's current role and the posting text for the target vacancy and returned a single integer score for technical skill fit. No embeddings, no projection networks, no training data from this firm, and no fine-tuning of any kind were involved. The model's assessment is based solely on the two posted job descriptions as text.

The same preprocessed posting texts used for E5 encoding were supplied to the model: postings exceeding the E5 encoder's 512-token limit were summarized by GPT-4.1 prior to the main analysis to preserve core responsibilities and qualifications; these summarized versions were used here as well, ensuring comparability across measures.

Prompt

GPT-4.1 Direct Scoring Prompt

You are evaluating internal job transfer compatibility at a large financial firm.
Current role job description:
`{jd_current}`

Target role job description:
`{jd_target}`

Rate the technical skill fit between these two roles on a scale of 0 to 10:

- 0–2: Minimal technical overlap; major retraining required
- 3–5: Some technical overlap; significant gaps exist
- 6–7: Good technical overlap; most skills transfer
- 8–10: Excellent technical alignment; near-perfect skill match

Consider: programming languages, tools, systems, technical methods, and domain knowledge.
Respond with a single integer only. No explanation.

Implementation parameters

- Model: GPT-4.1 (`gpt-4.1`)
- Temperature: 0.05
- Output format: single integer, 0–10
- Maximum tokens: 10
- Pairs scored: 1,034 (all application events with linked origin and target postings)

Scores are resume-safe: the scoring script writes results progressively to a `.jsonl` file and skips already-scored pairs on re-run, ensuring a complete and reproducible output file.

Use in the hiring model

The resulting score is used as a direct skill-fit measure in the same mixed-effects logistic regression reported for the trained projection measures (Section 4.3). Scores are standardized to the full-sample mean and standard deviation before entry, matching the standardization applied to the projection-based measures. No additional transformation is applied. The model enters the regression as a drop-in replacement for the trained projection similarity scores, with the same specification: vacancy fixed effects and applicant random intercepts.

Appendix D. Model Calibration Assessment

To ensure the reliability of the probabilistic predictions we performed a calibration assessment. A well-calibrated model produces predicted probabilities that correspond to the true underlying likelihood of the event. For instance, if we consider all instances where the model predicted a 20% chance of selection, we would expect to see an actual selection rate of approximately 20% within that group.

We evaluated calibration using a visual calibration plot, shown in Figure D.2. We grouped all 1,034 application events into deciles based on their predicted selection probability. For each decile, we plotted the mean predicted probability against the actual observed fraction of positive outcomes (i.e., the selection rate). A perfectly calibrated model would have all points lying on the 45-degree diagonal line.

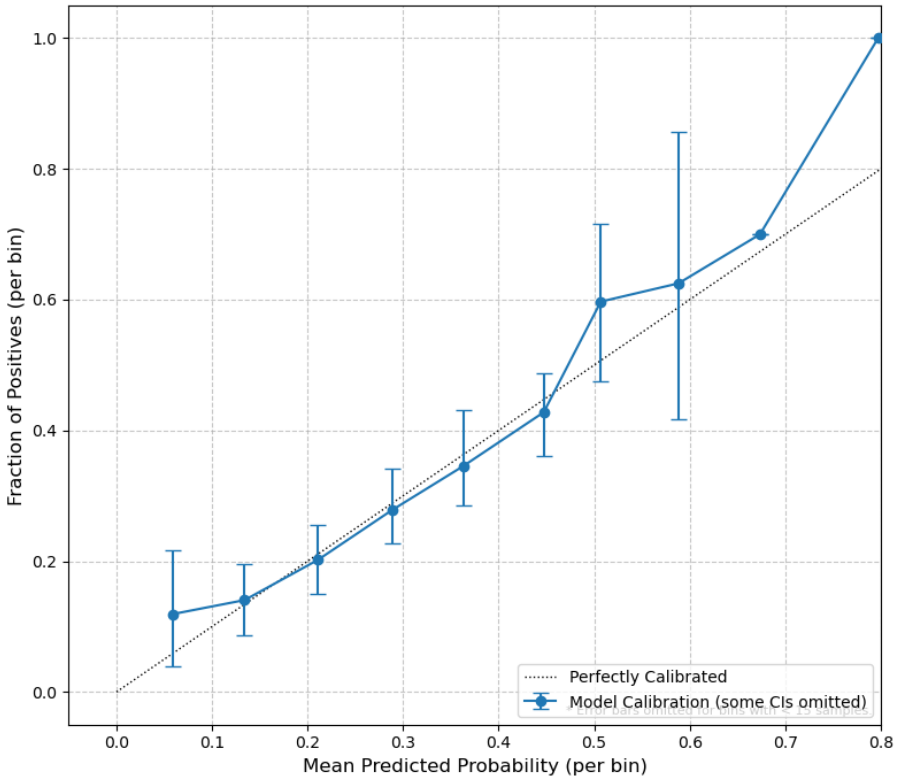


FIGURE D.2. Calibration Plot for the Full Interaction Model

Note: The plot compares mean predicted selection probabilities (x-axis) with observed selection rates (y-axis) across ten probability bins. The dotted line represents perfect calibration. Points represent the mean for each bin, with vertical bars indicating 95% bootstrap confidence intervals (1,000 replications). The model demonstrates good calibration, with most points lying close to the diagonal and within the confidence intervals of perfect calibration.

As shown in Figure D.2, the model is well-calibrated across the range of predictions. The points track the diagonal reference line closely, indicating that the model's confidence in its predictions is generally warranted. The slight deviations are well within the 95%

confidence intervals, suggesting they are not statistically significant.

To complement the visual analysis, we also calculated the Brier score, which is the mean squared error between the predicted probabilities and the actual outcomes. Our model achieved a Brier score of 0.193, indicating a good level of probabilistic accuracy.

Appendix E. Skill-Content Isolation Check

Each posting was reduced to a stripped version containing only its named skills, removing everything else a posting typically specifies. A two-tier regex-based extractor pulls the labeled skills section where one exists, falling back to a whole-document scan for skill and tool terms when it does not, with job-title lines, generic action verbs, and section headers excluded. Within each extracted item, proficiency and experience qualifiers are stripped ("Strong knowledge of Python" becomes "Python"), and the resulting terms are wrapped in a single fixed template ("This position requires the following core skills: X, Y, and Z.") common to every posting. Each stripped posting is then embedded using the same E5-large-v2 model and passed through the same trained projection network used to construct Technical Skill Fit in Section 3, and skill fit is computed as cosine similarity in that space, exactly as for the full-text measure.

E.1. Coverage and Reduction

Across the 1,329 unique postings underlying the 1,034-application estimation sample (current-role and target-role postings, deduplicated), this extraction sharply reduces posting length while preserving only named skills, tools, and technologies.

TABLE E.6. Posting length before and after skill extraction.

Statistic	Full posting (words)	Barebones posting (words)
Mean	248.6	21.5
Median	244	20
Min / Max	132 / 384	0 / 60

Note: $N = 1{,}329$ unique postings. Mean reduction is 91.1% (median 91.7%). 33 postings (2.5%) yielded no extractable skill terms and are excluded from the stripped-text analysis in Section 4.3. Neither location, education requirements, nor shift or schedule details survive extraction in any sampled posting.

E.2. Within-Competition Robustness Result

We recomputed skill fit on the stripped postings using the same trained projection, on the subset of the within-competition estimation sample with an extractable skills section ($N = 317$ of 339, 93.5% coverage). The comparison uses the identical specification reported in Table 6 (Equation 1): vacancy fixed effects, an applicant random effect, and the tenure-in-role and applying-up controls, estimated on the stripped skill-fit measure in place of the full-text measure. The stripped-text estimate remains positive and substantial, though it falls short of conventional significance (OR = 1.728, $p = .085$).

E.3. Representative Examples

Two representative postings illustrate the extraction. In both cases, education, location, and schedule details are present in the full posting and do not survive into the stripped version.

Example 1: Full Posting

Job Description:
The individual will be part of Mainframe Production Support team. He/She will be required to collaborate with both internal and external global teams across various locations for the project delivery.
Responsibilities: Support daily production Support activates which includes monitoring of jobs and fixing the job abends. Develop ad-hoc programs and JCL to solve or perform reports for the CQ request. Identify any gaps in the system and propose the solution. Collaborate with external programmers to coordinate delivery of software applications. Works on multiple systems or technologies underpinning a given application.
Requirements
Mandatory skills: Should have experience in Mainframe application development and Support. Should have experience in DB2 along with Store Proc. Should have knowledge in COBOL, DB2 programing. Should have Knowledge on CICS. Good interpersonal, written, and oral communication skills.
Desired Skills: Experience native Store Proc.
Experience Required: 4–6 yrs
Education Required: Any Graduate
Work Location: Hyderabad
Shift Timings: 10:00 AM - 7:30 PM

Example 1: Barebones

This position requires the following core skills: Core Java, DB2, Store Proc, COBOL, and CICS.

Example 2: Full Posting

The job requires working on applications catering to Trade Booking, Pricing, Risk Management and downstream reporting of Equity Linked Products. The role entails one to work closely with Business Users, Analytics and various Technology teams to develop correct and optimal solutions for one or more applications within the ELT space.
Responsibilities: The candidate would be working on the development of front office applications using cutting edge technology. The candidate would be interacting with Traders / Quants teams at other locations on a regular basis.
Requirements:
Mandatory Skills: Good experince in Core Java. Familiarity with Multithreading, Data Structures, Design Patterns, Databases.
Desired skills: Experience on Python. Relevant work experience in Investment Banking domain.
Experience Range: 3–6 years of Hands-on experience on Core Java
Education: B.E./B.Tech/M.S./M.Tech with good academic
Work Timings: 10:30 AM – 7:30 PM
Job Location: Malad, Mumbai

Example 2: Barebones

This position requires the following core skills: Core Java, Multithreading, Data Structures, Design Patterns, Databases, Python, and Investment Banking.

Appendix F. Encoder Choice

Section 3 builds the skill space on E5-large-v2 (Wang et al., 2022), a general-purpose sentence encoder, and the results could depend on that choice. This appendix repeats two of the tests of Section 4 with an alternative general-purpose encoder, all-mpnet-base-v2 (Reimers and Gurevych, 2019; Song et al., 2020): discrimination across all 1,034 application events (Table 4) and the within-competition odds ratio (Table 6). The postings, application events and specifications are those of the main text. Two measures are rebuilt for the new encoder. The first is the raw cosine similarity between the applicant's current posting and the target posting. The second is a skill space trained on the encoder's embeddings with the recipe of Appendix B: the same labels, split, loss and 128-dimension network, keeping the training run with the highest development-set correlation. Each score is standardized over the 1,034 events.

TABLE F.7. Alternative Encoder: Correlation, Odds Ratio and AUC

Measure	Correlation with E5	OR [95% CI]	AUC [95% CI]
E5-large-v2, raw cosine	—	1.633 [0.966, 2.762]	0.596 [0.558, 0.631]
E5-large-v2, skill space	—	2.140 [1.099, 4.169]	0.645 [0.608, 0.682]
all-mpnet-base-v2, raw cosine	0.665	1.459 [0.863, 2.466]	0.655 [0.620, 0.690]
all-mpnet-base-v2, skill space	0.847	1.641 [0.930, 2.895]	0.639 [0.602, 0.675]

Note: Correlation is the Pearson correlation across the 1,034 application events between the score and its E5 counterpart (raw with raw, skill space with skill space). OR is the odds ratio per one standard deviation, from the specification of Table 6 (vacancy fixed effects, applicant random effect, applying-up and tenure controls; 339 applications, 154 competitions). AUC is pooled over the 1,034 events, with 95% intervals from 1,000 bootstrap resamples. E5 rows reproduce Tables 6 and 4. all-mpnet-base-v2 reads the first 384 tokens of each posting; 594 of the 1,290 distinct posting texts are longer.

The two encoders order applicants similarly but not identically (correlation with E5 of 0.665 for the raw cosine and 0.847 for the skill space). Under both, higher scores are associated with selection: every odds ratio exceeds one, and AUC lies between 0.596 and 0.655. With all-mpnet-base-v2 the odds ratios are not significant at conventional levels ($p = .159$ for the raw cosine, $p = .088$ for the skill space). The gain from the skill space over the raw cosine is observed with E5 but not with all-mpnet-base-v2, whose raw cosine reaches an AUC of 0.655 against 0.639 for its skill space. These results are consistent with posting text carrying selection-relevant signal under either encoder; they do not establish that the projection adds discrimination for every encoder.