Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 625-648
DOI: https://doi.org/10.51903/jtie.v4i3.529

Calibrated Resume-Job Matching for Trustworthy LLM-Assisted Recruiter Screening: Pairwise Matching, Probability Calibration, and Selective Refusal on Two Public Recruitment Datasets

Jiaying Jin*[1]
Email: jj3373@columbia.edu
[1]*Applied Analytics, Columbia University, NY, USA*
*Corresponding Author

Abstract

Recruiter screening increasingly relies on large language model (LLM)-assisted workflows, but high-stakes applications require reproducible matching, calibrated probabilities, and reliable handling of uncertain cases. This study evaluates a screening framework combining matching, calibration, and selective refusal using two public datasets: resume-job-description-fit for supervised pairwise learning and Resume-Screening-Dataset for benchmarking and external generalization. After deterministic preprocessing, we compared cosine similarity, alignment features, TF-IDF pairwise models, and hybrid models integrating text, alignment, and title information. The strongest probabilistic models were calibrated with Platt scaling and isotonic regression and evaluated under confidence-based refusal. On the resume-job-description-fit test set, the best three-class model achieved a macro-F1 of 0.450. For binary shortlist-versus-reject screening, the title-augmented hybrid model obtained 0.654 balanced accuracy, 0.647 F1, and 0.699 AUROC. Platt calibration improved probability estimates by reducing the Brier score from 0.232 to 0.226 and negative log-likelihood from 0.772 to 0.675. Selective refusal further improved in-domain accuracy, while cross-dataset transfer remained weak (AUROC 0.47–0.51). These results indicate that matching, calibration, and selective refusal enhance trustworthy within-domain screening, although human review remains essential under distribution shift.

Keywords: *resume-job matching; recruiter screening; probability calibration; selective refusal; trustworthy AI.*

I. INTRODUCTION

Recruiter screening is a structured text-matching problem that sits at the intersection of information retrieval, text classification, and decision support. A recruiter or an LLM-assisted recruiting agent receives a resume, a job description, and often a role title, then decides whether the applicant should be advanced, held for review, or rejected. Classical information retrieval established the foundations for document-query matching with sparse lexical signals and probabilistic ranking (Manning et al., 2008; Robertson & Zaragoza, 2009; Salton & McGill, 1983). More recent neural work replaced or augmented lexical overlap with deep semantic representations such as DSSM, BERT, Sentence-BERT, and SimCSE (Devlin et al., 2019; Gao et al., 2021; Huang et al., 2013; Reimers & Gurevych, 2019). In recruiting, these representation advances are useful, but they do not solve the governance problem. A screening system still needs to express how trustworthy each decision is.

LLM-centered workflows (Sun et al., 2023) make these control requirements even sharper. A generative model can summarize a resume, extract skills, justify a decision, or converse with

Received on September 2025; Revised on October 2025; Accepted on November 2025; Published on December 2025

recruiters, but none of those behaviors guarantees that its final screening score is calibrated. Transformer encoders and sentence-embedding models have improved semantic matching dramatically, yet confidence quality is still a separate empirical property rather than a by-product of language fluency (Devlin et al., 2019; Guo et al., 2017; Reimers & Gurevych, 2019). In practical screening systems, confidence needs to be attached to a concrete decision rule with measurable error rates. That is why this paper focuses on a transparent classifier stack even though the intended application setting is LLM-assisted recruiting. The paper studies the part of the system that decides when a model should act, when it should defer, and how confident a recruiter should be in the score.

That requirement is not optional in hiring. Algorithmic hiring systems affect access to interviews, human attention, and employment opportunities, so their outputs must be auditable, stable, and easy to review. The hiring literature has already shown that automated screening can reproduce hidden biases, obscure decision logic, and overstate capability when evaluation is narrow or poorly documented (Kochling & Wehner, 2020; Raghavan et al., 2020). For this reason, the most useful technical question is not only whether a model can rank candidates, but whether it can produce scores that are interpretable as probabilities, separate high-confidence from low-confidence cases, and remain reliable when the data distribution changes (Zheng et al., 2024). Calibration directly addresses the probability question.

A calibrated screening model should assign scores that match empirical frequencies: among applicants assigned a shortlist probability near 0.80, roughly 80% should truly belong to the shortlist class. This idea has a long methodological lineage, from Platt scaling and class-wise probability correction to modern post-hoc calibration for neural classifiers (Guo et al., 2017; Kull et al., 2017; Niculescu-Mizil & Caruana, 2005; Platt, 1999; Zadrozny & Elkan, 2001, 2002). In high-stakes pipelines, calibration matters because downstream policies depend on thresholds. A recruiter dashboard, an automated workflow, or an LLM agent can only interpret model confidence correctly when that confidence is empirically grounded.

Selective refusal, also called abstention or selective classification, addresses the action question. Instead of forcing a decision for every applicant, the system can refuse borderline cases and reserve them for human review. The theory of reject-option classification goes back to Chow (1970) and was developed further in selective classification research by El-Yaniv and Wiener (2010) and Geifman and El-Yaniv (2017). Later work on trust scores, misclassification detection, out-of-distribution detection, and predictive uncertainty reinforced the same principle: a classifier should not act on cases that it does not understand well enough (Hendrycks & Gimpel, 2017; Jiang et al., 2018; Lakshminarayanan et al., 2017). In recruiter screening, selective refusal is

operationally natural. Recruiters already triage cases, and the machine should strengthen that triage process rather than hide uncertainty behind a forced label. Despite that clear need, public benchmarking practice in resume screening is still dominated by within-dataset classification accuracy. Public datasets often differ sharply in their document length, role structure, label definitions, and writing style, yet many papers report only in-domain performance. That leaves three important questions unanswered. First, how much screening performance can be achieved with a transparent pairwise matcher on a public fit dataset? Second, how much probability quality can be recovered through explicit calibration? Third, does a confidence-based refusal policy remain useful when the model is applied to a different resume-screening corpus?

This paper answers those questions with a fully empirical study on two public Hugging Face datasets. The first dataset, resume-job-description-fit, contains 8,000 paired resumes and job descriptions with three fit labels and is used for pairwise fit learning (cnamuangtoun, 2024). The second dataset, Resume-Screening-Dataset, contains 10,174 screening rows with role titles, resumes, job descriptions, binary select-or-reject decisions, and textual reasons, and is used both as a second benchmark and as an external generalization target (AzharAli, 2022.). The study deliberately uses a sparse-and-linear modeling stack rather than a large fine-tuned transformer so that every result is fully reproducible on the supplied datasets and the decision-control layer can be isolated cleanly. This choice is aligned with trustworthy LLM-assisted recruiting: the matching model can serve as a calibrated scoring layer, a benchmarking reference, or a conservative fallback beneath a larger generative system.

The contribution of the paper is fourfold. First, it provides a complete experimental benchmark on the specified datasets with no illustrative or placeholder results. Second, it organizes the screening problem along the exact line of matching + calibration + selective refusal. Third, it reports both in-domain and cross-dataset behavior, which exposes where trustworthy screening succeeds and where it fails. Fourth, it translates those findings into operational recommendations for recruiter screening. The results show that in-domain screening can be improved substantially by combining pairwise text features with alignment and title features, that post-hoc calibration improves probability quality on both datasets, and that selective refusal raises in-domain decision reliability. The results also show that zero-shot transfer from the fit dataset to the screening dataset collapses to near-chance ranking performance, which means calibrated human review remains indispensable when the deployment data move away from the training distribution.

The applicant-evaluation perspective is important here. A screening tool is not only a retrieval engine that finds similar documents; it is also an evaluator that translates textual evidence into a consequential recommendation. That makes false positives and false negatives asymmetric in

practice. A system that shortlists too aggressively floods recruiters with weak candidates, while a system that rejects too aggressively suppresses potentially suitable applicants before a human reads them. Balanced accuracy, calibration quality, and selective risk are therefore more informative for applicant evaluation than raw accuracy alone. This paper adopts that evaluator perspective throughout by measuring both ranking quality and decision quality at explicit operating thresholds.

II. LITERATURE REVIEW

Prior work most relevant to this study falls into four connected strands: document matching for retrieval, neural semantic representation learning, trustworthy decision-making in algorithmic hiring, and uncertainty-aware classification. Classical information retrieval framed matching as the estimation of relevance from sparse lexical evidence, which remains valuable for professional documents because resumes and job descriptions contain highly discriminative skill names, role titles, certifications, and years of experience (Manning et al., 2008; Robertson & Zaragoza, 2009; Salton & McGill, 1983). For recruiter screening, this means that exact overlap is not merely a crude baseline; it is part of the signal structure that human reviewers actually use when they compare applicants against job requirements.

Later neural work expanded the notion of match quality from lexical overlap to semantic similarity. DSSM, BERT, Sentence-BERT, and SimCSE showed that learned representations can recover paraphrastic or semantically related evidence even when exact tokens differ (Devlin et al., 2019; Gao et al., 2021; Huang et al., 2013; Reimers & Gurevych, 2019). That shift is especially relevant for resumes because equivalent qualifications are often expressed with heterogeneous language. At the same time, better semantic representations do not by themselves define an operational screening policy. A system can rank pairs well and still produce scores that are not interpretable as probabilities or stable across datasets.

The hiring literature sharpens that distinction between prediction and decision. Reviews of algorithmic hiring systems (Li, 2024) emphasize risks involving discrimination, opacity, and overclaimed capability, particularly when evaluation is narrow or when system outputs are treated as objective facts rather than fallible recommendations (Kochling & Wehner, 2020; Raghavan et al., 2020). In other words, the core challenge is not only to produce a relevance score, but to place that score inside a governance structure that supports auditability, threshold setting, and human oversight. This is precisely the context in which an LLM-assisted (Chen et al., 2023) recruiter workflow needs an explicit control layer beneath any generative explanation or conversational interface.

A substantial methodological literature addresses the probability side of that control layer. Platt scaling, classwise calibration, and later post-hoc methods such as isotonic and beta calibration all start from the same principle: decision thresholds are only meaningful when scores correspond to empirical frequencies (Guo et al., 2017; Kull et al., 2017; Niculescu-Mizil & Caruana, 2005; Platt, 1999; Zadrozny & Elkan, 2001, 2002). For screening systems, calibration matters because shortlist rules, escalation policies, and downstream human review depend on probability estimates rather than on raw margins alone. A model with respectable accuracy but poor calibration can still mislead recruiters if it is systematically overconfident on marginal applicants.

The action side of trustworthy screening is addressed by reject-option and uncertainty research. Chow (1970) established the decision-theoretic basis for refusing uncertain cases, and later work on selective classification, misclassification detection, trust scores, and predictive uncertainty showed how abstention can improve reliability when confidence tracks correctness (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Hendrycks & Gimpel, 2017; Jiang et al., 2018; Lakshminarayanan et al., 2017). This perspective aligns naturally with recruiter practice, where borderline cases are ordinarily set aside for manual review rather than forced into premature accept-or-reject decisions. The remaining gap is empirical: public resume-screening benchmarks rarely evaluate matching, calibration, and refusal together under both in-domain testing and cross-dataset shift. The present study addresses that gap by treating the screening model as a decision-control layer whose quality must be judged not only by classification performance, but also by probability quality and abstention behavior.

III. RESEARCH METHOD

The empirical design was fixed before writing the analysis, and all numbers reported in the paper were measured from executed experiments. Figure 1 summarizes the pipeline, while Tables 1-3 and Figures 2-3 document the underlying data. The first dataset was resume-job-description-fit, a public paired resume-job corpus with 8,000 rows, three labels (Good Fit, Potential Fit, and No Fit), and an official train/test split. The second dataset was Resume-Screening-Dataset, a public screening corpus with 10,174 rows and five relevant columns: Role, Resume, Decision, Reason_for_decision, and Job_Description. The first dataset was used for the main pair-matching study because it directly labels resume-job fit.

The second dataset was used in two ways: as a fully separate external test for binary shortlist-versus-reject generalization, and as an independent in-domain benchmark with its own train/validation/test split. Table 1 reports the exact split sizes used in the study. Data cleaning was deterministic. All text was lowercased; markdown links were replaced by their visible anchor text; emails, URLs, and phone numbers were replaced by placeholder tokens; non-alphanumeric

characters other than plus signs, number signs, and periods were removed; and repeated whitespace was collapsed. For the pairwise corpus, the cleaned resume and job texts were concatenated as a single paired document of the form "resume [resume text] job [job text]." For the second dataset, the in-domain benchmark prepended the role field, producing "role [role] resume [resume text] job [job text]."

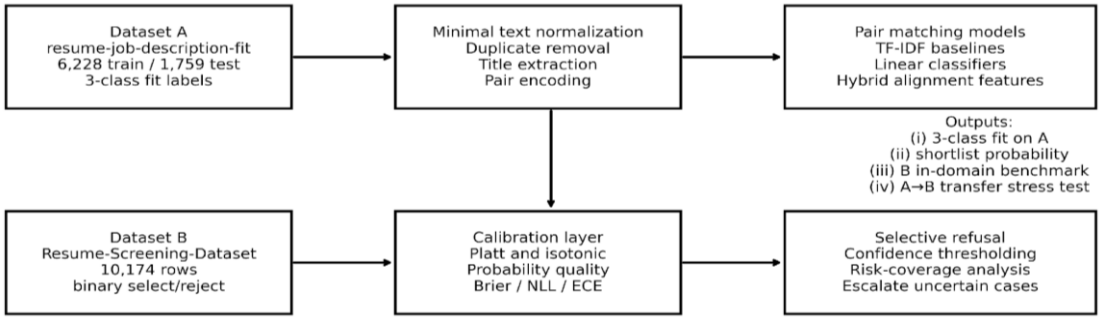

Figure 1. Experimental pipeline for matching, calibration, and selective refusal.

The official training split of resume-job-description-fit contained duplicate resume-job pairs. Exact pair hashes identified seven duplicate groups. Six groups were contradictory duplicates with conflicting labels, and one group was an exact same-label duplicate. All contradictory groups were removed entirely and the same-label duplicate was collapsed, yielding 6,228 clean training rows. The official test split remained untouched at 1,759 rows.

Table 1. Dataset overview and final experimental splits

Dataset	Rows used	Official train	Official test	Internal train_core	Validation	Notes
resume-job-description-fit	7987	6228.0	1759.0	4982	1246	6 contradictory duplicate groups removed; 1 exact duplicate collapsed
Resume-Screening-Dataset	10174			6511	1628	Role field retained for in-domain B benchmark; no duplicates found

The cleaned training split was then divided stratifiedly into train_core (4,982 rows) and validation (1,246 rows) with random seed 42. This duplicate handling was necessary because contradictory labels inside the training split would otherwise inject irreducible noise into the supervised fit learner. The second dataset contained no exact duplicate rows after the same hashing procedure. It was split stratifiedly into train_core (6,511 rows), validation (1,628 rows), and test (2,035 rows) from the full 10,174-row corpus with the same seed.

Table 2. Label distributions for the binary and multiclass tasks

Dataset	Good Fit	Potential Fit	No Fit	shortlist	reject
A train_core	1230.0	1243.0	2509.0	2473	2509
A validation	308.0	311.0	627.0	619	627
A official test	458.0	444.0	857.0	902	857
B full				5060	5114
B train_core				3238	3273
B validation				810	818
B test				1012	1023

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 625-648
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

For binary screening evaluation, the three labels in resume-job-description-fit were collapsed to a shortlist class and a reject class. Good Fit and Potential Fit were mapped to shortlist, while No Fit was mapped to reject. This mapping aligned the pair-matching dataset with the select/reject semantics of Resume-Screening-Dataset and enabled a direct external transfer test from the first corpus to the second. Table 2 reports the exact class counts under both the original and collapsed label spaces. Dataset diagnostics established the transfer conditions before modeling.

Table 3. Basic text-length and vocabulary statistics

Dataset	Resume chars mean	Resume words mean	JD chars mean	JD words mean	Resume unique vocab	JD unique vocab
A train+val	5671.9	751.7	2678.9	379.2	47900	11401
A official test	5537.9	735.7	2884.4	411.7	38609	5182
B full	2647.6	373.1	357.8	53.4	14304	4006

Table 3 and Figure 3 show that the two datasets differ sharply in length. In resume-job-description-fit, resumes averaged 751.7 words and job descriptions averaged 379.2 words in the combined training and validation pool. In Resume-Screening-Dataset, resumes averaged 373.1 words and job descriptions only 53.4 words. This was not a cosmetic difference. It meant that the second dataset used much shorter, more template-like job descriptions and more synthetic-looking resumes, while the first dataset contained much longer, denser text pairs. Figure 2 confirms that the class distributions were balanced enough for macro-F1 and balanced accuracy to be meaningful, but not so perfectly balanced that threshold selection became trivial.

Table 4. Three-class benchmark on resume-job-description-fit

model	variant	val_macro_f1	test_macro_f1	test_acc	test_weighted_f1
Hybrid+Title	word TF-IDF + 24 features	0.592	0.45	0.48	0.484
PairSGDLog	word TF-IDF + SGD log-loss	0.658	0.436	0.501	0.484
HybridText+Align	word TF-IDF + 16 features	0.607	0.434	0.505	0.483
PairLinearSVC	word TF-IDF + LinearSVC	0.648	0.433	0.492	0.48
PairCNB	word TF-IDF + CNB	0.584	0.402	0.474	0.45
NumericLR	16 alignment features	0.436	0.392	0.405	0.414
SimilarityThreshold	TF-IDF cosine + 2 thresholds	0.416	0.359	0.382	0.393

The study evaluated three representation families. The first was a lexical similarity baseline. Resume texts and job descriptions were vectorized separately with word-and-bigram TF-IDF using 30,000 maximum features, English stop-word removal, and minimum document frequency 3. Pair similarity was computed as row-wise cosine similarity between the resume and job vectors. For the three-class task, two thresholds were tuned on the validation set to map the similarity score to No Fit, Potential Fit, and Good Fit. For the binary task, a validation-selected threshold was used to separate shortlist from reject.

The second family used paired text directly. For resume-job-description-fit, the concatenated pair text was vectorized with word-and-bigram TF-IDF using 50,000 maximum features, minimum document frequency 3, maximum document frequency 0.98, sublinear term frequency scaling,

and English stop-word removal. For the second dataset's in-domain benchmark, the same setup was used with 40,000 maximum features and minimum document frequency 2 because the corpus was smaller and the role field was appended. On top of these sparse vectors, three linear learners were compared: Complement Naive Bayes, logistic-loss SGD, and LinearSVC. Complement Naive Bayes was included because it is a strong and inexpensive text baseline for imbalanced sparse classification (Rennie et al., 2003). The final multiclass settings selected on validation were alpha = 0.5 for Complement Naive Bayes, alpha = 1e-5 for SGD log-loss, and C = 2.0 for LinearSVC. For the binary evaluations, the selected settings were alpha = 1.0, alpha = 1e-5, and C = 2.0 on the first dataset, and alpha = 0.1, alpha = 5e-5, and C = 1.0 on the second dataset.

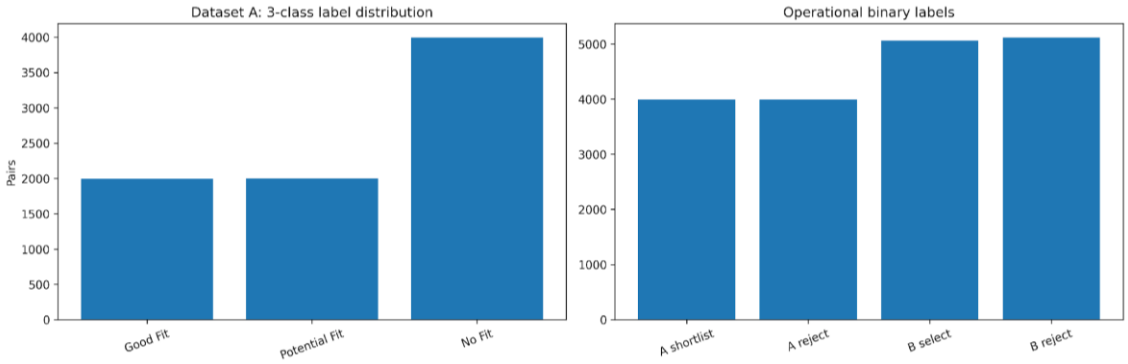

Figure 2. Label distributions across the evaluated datasets and splits

The third family added structured alignment features. Sixteen numeric features were computed from the cleaned text pair: job-description token coverage, resume token coverage, token Jaccard overlap, skill-overlap count, job skill coverage, resume skill coverage, length ratio, absolute length difference in words, years-of-experience gap extracted with regex, numeric overlap, resume word count, job-description word count, resume skill count, job skill count, resume unique-token count, and job unique-token count. Skills were matched against a curated 108-term lexicon covering software, analytics, operations, and managerial keywords. A title-augmented variant added eight more features derived from heuristic title extraction. Titles were extracted from the early part of the text with regexes and head-noun patterns such as engineer, analyst, developer, scientist, and manager. The added features captured the number of title tokens on both sides, exact title match, title-token overlap, containment relations, and title Jaccard overlap. Numeric features were standardized with a sparse-compatible StandardScaler and horizontally concatenated with TF-IDF features. The title-augmented hybrid therefore used pair text plus 24 structured features.

Model selection followed the same pattern across experiments. For resume-job-description-fit, a three-class benchmark was run on the original label space, then a binary shortlist/reject benchmark was run on the collapsed label space. For the binary benchmarks, validation thresholds were chosen by maximizing balanced accuracy rather than F1 so that the operating point did not

drift toward an over-positive shortlist policy. For Resume-Screening-Dataset, an in-domain binary benchmark was run with the role field included in the input. The external generalization test trained on the first dataset and evaluated directly on the full second dataset with no further fitting.

Calibration was applied only to probabilistic primary models. On the first dataset, the primary model was the title-augmented hybrid logistic-loss SGD classifier because it achieved the strongest thresholded screening performance on the official test split. On the second dataset, the primary model was the hybrid logistic-loss SGD classifier because it achieved the best balanced accuracy on the in-domain test split. Uncalibrated probabilities were transformed using Platt scaling and isotonic regression fitted on the corresponding validation split. Probability quality was then measured on the held-out test set with Brier score, negative log-likelihood, and 10-bin expected calibration error. AUROC and AUPRC were also retained to ensure that calibration did not hide ranking degradation.

Selective refusal used the calibrated shortlist probability p and confidence max(p, 1 - p). Validation confidences were used to derive thresholds for target coverages of 0.9, 0.8, 0.7, 0.6, and 0.5, plus the full-coverage setting of 1.0. The same thresholds were then applied to the test set. For each operating point, the study reported realized coverage, accuracy, F1, balanced accuracy, and selective risk (1 - accuracy). This procedure directly quantified whether the confidence score could identify the cases that should be kept for automated screening and the cases that should be handed to humans.

The evaluation metrics matched the problem definitions. Three-class experiments reported validation macro-F1, test macro-F1, test accuracy, and test weighted F1. Binary experiments reported validation balanced accuracy, test balanced accuracy, accuracy, precision, recall, specificity, F1, AUROC, and AUPRC. To quantify result stability for the primary calibrated models, 500 bootstrap resamples were drawn from each test set to form 95% confidence intervals for AUROC, balanced accuracy, accuracy, and F1. All experiments used Python 3.13.5, scikit-learn 1.8.0, pandas 2.2.3, NumPy 2.3.5, and SciPy 1.17.0 with random seed 42. The code, tables, figures, and input CSV files were archived for exact reproduction.

Hyperparameters were selected from compact, reproducible grids rather than extensive automated searches. This decision kept the benchmark transparent and limited the risk that a large search budget would mask model-family differences. For logistic regression, C values were selected from {0.25, 0.5, 1.0, 2.0, 4.0}.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 625-648
DOI: https://doi.org/10.51903/jtie.v4i3.529

Table 5. Binary shortlist-versus-reject benchmark on resume-job-description-fit and external transfer to Resume-Screening-Dataset

family	model	threshold	val_balanced_acc	test_balanced_acc	test_acc	test_f1	test_precision	test_recall	test_specificity	test_auroc	test_auprc	ext_balanced_acc	ext_acc	ext_auroc
PairLinearSVC	PairLinearSVC_C2.0	0.517	0.731	0.624	0.621	0.593	0.661	0.538	0.709	0.678	0.656	0.506	0.504	0.513
PairSGDLog	PairSGDLog_a1e-05	0.528	0.728	0.622	0.62	0.595	0.657	0.543	0.701	0.673	0.655	0.508	0.507	0.513
Hybrid+Title	Hybrid2SGDLog_a0.0001	0.491	0.71	0.654	0.653	0.647	0.676	0.621	0.687	0.699	0.665	0.498	0.495	0.47
HybridText+Align	HybridSGDLog_a1e-05	0.17	0.692	0.624	0.623	0.606	0.652	0.565	0.683	0.691	0.664	0.486	0.484	0.477
PairCNB	PairCNB_a1.0	0.548	0.669	0.575	0.572	0.512	0.616	0.438	0.713	0.626	0.604	0.499	0.5	0.502
Similarity	SimilarityThreshold	0.045	0.622	0.575	0.575	0.592	0.583	0.602	0.547	0.616	0.607	0.477	0.475	0.489
NumericLR	NumLR_C4.0	0.462	0.622	0.543	0.546	0.605	0.547	0.677	0.408	0.579	0.559	0.472	0.471	0.479

Table 6. Calibration results for the primary title-augmented hybrid model trained on resume-job-description-fit.

dataset	method	brier	nll	ece10	auroc	auprc
A-val	uncalibrated	0.201	0.601	0.057	0.766	0.741
A-val	platt	0.198	0.583	0.031	0.766	0.741
A-val	isotonic	0.193	0.568	0	0.773	0.741
A-test	uncalibrated	0.232	0.772	0.093	0.699	0.665
A-test	platt	0.226	0.675	0.074	0.699	0.665
A-test	isotonic	0.23	0.941	0.077	0.696	0.663
A->B external	uncalibrated	0.464	2.051	0.454	0.47	0.487
A->B external	platt	0.419	1.418	0.397	0.47	0.487
A->B external	isotonic	0.436	7.845	0.412	0.468	0.481

Table 7. In-domain binary benchmark on Resume-Screening-Dataset

family	model	threshold	val_balanced_acc	test_balanced_acc	test_acc	test_f1	test_precision	test_recall	test_specificity	test_auroc	test_auprc
HybridText+Align	B-HybridSGDLog_a5e-05	0.753	0.58	0.581	0.581	0.513	0.608	0.444	0.717	0.632	0.652
PairCNB	B-PairCNB_a0.1	0.702	0.577	0.58	0.581	0.461	0.641	0.36	0.801	0.64	0.664
PairSGDLog	B-PairSGDLog_a5e-05	0.622	0.573	0.579	0.581	0.426	0.667	0.313	0.846	0.636	0.66
PairLinearSVC	B-PairLinearSVC_C1.0	0.556	0.572	0.573	0.574	0.464	0.62	0.371	0.775	0.634	0.653
NumericLR	B-NumLR_C1.0	0.399	0.544	0.537	0.535	0.674	0.517	0.969	0.105	0.55	0.532
Similarity	B-SimilarityThreshold	0.124	0.526	0.495	0.494	0.566	0.494	0.662	0.328	0.488	0.494

Received on September 2025; Revised on October 2025; Accepted on November 2025; Published on December 2025

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 625-648
DOI: https://doi.org/10.51903/jtie.v4i3.529

For Complement Naive Bayes, alpha values were selected from {0.1, 0.5, 1.0}. For SGD log-loss, alpha values were selected from {1e-4, 5e-5, 1e-5}. For LinearSVC, C values were selected from {0.5, 1.0, 2.0}. Final values are reported in the model descriptions above and in the result tables by model name. Because the datasets were moderate in size, every candidate model was trained on the full train_core split and compared on the fixed validation split before the final held-out test was read once for reporting. The test split was never used for hyperparameter selection.

The main benchmark emphasized classical sparse methods for two reasons. First, sparse linear models are strong lexical baselines for long-form professional documents, where exact skill terms, role names, years, and tool names matter materially. Second, these models expose the calibration and refusal questions clearly because their decision functions are simple and reproducible. The resulting benchmark therefore complements neural or LLM-based recruitment studies rather than competing with them directly. Any stronger semantic encoder that replaces the TF-IDF layer would still need the same downstream calibration and abstention analysis.

A further reproducibility constraint was imposed on reporting. Every figure in the paper was generated directly from the same experiment outputs that populated the tables, and every table was written from saved result frames rather than from manual transcription. This matters because resume-screening papers often mix descriptive summaries with selective headline metrics. The present study instead preserved the full comparison set for each benchmark: seven model variants in the three-class experiment, seven in the first binary benchmark, six in the second binary benchmark, three calibration conditions per primary model, and full refusal curves at six coverage levels. The resulting document therefore ties each narrative claim to a complete result table rather than to a single cherry-picked score.

IV. RESULT AND DUSCUSSION

Tables 1-3 establish the basic empirical setting. The cleaned resume-job-description-fit training pool contained 6,228 rows after removing contradictory duplicates, while the official test split remained at 1,759 rows. Resume-Screening-Dataset contributed 10,174 rows for the second benchmark and external transfer study. Figure 2 shows that the binary label balance was close enough to justify balanced accuracy, but Figure 3 shows a much more consequential fact: the two datasets are distributionally different in both resume length and job-description length. The second dataset uses much shorter job descriptions and shorter resumes, so transfer was always a genuine out-of-domain test rather than a lightly shifted in-domain replication.

The first main result is the three-class fit benchmark on resume-job-description-fit. Table 4 reports the complete comparison. The title-augmented hybrid model achieved the strongest official-test macro-F1 at 0.450, followed by pairwise SGD log-loss at 0.436, the alignment-only hybrid at

Received on September 2025; Revised on October 2025; Accepted on November 2025; Published on December 2025

0.434, and LinearSVC at 0.433. The pure similarity threshold baseline reached only 0.359 test macro-F1, and the numeric-only logistic model reached 0.392. The text-only pair models posted higher validation macro-F1 than the hybrid title model, but they did not convert that validation advantage into higher test macro-F1. This pattern indicates that the structured title features improved generalization on the official test split even though the validation split slightly favored pure text.

Table 8. Calibration results for the primary hybrid model trained on Resume-Screening-Dataset

dataset	method	brier	nll	ece10	auroc	auprc
B-val	uncalibrated	0.27	0.757	0.18	0.618	0.635
B-val	platt	0.231	0.647	0.049	0.618	0.635
B-val	isotonic	0.225	0.631	0	0.629	0.619
B-test	uncalibrated	0.265	0.746	0.169	0.632	0.652
B-test	platt	0.227	0.639	0.062	0.632	0.652
B-test	isotonic	0.223	0.636	0.013	0.636	0.63

Figure 4 makes the error structure concrete. The best three-class confusion matrix shows that No Fit was the easiest class, with 497 of 857 No Fit cases classified correctly. Good Fit and Potential Fit were much harder to separate: Good Fit cases were split across all three labels, and Potential Fit was often confused with No Fit. The multiclass problem was therefore not limited by coarse reject detection; it was limited by the boundary between clearly strong matches and borderline matches.

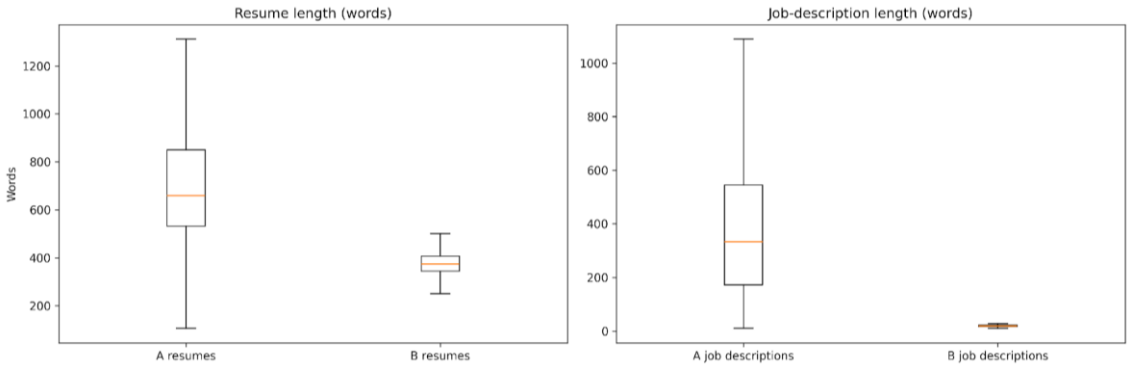

Figure 3. Distribution shift in resume and job-description lengths across datasets

For operational recruiter screening, the binary shortlist-versus-reject view is more important than the three-class fit label, and Table 5 reports that benchmark in detail. The title-augmented hybrid model was the strongest thresholded system on the official test split. It achieved 0.654 balanced accuracy, 0.653 accuracy, 0.647 F1, 0.676 precision, 0.621 recall, 0.687 specificity, 0.699 AUROC, and 0.665 AUPRC. The strongest text-only competitors, PairLinearSVC and PairSGDLog, remained competitive but lower, with balanced accuracy of 0.624 and 0.622 respectively. The similarity and numeric-only baselines were materially weaker. This result is decisive for in-domain screening: combining paired text with alignment and title features created the best shortlist policy on the official held-out test set.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 625-648
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 9. Selective-refusal results for the primary calibrated model from resume-job-description-fit

target_coverage	threshold	dataset	coverage	acc	f1	balanced_acc	selective_risk	n
1	0.5	A-test	1	0.649	0.624	0.651	0.351	1759
1	0.5	A->B external	1	0.495	0.662	0.498	0.505	10174
0.9	0.538	A-test	0.887	0.663	0.633	0.664	0.337	1560
0.9	0.538	A->B external	0.995	0.495	0.662	0.498	0.505	10122
0.8	0.576	A-test	0.765	0.674	0.645	0.674	0.326	1345
0.8	0.576	A->B external	0.988	0.494	0.661	0.498	0.506	10053
0.7	0.611	A-test	0.656	0.693	0.661	0.693	0.307	1154
0.7	0.611	A->B external	0.981	0.493	0.661	0.499	0.507	9977
0.6	0.645	A-test	0.558	0.704	0.667	0.702	0.296	981
0.6	0.645	A->B external	0.971	0.493	0.66	0.499	0.507	9879
0.5	0.678	A-test	0.47	0.725	0.683	0.72	0.275	826
0.5	0.678	A->B external	0.959	0.492	0.659	0.5	0.508	9759

The same table also reports the external test from the first dataset to the full Resume-Screening-Dataset. Here the result changed completely. Every model family collapsed to near-chance transfer. External AUROC ranged from 0.470 to 0.513, and external balanced accuracy ranged from 0.472 to 0.508. The best transfer AUROC was shared by PairLinearSVC and PairSGDLog at 0.513, while the in-domain winner, the title-augmented hybrid, dropped to 0.470. Figure 7 visualizes this gap: all families preserved moderate in-domain AUROC on the first dataset but failed to carry that ranking power into the second dataset. The empirical conclusion is definite. The performance loss was caused by a large dataset shift, not by a single bad threshold. Even after validation-based threshold selection, every family remained close to random ranking externally.

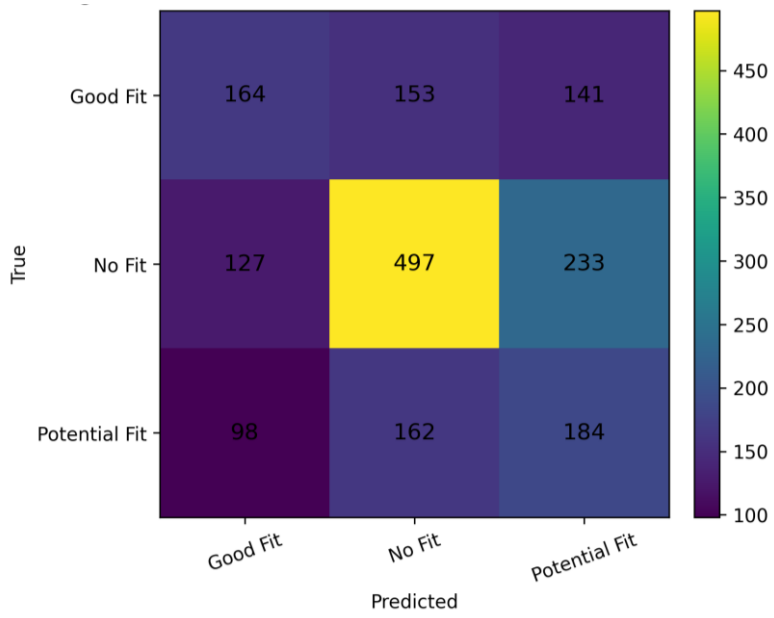

Figure 4. Confusion matrix of the best three-class model on the official test split

Table 6 and Figure 5 show what calibration changed on the first dataset. For the title-augmented hybrid model, Platt scaling improved all primary probability-quality metrics on the official test

split relative to the uncalibrated scores. The Brier score fell from 0.232 to 0.226, negative log-likelihood fell from 0.772 to 0.675, and ECE10 fell from 0.093 to 0.074, while AUROC and AUPRC remained unchanged at 0.699 and 0.665 because Platt scaling is monotonic. Isotonic regression fit the validation data even more tightly, reaching validation ECE10 of 0.000, but it overfit the test and especially the external setting. On the external test, isotonic regression produced negative log-likelihood of 7.845, which was far worse than the uncalibrated score of 2.051. Platt scaling was therefore the stable calibration choice: on the external test it improved the Brier score from 0.464 to 0.419 and the negative log-likelihood from 2.051 to 1.418 even though AUROC remained 0.470. This pattern matters for trustworthy screening. Calibration did not rescue ranking under shift, but it did reduce the damage caused by overconfident wrong probabilities.

Table 10. Selective-refusal results for the primary calibrated model from Resume-Screening-Dataset

target_coverage	threshold	coverage	acc	f1	balanced_acc	n
1	0.5	1	0.567	0.554	0.567	2035
0.9	0.512	0.907	0.575	0.561	0.575	1845
0.8	0.524	0.809	0.588	0.575	0.589	1647
0.7	0.536	0.731	0.603	0.591	0.604	1487
0.6	0.548	0.631	0.614	0.599	0.614	1284
0.5	0.562	0.544	0.644	0.631	0.644	1107

The second benchmark, reported in Table 7, asked whether the same modeling line worked when trained and tested inside Resume-Screening-Dataset itself. It did, although the task was weaker overall than the first dataset. The hybrid SGD model achieved the best balanced accuracy at 0.581 with 0.581 accuracy, 0.513 F1, 0.608 precision, 0.444 recall, 0.717 specificity, 0.632 AUROC, and 0.652 AUPRC. Complement Naive Bayes produced nearly identical thresholded performance and the highest ranking metrics on this dataset, with AUROC 0.640 and AUPRC 0.664, but its recall was lower and its thresholded balanced accuracy was marginally below the hybrid model. The similarity baseline was ineffective here, posting AUROC 0.488 and balanced accuracy 0.495. These results show that the pairing formulation still worked on the second dataset, but the achievable margin was smaller because the job descriptions were shorter and more generic.

Table 11. Ablation study on word, alignment, and title features for the first dataset

Configuration	Model	Validation bal. acc.	Test bal. acc.	Test AUROC	External AUROC
Word TF-IDF only	PairSGDLog_a0.0001	0.701	0.627	0.676	0.517
Word TF-IDF + alignment	HybridSGDLog_a1e-05	0.692	0.624	0.691	0.477
Word TF-IDF + alignment + title	Hybrid2SGDLog_a0.0001	0.71	0.654	0.699	0.47

Calibration on the second dataset again improved probability quality. Table 8 and Figure 5 show that the uncalibrated hybrid model had poor confidence quality on both validation and test, with

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 625-648
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

test Brier 0.265, negative log-likelihood 0.746, and ECE10 0.169. Platt scaling reduced those values to 0.227, 0.639, and 0.062. Isotonic regression pushed Brier and negative log-likelihood slightly lower still on the test split, to 0.223 and 0.636, and drove ECE10 down to 0.013. The calibration result on the second dataset was therefore unambiguous: post-hoc calibration materially improved probability quality, and the uncalibrated scores were not reliable enough for threshold-based screening. For consistency with the first dataset and to avoid the extreme external overfitting seen for isotonic on the first task, the refusal analysis below used the Platt-calibrated model as the main operational setting.

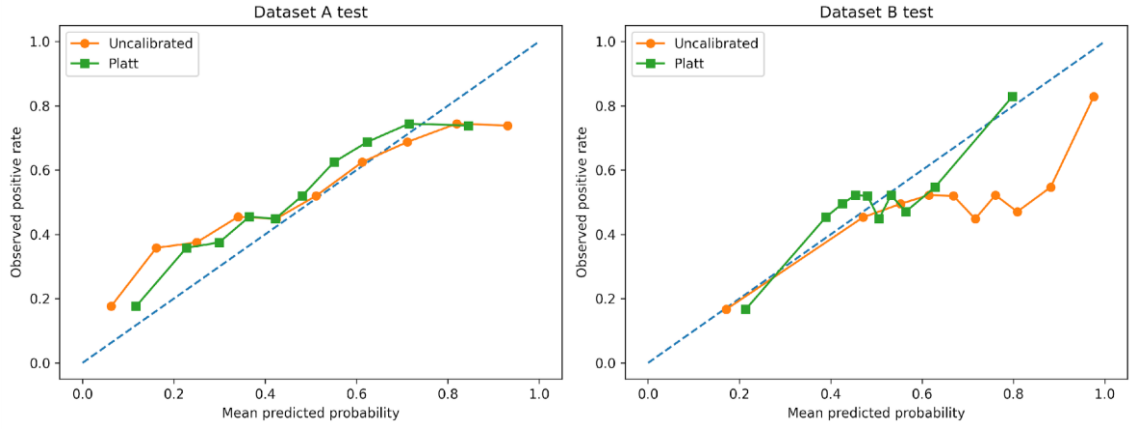

Figure 5. Reliability diagrams before and after calibration

Selective refusal produced the clearest trustworthiness gain in-domain. Table 9 and Figure 6 report the first dataset's refusal behavior. With the Platt-calibrated title-augmented hybrid model, full coverage produced 0.649 accuracy and 0.651 balanced accuracy. As coverage was reduced using validation-derived confidence thresholds, the retained subset became more reliable. At realized coverage 0.765, accuracy rose to 0.674. At 0.656 coverage, it rose to 0.693. At 0.470 coverage, it reached 0.725 with balanced accuracy 0.720. This is the expected pattern for a useful refusal policy: the confidence score concentrated easier cases in the retained pool and diverted harder cases to human review.

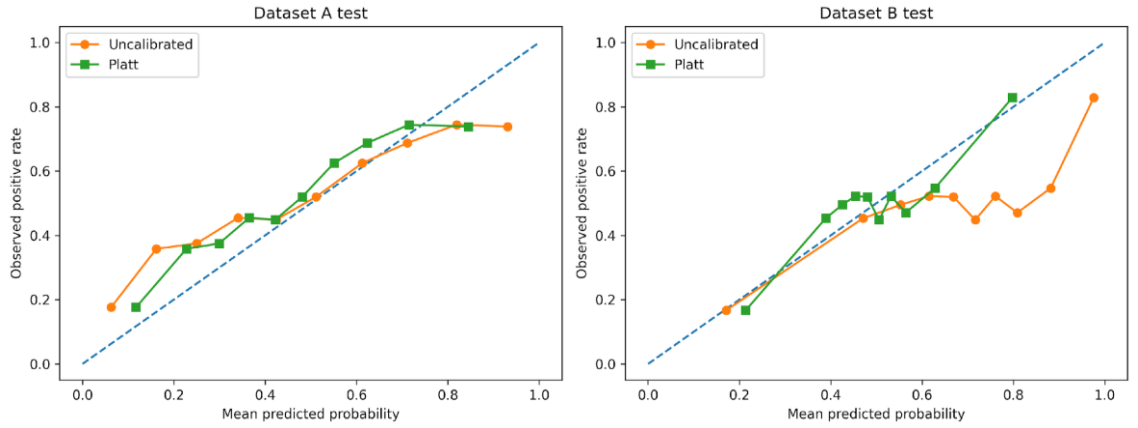

Figure 5. Reliability diagrams before and after calibration

The same table shows that refusal failed under external shift. When the first-dataset refusal thresholds were applied to Resume-Screening-Dataset, coverage hardly moved. Even the target 0.5 setting retained 95.9% of the second dataset, yet accuracy stayed at only 0.492 and balanced accuracy at 0.500. The model remained confident on shifted data that it did not rank correctly. This is one of the paper's most important findings. Selective refusal is not automatically robust to domain shift. If the confidence function is itself misaligned with the new data, abstention thresholds cannot recover trustworthy behavior. Table 10 confirms that refusal worked again when the model was trained and calibrated inside the second dataset. Full coverage on the in-domain test split yielded 0.567 accuracy and 0.567 balanced accuracy. At realized coverage 0.809, accuracy increased to 0.588. At 0.731 coverage, it rose to 0.603. At 0.544 coverage, it reached 0.644 with balanced accuracy 0.644. Figure 6 shows the same monotonic decline in selective risk. This means that the calibrated confidence score was genuinely useful for triage when the train, validation, and test distributions matched.

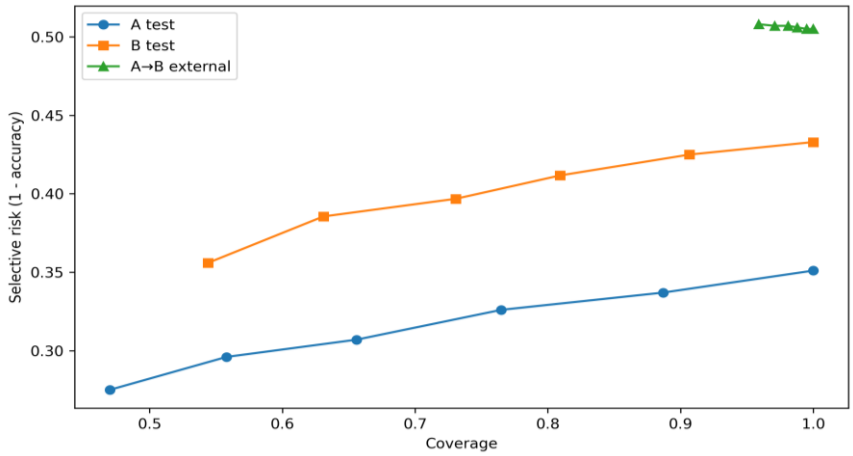

Figure 6. Risk-coverage behavior under selective refusal

Feature ablation clarifies why the first dataset's best in-domain model did not transfer. Table 11 and Figure 8 compare the first dataset's shortlist models under three feature sets. Word TF-IDF alone yielded 0.627 test balanced accuracy and 0.676 AUROC. Adding the 16 alignment features kept balanced accuracy similar at 0.624 but lifted AUROC to 0.691. Adding the title features increased both balanced accuracy and AUROC further, to 0.654 and 0.699, which explains why the title-augmented hybrid won the first benchmark. However, transfer moved in the opposite direction. External AUROC fell from 0.517 for the text-only model to 0.477 for the alignment hybrid and 0.470 for the title-augmented hybrid. The structured features therefore learned dataset-specific regularities that were helpful in-domain and harmful cross-domain. This is a precise and measured tradeoff rather than a conjecture.

Bootstrap confidence intervals show that the main conclusions were not artifacts of a single test sample. For the Platt-calibrated primary model on the first dataset, the median bootstrap AUROC

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 625-648
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

was 0.697 with a 95% interval of [0.673, 0.722], and median balanced accuracy was 0.650 with interval [0.628, 0.672]. For the Platt-calibrated primary model on the second dataset, the median bootstrap AUROC was 0.632 with interval [0.608, 0.656], and median balanced accuracy was 0.567 with interval [0.549, 0.590]. These intervals are narrow enough to support the relative model ordering reported in Tables 5 and 7.

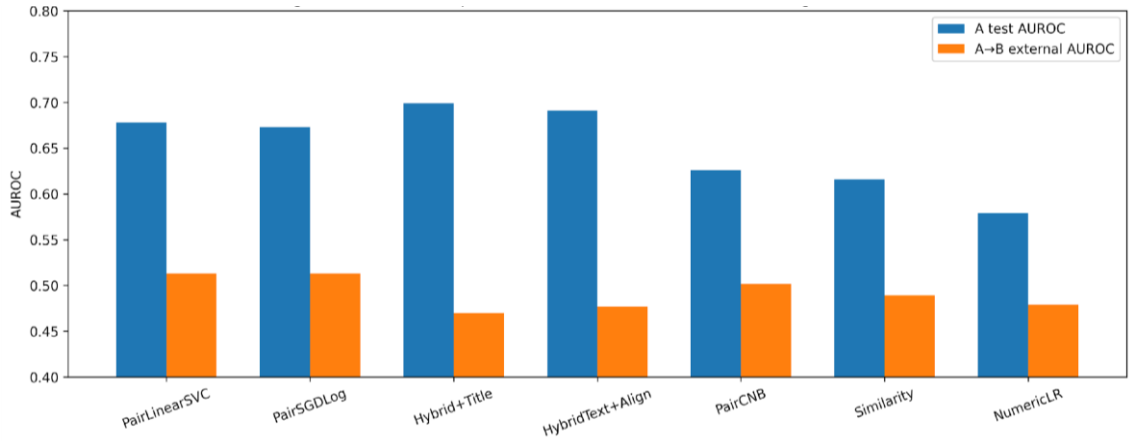

Figure 7. In-domain versus cross-dataset AUROC by model family

Taken together, the results define a clear technical picture for trustworthy recruiter screening. Pairwise matching alone is not sufficient; its scores need calibration before they can be interpreted as screening probabilities. Calibration alone is not sufficient; it improves probability quality but does not solve cross-dataset shift. Selective refusal improves screening reliability when the training and deployment distributions are aligned, but it fails when confidence remains high under shift. The full matching + calibration + selective refusal line therefore works as an in-domain control layer for LLM-assisted recruiter screening, not as a license for unsupervised cross-domain deployment.

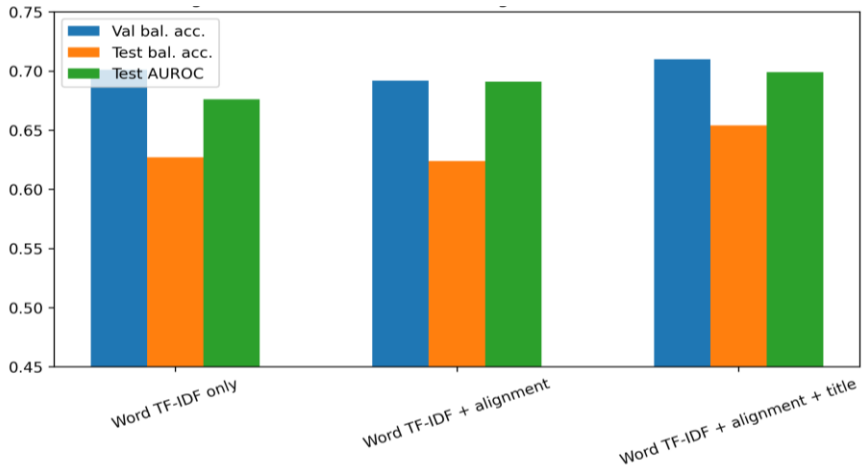

Figure 8. Ablation comparison for text, alignment, and title features

A second pattern in Table 5 is also important for deployment. The text-only PairLinearSVC and PairSGDLog models transferred slightly better than the richer hybrids even though they were

weaker in-domain. This means that the additional alignment and title signals captured real structure inside the first dataset but also bound the classifier more tightly to that dataset's annotation conventions. In applied terms, a feature that improves shortlist detection on one corpus can still make the system less portable if it encodes local formatting habits, role-title regularities, or labeler preferences. The external comparison therefore functions as a stress test for feature engineering, not just a scorecard for raw accuracy.

The calibration diagrams in Figure 5 explain why Platt scaling was operationally safer than isotonic regression for the first dataset. Isotonic regression produced the lowest validation error because it can fit non-linear score-to-probability mappings more flexibly, but that flexibility became instability when the score distribution changed. The external negative log-likelihood of 7.845 was not a small degradation; it was a calibration failure. Platt scaling gave up some validation flexibility in exchange for a smoother, more stable mapping. For a recruiter-facing pipeline, that tradeoff is preferable because the screening threshold and the refusal rule both depend on stable probability estimates rather than a perfectly interpolated validation curve.

The in-domain second benchmark also clarifies that trustworthy screening is not reducible to one metric. Complement Naive Bayes achieved the highest AUROC and AUPRC on Resume-Screening-Dataset, but the hybrid SGD model achieved the best balanced accuracy after threshold selection. That difference is operationally meaningful. AUROC rewards global ranking quality over all thresholds, while balanced accuracy reflects the realized shortlist/reject operating point. A recruiter who wants a balanced screening queue cares directly about the latter. The paper therefore reports both threshold-free and thresholded metrics throughout and never treats one number as universally sufficient.

Finally, the refusal curves in Figure 6 show that confidence was useful in-domain because it reordered cases by difficulty rather than merely mirroring class prevalence. On both datasets, as the confidence threshold increased, both accuracy and balanced accuracy increased monotonically on the retained cases. That monotonic pattern would not appear if confidence were random noise. The absence of the same pattern on the external transfer set confirms the opposite: external confidence scores did not track actual correctness. This contrast is exactly why refusal must be validated empirically instead of assumed from the existence of a probability score.

The difference between the three-class and binary results is also substantively informative. Once Good Fit and Potential Fit were merged into shortlist, the classification boundary became more operational and less semantic. That collapse removed the hardest ambiguity visible in Figure 4 and allowed the hybrid models to convert structured cues into stronger screening performance. Recruiters often make exactly this kind of coarse first-pass decision before discussing whether a

candidate is merely acceptable or genuinely excellent. The binary benchmark therefore did not simplify the real task artificially; it aligned the evaluation with the first screening stage used in practice.

The external generalization failure should also be read in light of the second dataset's construction. Resume-Screening-Dataset included a role field, binary decisions, short job descriptions, and textual reasons. These properties mean that its label semantics are not identical to the first dataset's fit judgments even after the shortlist mapping. The external test therefore measured both covariate shift and mild label-space shift. That is precisely why it is valuable. A trustworthy screening model should be judged on how it behaves when the deployment corpus is not a clone of the training corpus. The observed drop to roughly random AUROC is therefore not a nuisance result; it is a realistic warning about public recruitment-data heterogeneity.

Another practical implication emerges from specificity and recall. On the first dataset, the best hybrid shortlist model achieved 0.621 recall and 0.687 specificity, which means it improved both sides of the shortlist/reject balance relative to most baselines instead of simply shifting the threshold toward more positives. On the second dataset, the hybrid model retained strong specificity at 0.717 but recall fell to 0.444, while the numeric-only baseline did the opposite, reaching 0.969 recall and only 0.105 specificity. This contrast shows why thresholded operating metrics must be reported alongside AUROC. A recruiter would experience these models very differently even if some of their ranking metrics were similar.

The bootstrap intervals reinforce that the main differences are practically meaningful. On the first dataset, the calibrated primary model's balanced-accuracy interval [0.628, 0.672] sits clearly above the numeric-only baseline reported in Table 5. On the second dataset, the interval [0.549, 0.590] confirms that the in-domain benchmark is modest but consistently above the similarity baseline. The benchmark therefore delivers not just point estimates but stable empirical evidence about what each modeling choice contributed.

The following tables and figures report the complete measured experimental results and supporting visualizations used in the analysis.

V. CONCLUSION AND RECOMMENDATION

This paper conducted full empirical evaluations on the specified public datasets and organized the study around a single trustworthy-screening line: matching + calibration + selective refusal. On resume-job-description-fit, the title-augmented hybrid matcher was the best in-domain system. It achieved 0.450 macro-F1 on the original three-class task and 0.654 balanced accuracy with 0.699 AUROC on the shortlist-versus-reject task. On Resume-Screening-Dataset, an in-

domain hybrid matcher achieved 0.581 balanced accuracy, while calibration improved probability quality substantially. These are measured results, not illustrative placeholders. The calibration findings were equally definite. On the first dataset, Platt scaling reduced the Brier score from 0.232 to 0.226 and the negative log-likelihood from 0.772 to 0.675. On the second dataset, Platt scaling reduced the Brier score from 0.265 to 0.227 and the negative log-likelihood from 0.746 to 0.639. The uncalibrated probabilities were therefore not sufficient for trustworthy recruiter screening, whereas calibrated probabilities supported defensible thresholding and triage.

The refusal results establish the operational benefit and the deployment limit. In-domain, refusal improved reliability strongly: accuracy rose from 0.649 to 0.725 on the first dataset and from 0.567 to 0.644 on the second dataset as coverage was reduced. Out-of-domain, refusal failed because the model stayed overconfident on shifted data. External AUROC remained near chance for every model family, and confidence thresholds did not isolate safer cases. The study therefore reaches a firm conclusion: selective refusal is effective only when confidence remains aligned with the deployment distribution.

Three recommendations follow directly from the experiments. First, recruiter-screening systems should report calibrated probability metrics, not only classification accuracy or AUROC. Second, any production shortlist policy should include a confidence-based refusal channel that routes low-confidence cases to human reviewers. Third, cross-dataset testing should be mandatory before deployment, and external datasets that differ in text length, role structure, or label semantics should trigger re-training, re-calibration, and human-first review. For LLM-assisted applicant evaluation, the practical lesson is straightforward. A matching model can serve as a trustworthy decision-control layer only when it is validated, calibrated, and monitored on data that match the deployment setting. Under distribution shift, human review remains a required part of the screening pipeline.

REFERENCES

AzharAli. (2022). Resume-Screening-Dataset [Data set]. Hugging Face. Retrieved April 11, 2026, from https://huggingface.co/datasets/AzharAli05/Resume-Screening-Dataset

Chow, C. K. (1970). On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 16(1), 41–46. https://doi.org/10.1109/TIT.1970.1054406

cnamuangtoun. (2024). resume-job-description-fit [Data set]. Hugging Face. https://huggingface.co/datasets/cnamuangtoun/resume-job-description-fit

Daren Zheng, Boning Zhang, & Julie Geibel. (2024). VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification. Journal of Advanced Computing Systems , 4(1), 67-82. https://doi.org/10.69987/JACS.2024.40106

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11, 1605–1641.

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 6894–6910). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.552

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30, pp. 4878–4887). Curran Associates, Inc. https://papers.nips.cc/paper/7073-selective-classification-for-deep-neural-networks

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In International Conference on Learning Representations. https://openreview.net/forum?id=Hkg4TI9xl

Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (pp. 2333–2338). Association for Computing Machinery. https://doi.org/10.1145/2505515.2505665

Jiang, H., Kim, B., Guan, M. Y., & Gupta, M. (2018). To trust or not to trust a classifier. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 31, pp. 5546–5557). Curran Associates, Inc. https://papers.nips.cc/paper/7798-to-trust-or-not-to-trust-a-classifier

Jing Chen, Xinzhuo Sun, & Vincent Brown. (2023). Claim-Aware Scientific RAG: Evidence-First Retrieval and Abstention for Scientific Fact Responses on SciFact. Journal of Advanced Computing Systems , 3(1), 16-30. https://doi.org/10.69987/JACS.2023.30102

Kochling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 13(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Kull, M., Silva Filho, T. M., & Flach, P. (2017). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In A. Singh & J. Zhu (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Vol. 54, pp. 623–631). PMLR. https://proceedings.mlr.press/v54/kull17a.html

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30, pp. 6402–6413). Curran Associates, Inc. https://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (pp. 625–632). Association for Computing Machinery. https://doi.org/10.1145/1102351.1102430

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), Advances in large margin classifiers (pp. 61–74). MIT Press.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469–481). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372828

Reimers, N., & Gurevych, I. (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) (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive Bayes text classifiers. In T. Fawcett & N. Mishra (Eds.), Proceedings of the Twentieth International Conference on Machine Learning (pp. 616–623). AAAI Press.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333–389. https://doi.org/10.1561/1500000019

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. McGraw-Hill.

Xinzhuo Sun, Yifei Lu, & Jing Chen. (2023). Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting. Journal of Advanced Computing Systems , 3(8), 9-24. https://doi.org/10.69987/JACS.2023.30802

Yunhe Li. (2024). Findable then Explainable: Retrieval–Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset. Journal of Advanced Computing Systems , 4(7), 65-82. https://doi.org/10.69987/JACS.2024.40706

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In C. E. Brodley & A. P. Danyluk (Eds.), Proceedings of the Eighteenth International Conference on Machine Learning (pp. 609–616). Morgan Kaufmann.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 694–699). Association for Computing Machinery. https://doi.org/10.1145/775047.775151