

An Internet Search--Augmented Soft Prompt Tuning Framework for Explainable Resume Screening in Healthcare Recruitment

Yanting Wu

Shanghai Chest Hospital, Shanghai Jiao Tong University

Tianao Hu

Shanghai University

Siyu Gu

Shanghai Chest Hospital, Shanghai Jiao Tong University

Yilin Mao

Shanghai Chest Hospital, Shanghai Jiao Tong University

Xingxing Cen

`cxx2347@163.com`

Shanghai Chest Hospital, Shanghai Jiao Tong University

Article

Keywords: large language model, health care recruitment, person–job fit, soft prompt tuning, evidence grounding, explainable ranking

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11034303/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

An Internet Search–Augmented Soft Prompt
Tuning Framework for Explainable Resume
Screening in Healthcare Recruitment

Yanting Wu^{1†}, Tianao Hu^{2†}, Siyu Gu¹, Yilin Mao^{1*},
Xingxing Cen^{1*}

¹Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai,
China.

²Shanghai University, Shanghai, China.

*Corresponding author(s). E-mail(s):
linlin_harvest@sina.com (Yilin Mao);
cxx2347@163.com (Xingxing Cen);

[†]These authors contributed equally to this work.

Abstract

Background: Large-scale hospital recruitment requires accurate and auditable matching between heterogeneous resume evidence and job-specific requirements. **Methods:** This study developed an internet search–augmented, soft-prompt-tuned large language model (LLM) framework integrating structured resume extraction, targeted retrieval of publication evidence, continuous task-specific prompt embeddings attached to a frozen Qwen backbone, deterministic expert rules, and weighted ranking. Institutional recruitment records comprised 5,419 applications across 38 positions. Ranking experiments included nine positions with complete archived reference shortlists, encompassing 4,339 applications and 150 reference candidates. The pipeline-based model was compared with a Raw Prompt baseline across multiple manual-review pool sizes, and an ablation study of the abstract module was conducted. Model interpretability was evaluated based on field-level agreement with manually reviewed reference standards and correlations between job-specific module scores and archived expert shortlists. **Results:** The overall F1 score for key-information extraction was 98.45%. Across the nine positions, macro-average R-Precision increased from 61.18%

with the Raw Prompt baseline to 82.64% with the pipeline-based model (paired mean improvement, 21.46 percentage points; position-level bootstrap 95% confidence interval (CI), 8.73–37.51; exact Wilcoxon test, $P = 0.0078$). Macro-average NDCG@ R increased from 0.6632 to 0.8702 ($P = 0.0039$). When the manual-review pool was approximately $2R$, macro-average recall reached 97.83%. Removal of the publication-abstract module reduced macro-average R-Precision for research positions by 10.32 percentage points. The Research/Experience module showed the strongest association with archived expert shortlists (macro-average Spearman correlation across positions, $\rho = 0.587$; 95% CI, 0.521–0.642; $P = 0.002$).

Conclusions: The framework provides a valuable, scalable, and auditable decision-support system for healthcare recruitment: it can standardize resume evidence extraction, prioritize qualified candidates within a manageable review pool, and expose the evidence and module contributions underlying each ranking. It is designed to reduce repetitive screening effort while preserving accountable human oversight. External and prospective validation are required before deployment beyond the study institution.

Keywords: large language model, healthcare recruitment, person–job fit, soft prompt tuning, evidence grounding, explainable ranking

1 Background

Person–job fit is a core component of recruitment and aims to determine the degree of alignment between applicants’ qualifications, experience, professional profiles, and job requirements [1, 2]. This task is particularly complex in hospital recruitment: while processing large volumes of application materials that vary in format and content, recruiters must simultaneously evaluate professional licensure, clinical training, specialty experience, research output, and educational background [3]. Manual review is time-consuming and may be subject to inter-reviewer variability; although purely rule-based systems are efficient, they have limited ability to interpret implicit or context-dependent information.

Early automated systems primarily used keyword matching, support vector machines, random forests, and neural text classifiers [4–6]. These approaches improved evaluation consistency but typically relied on task-specific labeled data and often represented resumes as fixed feature vectors. Consequently, they may not generalize

reliably when equivalent qualifications are expressed using different wording or when
evaluation criteria vary substantially across hospital job categories.

Large language models provide an alternative adaptation mechanism. Continuous
soft prompts can transfer linguistic and general domain knowledge acquired dur-
ing pretraining to a downstream task while leaving the foundation-model parameters
unchanged [7–11]. This is attractive when only a limited amount of task-specific labeled
data is available. However, an unconstrained single-pass LLM call may combine extrac-
tion, interpretation, and scoring within the same process, making the output difficult
to audit. The static parametric knowledge encoded within a model may also be out-
dated or unverifiable, which is particularly problematic when academic publications
and professional evidence influence recruitment decisions [11, 12].

Accordingly, this study developed an internet search-augmented soft prompt
tuning framework for resume screening in healthcare recruitment. Its principal con-
tribution is a hybrid decision architecture in which a compact trainable soft prompt
adapts a frozen Qwen backbone, while document parsing, external evidence ground-
ing, deterministic rule assessment, weighted score fusion, and ranking remain explicit
and auditable. The framework was compared with a Raw Prompt baseline, evalu-
ated through an ablation study of the abstract module, and further examined using
module-level interpretability and statistical agreement analyses.

1.1 Automated Resume Screening and Person–Job Fit

Research on person–job fit has progressed from exact keyword matching to represen-
tation learning and bidirectional preference modeling [2, 4–6]. When sufficient labeled
data are available, conventional supervised-learning systems can achieve favorable
classification performance; however, the decision boundaries they learn are often dif-
ficult to transfer to new positions with different qualification requirements. Moreover,

recruitment decisions are inherently ranking problems rather than simple binary-classification tasks: a system must prioritize relevant candidates within a small manual-review pool while preserving a reasonable relative ordering.

1.2 Soft Prompt Tuning of Large Language Models

Soft prompt tuning represents a downstream task with a short sequence of continuous virtual-token embeddings learned from labeled data, while the original parameters of the foundation model remain frozen [9]. It differs from manually written hard prompts because the learned vectors are optimized through backpropagation and are not required to correspond to human-readable tokens. It also differs from full-parameter fine-tuning and LoRA because only the prompt embedding matrix is updated. This parameter-efficient design is suitable for hospital recruitment, where task-specific labels are limited and the same linguistic backbone must support positions with different qualification requirements. Human evaluation, transparent criteria, and safeguards against unsupported outputs nevertheless remain essential [10, 11, 13].

1.3 External Evidence Grounding and Explainable Decision Support

Retrieval-augmented generation improves performance on knowledge-intensive tasks by conditioning generation on externally retrieved evidence [12]. Tool-use approaches similarly allow models to request information from databases or web services during inference [14, 15]. In recruitment settings, restricted external evidence grounding can verify publication metadata and retrieve abstracts not provided in resumes. Interpretability should not rely solely on a model-generated explanation; decisions should be decomposable into input evidence, triggered rules, module scores, soft-prompt-derived

semantic contributions, and their effects on the final ranking. This study operationalized these requirements through an auditable soft-prompt-augmented rule-LLM scoring pipeline.

2 Methods

2.1 Study Design and System Overview

As presented in Figure 1, we introduce an internet search-augmented soft prompt tuning framework for candidate-position matching and ranking. The framework comprises three components. (I) Evidence construction and serialization, which organize the position description, structured resume fields, and publication evidence obtained through restricted title-based retrieval using explicit field identifiers; (II) Soft prompt adaptation, in which trainable continuous embeddings are prepended to each serialized instance and processed by a frozen Qwen backbone; and (III) Match estimation and hybrid ranking, in which candidate-level binary cross-entropy and within-position pairwise supervision jointly optimize the semantic matching signal. During inference, this signal is incorporated into the applicable job-specific module, combined with deterministic rules and predefined module weights, and then used to produce the final auditable score, ranking, and Top- K review pool described in Section 2.6.

This retrospective study was approved by the Ethics Committee of Shanghai Chest Hospital (approval no. KS26065). All methods were carried out in accordance with relevant guidelines and regulations and the Declaration of Helsinki. The requirement for informed consent was waived by the Ethics Committee of Shanghai Chest Hospital due to the retrospective nature of the study.

For clarity, the position description, structured resume, and retrieved publication evidence for candidate-position instance i are denoted by J_i , R_i , and C_i , respectively. The trainable soft prompt matrix $\mathbf{P}$ contains m virtual-token embeddings of dimension d and is parameterized by θ , whereas Φ denotes all original parameters of the

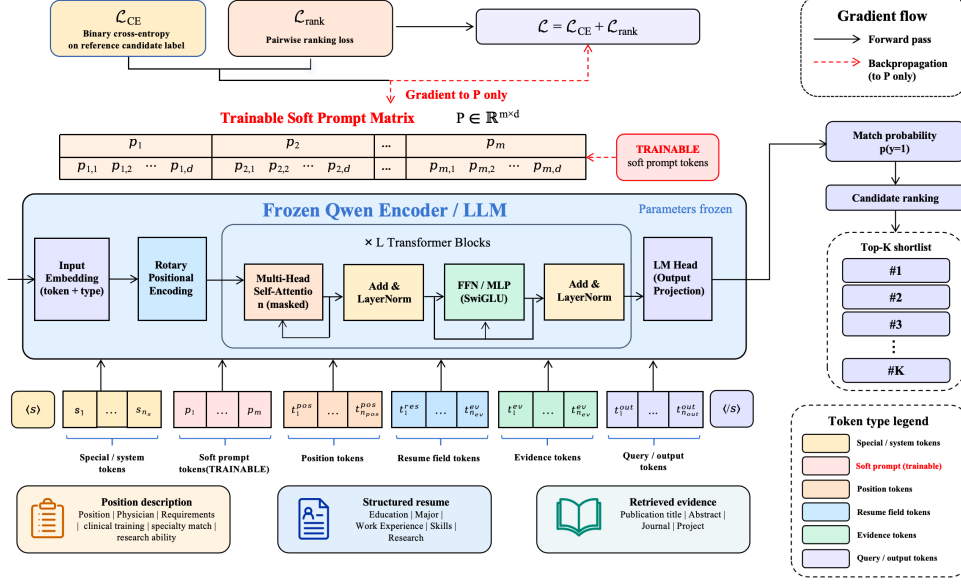


Fig. 1 Internet search-augmented soft prompt tuning framework. The framework comprises (I) evidence construction and serialization, (II) soft prompt adaptation with a frozen Qwen backbone, and (III) hybrid candidate ranking. Candidate-level binary cross-entropy and a within-position pairwise logistic loss are combined with fixed equal contributions, and gradients update only the soft prompt matrix $\mathbf{P}$. During inference, deterministic rules and predefined module weights transform the semantic matching probability into the final score used for ranking and Top- K selection.

Qwen3-235B backbone. Model operation is divided into training and inference phases.

During training, $\mathbf{P}$ is combined with the serialized evidence sequence and passed through the frozen embedding layer, rotary positional encoding, Transformer blocks, and language-model head. Binary cross-entropy provides candidate-level supervision, while the within-position pairwise logistic loss encourages a reference candidate to receive a larger ranking logit than a paired nonreference candidate applying for the same position. The two mean log losses are added using fixed equal contributions, without an additional loss-weight coefficient. Gradients are propagated through the frozen computation graph only to θ , and Φ remains unchanged. During inference, both θ and Φ are frozen. The model probability is used as a bounded semantic contribution within the applicable job-specific module; deterministic rules and predefined module weights then produce the final score F_{iJ} , from which the position-specific ranking and Top- K review pool are obtained. The fixed checkpoint, learned prompt, input

schema, retrieval configuration, scoring rules, and ranking procedure together define the reproducible analytical unit; deterministic inference uses `temperature = 0` and limits textual inputs to 20,000 characters before tokenization.

2.2 Structured Resume Parsing and Publication Evidence

For N candidate resumes provided as PDF, DOC, or TXT files, an unstructured file loader implemented in Python first converted each document into machine-readable text. A separate parsing prompt containing few-shot examples then extracted educational background, specialty, work experience, professional skills, publications, and research experience into a constrained JSON schema. The workflow is shown in Figure 2. This hard-prompt parsing stage standardized the evidence supplied to the matching model and was distinct from the trainable soft prompt used for candidate–position matching.

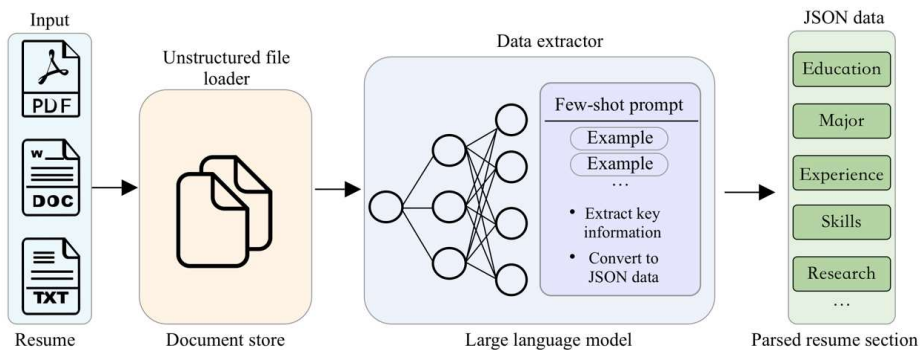


Fig. 2 LLM-based resume-parsing workflow.

The parser retained missing fields as blank or missing rather than inferring them from unrelated content. Personally identifying attributes were used only for record linkage and were excluded from scoring. The few-shot template and JSON-structured output are illustrated in Figure 3.

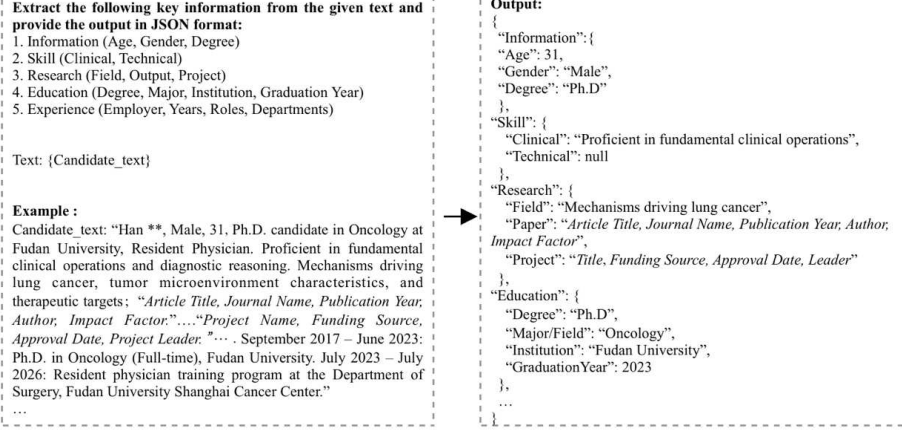


Fig. 3 Few-shot prompt template and JSON-structured output.

For research-related positions, publication titles explicitly reported in the resumes were used as restricted retrieval queries. The primary source was the Semantic Scholar API; publicly accessible publisher and institutional pages were used only when further confirmation was required. Retrieved titles, authors, years, journals or conferences, and abstracts were linked to the corresponding resume record. The system did not conduct unrestricted searches based on names, demographic attributes, or unrelated personal information. Retrieved abstracts were incorporated as external evidence and, when required by a research-position scoring scheme, were evaluated using a constrained binary relevance criterion $a_i \in \{0, 1\}$.

2.3 Candidate–Position Instance Construction

Hospital resume screening was formulated as a candidate–position matching task. For candidate–position instance i , J_i denotes the target-position description, R_i the structured resume, C_i the retrieved academic evidence when available, and $y_i \in \{0, 1\}$ whether the candidate belongs to the archived reference shortlist for that position. The recruitment dataset is defined as

$$\mathcal{D} = \{(J_i, R_i, C_i, y_i)\}_{i=1}^N \quad (1)$$

Equation (1) defines $\mathcal{D}$ at the candidate–position level. Here, N is the number of candidate–position instances rather than the number of unique applicants, and index i identifies one applicant evaluated for one target position. This distinction permits the same resume to form separate instances when an applicant is considered for different positions. J_i contains the position-specific requirements, R_i the corresponding structured resume, C_i only publication evidence traceable to the applicant, and $y_i \in \{0, 1\}$ the archived reference-shortlist label for that particular position.

Position requirements, resume fields, retrieved evidence, and the expected output were concatenated using explicit field identifiers. The resulting textual representation was

$$\mathbf{c}_i = [\text{Position}; J_i; \text{Resume}; R_i; \text{Evidence}; C_i; \text{Output}] \quad (2)$$

In Equation (2), lower-case $\mathbf{c}_i$ denotes the complete serialized representation, whereas upper-case C_i denotes only the retrieved publication evidence. Square brackets and semicolons indicate ordered concatenation rather than an arithmetic operation. The literal field identifiers provide stable boundaries between position requirements, resume attributes, supporting evidence, and the constrained output instruction. A unified missing-value token was used for absent information so that the segment order remained identical across instances.

Candidate records from the same individual were assigned to only one data subset to prevent leakage. The optimization subset was used to learn θ , the evaluation subset to select the checkpoint, and the test candidate pools exclusively for final ranking. Because reference candidates were less common than nonreference candidates, all positive instances were retained and negative instances were sampled within positions during optimization; evaluation and test rankings preserved the complete candidate pools and their original distributions.

415 2.4 Soft Prompt Initialization and Frozen Qwen Backbone

416
417 A soft prompt is a sequence of continuous task-specific embeddings learned from
418
419 labeled instances. Unlike a human-readable hard prompt, its vectors need not corre-
420
421 spond to natural-language tokens. Let s denote an initial instruction describing priority
422
423 review according to position requirements and candidate education, experience, qual-
424
425 ifications, and research ability. If $f_\theta(\cdot)$ is the trainable prompt-embedding layer, the
426
427 soft prompt matrix is initialized from s as

$$428 \quad \mathbf{P} = f_\theta(s) = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_m] \in \mathbb{R}^{m \times d} \quad (3)$$

430
431 Each row $\mathbf{p}_j \in \mathbb{R}^d$ in Equation (3) represents one virtual prompt token; con-
432
433 sequently, $\mathbf{P}$ contains m trainable vectors and has shape $m \times d$. The initialization
434
435 function $f_\theta(\cdot)$ uses instruction s to initialize these vectors, after which θ parameter-
436
437 izes their values directly. Because $\mathbf{p}_j$ is continuous, it is not required to coincide with
438
439 an embedding in the Qwen vocabulary. The matrix $\mathbf{P}$, rather than any original Qwen
440
441 parameter, is optimized for the matching task.

442 For each candidate–position instance, the shared initialization instruction and the
443
444 instance-specific serialized evidence formed

$$445 \quad \mathbf{t}_i = [s, \mathbf{c}_i] \quad (4)$$

446
447 In Equation (4), s supplies task-level initialization information shared across
448
449 instances, whereas $\mathbf{c}_i$ supplies the evidence specific to candidate–position instance i .
450
451 The brackets indicate sequence composition in a fixed order; no numerical addition is
452
453 performed.

454 The complete instance was tokenized into token identifiers

$$455 \quad \mathbf{z}_i = [z_{i,1}, z_{i,2}, \dots, z_{i,L_i}] \quad (5)$$

In Equation (5), $\mathbf{z}_i$ is the token-identifier sequence, $z_{i,j}$ is the j th token identifier, and L_i is the sequence length after tokenization. The same Qwen tokenizer and segment order were used for every instance, ensuring that input differences arose from candidate and position content rather than from changes in serialization format.

Virtual prompt tokens were mapped by $f_\theta(\cdot)$, whereas position requirements, resume fields, retrieved evidence, and the output instruction were mapped by the frozen Qwen embedding layer. The resulting ordered embedding sequence was

$$\mathbf{E}_i = [f_\theta(s); E_\Phi(J_i); E_\Phi(R_i); E_\Phi(C_i); E_\Phi(T^{\text{out}})] \quad (6)$$

In Equation (6), each term is a sequence of d -dimensional vectors, and the semicolons denote concatenation along the token dimension. The first m positions are supplied by $f_\theta(s) = \mathbf{P}$, whereas $E_\Phi(\cdot)$ maps the position, resume, evidence, and output tokens through the frozen Qwen embedding table. T^{out} denotes the constrained output instruction. This construction allows the trainable prompt to condition all evidence tokens without changing their discrete token identities.

The concatenated sequence was then passed through K frozen Transformer blocks. For token j of instance i , the layerwise hidden representation was

$$\mathbf{h}_{i,j} = [\mathbf{h}_{i,j}^{(1)}; \mathbf{h}_{i,j}^{(2)}; \dots; \mathbf{h}_{i,j}^{(K)}] \quad (7)$$

In Equation (7), $\mathbf{h}_{i,j}^{(k)} \in \mathbb{R}^d$ is the hidden representation of token j after Transformer block k , and K is the total number of frozen blocks. The bracketed notation collects the layerwise states to describe how the token representation evolves through the backbone; the calculation at each block uses the output of the preceding block. Self-attention therefore allows professional requirements to interact with specialty and credentials, clinical requirements with work and training experience, and research requirements with publications, abstracts, and projects.

All original Qwen backbone parameters were frozen and excluded from the optimizer. Gradients were propagated through the backbone only to update the trainable soft prompt embeddings.

2.5 Matching Probability and Training Objective

The trainable matching task used two output labels. Label token '1' denoted membership in the archived reference shortlist, whereas label token '0' denoted non-membership. Given instance $\mathbf{t}_i$, let $\ell_i^{(0)}$ and $\ell_i^{(1)}$ be the logits of the two label tokens.

The candidate matching probability was

$$p_i = p_{\theta, \Phi}(y_i = 1 \mid \mathbf{t}_i) = \frac{\exp(\ell_i^{(1)})}{\exp(\ell_i^{(0)}) + \exp(\ell_i^{(1)})} \quad (8)$$

Equation (8) applies a two-class softmax to the label logits $\ell_i^{(0)}$ and $\ell_i^{(1)}$. The denominator normalizes the two scores so that $p_i \in (0, 1)$, while the difference $\ell_i^{(1)} - \ell_i^{(0)}$ determines the odds assigned to the reference label. A larger p_i therefore indicates stronger candidate–position correspondence. The soft prompt was optimized by maximizing the conditional probability of the observed reference label for each training instance:

$$\max_{\theta} \log p_{\theta, \Phi}(y_i \mid \mathbf{t}_i) \quad (9)$$

Equation (9) gives the instance-level maximum-likelihood criterion with respect to θ while Φ remains fixed. Because the logarithm is monotonic, maximizing log probability has the same optimum as maximizing probability, while providing additive and numerically stable instance contributions. For the complete training subset $\mathcal{D}^{\text{train}}$, containing $|\mathcal{D}^{\text{train}}|$ instances, the objective was

$$\max_{\theta} \sum_{i=1}^{|\mathcal{D}^{\text{train}}|} \log p_{\theta, \Phi}(y_i \mid \mathbf{t}_i) \quad (10)$$

Equation (10) sums the log-likelihood contributions across all training instances and favors prompt parameters that assign high probability to each observed label. Equivalently, the negative log-likelihood minimized by the optimizer was

$$\mathcal{L} = - \sum_{i=1}^{|\mathcal{D}^{\text{train}}|} \log p_{\theta, \Phi}(y_i | \mathbf{t}_i) \quad (11)$$

The leading minus sign in Equation (11) converts likelihood maximization into loss minimization. For the binary label task, expanding the conditional likelihood yields the binary cross-entropy objective:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log p_i + (1 - y_i) \log(1 - p_i)] \quad (12)$$

In Equation (12), N denotes the number of instances included in the loss calculation, $y_i \in \{0, 1\}$, and p_i is the probability defined by Equation (8). The first logarithmic term penalizes a low probability for a reference candidate, whereas the second penalizes a high probability for a nonreference candidate. The averaging factor $1/N$ keeps the loss scale comparable across optimization batches of different sizes.

Because candidate ordering was performed separately for each position, positive-negative pairs were constructed only among applicants for the same position. Candidate-level data splitting was completed before pair construction to prevent information leakage. During each training epoch, every reference candidate was paired with one nonreference candidate sampled uniformly from the corresponding position, and negative candidates were resampled in subsequent epochs. This one-to-one strategy exposed the model to different within-position comparisons without constructing the complete Cartesian product of positive and negative candidates. Cross-position pairs and hard-negative mining were not used.

Let $r_i = \ell_i^{(1)} - \ell_i^{(0)}$ denote the ranking logit of instance i . Because $p_i = \sigma(r_i)$, ordering candidates by r_i is equivalent to ordering them by the matching probability p_i , while the unbounded logit difference provides a more stable pairwise optimization

599 signal. For each optimization batch, a position-specific set of sampled reference–
600 nonreference pairs and the set of positions containing at least one valid pair were
601 defined. The position-balanced pairwise ranking loss was calculated as
602
603

$$\mathcal{L}_{\text{rank}} = -\frac{1}{|\mathcal{J}_B|} \sum_{J \in \mathcal{J}_B} \frac{1}{|\mathcal{P}_J|} \sum_{(i^+, i^-) \in \mathcal{P}_J} \log \sigma(r_{i^+} - r_{i^-}) \quad (13)$$

604
605 In Equation (13), i^+ and i^- denote a reference candidate and a nonreference can-
606 didate from the same position, respectively; the inner sum is taken over the sampled
607 pair set for position J , the outer sum is taken over positions represented by at least
608 one valid pair in the batch, and $\sigma(\cdot)$ is the sigmoid function. The loss increases when
609 r_{i^+} does not exceed r_{i^-} and decreases as the positive–negative logit difference becomes
610 larger. Computing a mean within each position and then averaging across positions
611 prevents positions with larger applicant pools from dominating prompt optimization.
612 A position with no valid reference–nonreference pair in a batch contributes no pairwise
613 term for that batch.
614

615 The binary cross-entropy term preserved candidate-level discrimination, whereas
616 the pairwise term regularized relative ordering within each position. Because both
617 terms were calculated as mean log losses, the total objective used fixed equal
618 contributions without an additional tunable coefficient:
619

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{rank}} \quad (14)$$

620 Equation (14) adds the two normalized objectives directly, thereby avoiding unre-
621 ported loss weights. During backpropagation, gradients pass through the frozen Qwen
622 computation graph only to θ , the parameters of $\mathbf{P}$; Φ is not included in the optimizer.
623 The pairwise term therefore regularizes the soft-prompt-derived semantic matching
624 signal rather than replacing the deterministic score fusion in Section 2.6.
625
626
627
628
629
630
631
632
633
634

The checkpoint with the best evaluation performance was retained as θ^* . Only the learned prompt embeddings were saved, and the original Qwen checkpoint remained unchanged.

2.6 Job-Specific Hybrid Scoring and Candidate Ranking

The matching probability was used as a bounded semantic contribution within the applicable job-specific module rather than as an unverified final hiring decision. Each position used a scoring scheme fixed before evaluation that defined allowable evidence, maximum module scores, deterministic penalties, semantic items, and module weights. Deterministic rules handled explicitly verifiable criteria, including age range, educational qualification, professional credentials, journal quartile, impact factor, project level, and specified clinical experience. Soft-prompt-derived probabilities were used only for context-dependent judgments such as specialty–position, research-direction, and abstract relevance.

The predefined module weights for the nine evaluated positions are presented in Table 1.

Table 1: Predefined module weights used for the nine evaluated positions

Cat.	Pos.	BI	E/F	R/E
Administrative	Medical Affairs Department Officer	0.40	0.20	0.40
Administrative	Finance Department Officer	0.40	0.20	0.40
Administrative	Information Center Officer	0.30	0.30	0.40
Health Technical	Nurses for Clinical Nursing Units	0.30	0.20	0.50
Health Technical	Radiology Technician	0.20	0.30	0.50
Health Technical	Thoracic Surgery Physician	0.20	0.20	0.60
Research	Central Laboratory Technician	0.15	0.20	0.65
Research	Thoracic Oncology Institute Laboratory Technician	0.15	0.20	0.65
Research	Cardiovascular Basic Research Center Laboratory Technician	0.15	0.20	0.65

Abbreviations: Cat., category; Pos., position; BI, Basic Information; E/F, Education/Fit; R/E, Research/Experience.

Let s_{ik} denote the raw score of candidate i in module k , and let M_k^J denote the maximum score of that module for position J . The normalized module score was

$$\tilde{s}_{ik} = 100 \frac{\min\{\max(s_{ik}, 0), M_k^J\}}{M_k^J} \quad (15)$$

Equation (15) applies nested clipping before normalization: $\max(s_{ik}, 0)$ removes negative module scores, and $\min\{\cdot, M_k^J\}$ caps values at the maximum defined for module k in position J . Division by M_k^J and multiplication by 100 then map the normalized score $\tilde{s}_{ik}$ to $[0, 100]$. This operation makes modules with different raw point ranges directly comparable before weighting.

The final candidate score combined the normalized modules and the recorded hard-rule penalty:

$$F_{iJ} = \sum_k w_k^J \tilde{s}_{ik} + H_{iJ}, \quad \sum_k w_k^J = 1 \quad (16)$$

In Equation (16), w_k^J is the predefined nonnegative weight of module k for position J , and $\sum_k w_k^J = 1$ makes the weighted component a convex combination of normalized module scores. The term H_{iJ} is the separate additive hard-rule adjustment for candidate i under position J : it is zero when no hard rule is triggered and negative when a mandatory criterion is violated. Keeping this adjustment outside the weighted sum prevents an exclusionary rule from being diluted by high semantic scores in other modules.

Candidates within each position were ordered in descending final-score order:

$$\pi_J = \underset{i}{\operatorname{argsort}}(F_{iJ}; \text{descending}) \quad (17)$$

In Equation (17), π_J denotes the position-specific permutation of candidate indices for position J , and $\pi_J(r)$ denotes the candidate placed at rank r . Descending sorting

places larger final scores earlier in the review order; predefined secondary keys are invoked only when two candidates have exactly the same final score.

The first K candidates formed the priority manual-review pool:

$$\mathcal{C}_J^K = \{\pi_J(1), \pi_J(2), \dots, \pi_J(K)\} \quad (18)$$

Equation (18) selects the first K indices from the position-specific permutation to form the Top- K set. Here, K is an operational manual-review capacity rather than a learned model parameter, and the resulting set is a priority review pool rather than an automated final hiring decision. K was set to R , approximately $1.25R$, $1.5R$, $2R$, or $3R$, where R was the archived number of reference candidates for the position, thereby reproducing the manual-review capacities evaluated in Section 2.9.

2.7 Inference, Interpretability, and Audit Trail

During inference, the learned soft prompt $f_{\theta^*}(s)$ was prepended to each candidate–position instance. Both θ^* and Φ were fixed, and neither a loss nor gradients were calculated. The operational sequence was: structured resume and position requirements $\rightarrow$ restricted publication evidence $\rightarrow$ candidate–position instance $\rightarrow$ learned soft prompt plus frozen Qwen $\rightarrow$ bounded semantic contributions $\rightarrow$ deterministic module fusion $\rightarrow$ ranked manual-review pool.

For each candidate, the system stored an explanation object containing (1) extracted evidence; (2) the rule or semantic criterion applied to that evidence; (3) soft-prompt-derived probabilities and raw module scores; (4) normalized module scores and triggered hard rules; (5) weighted contributions; and (6) the final score and rank. This design supported local inspection of individual candidates and global analysis of which modules were most strongly associated with archived expert shortlists. The LLM therefore did not directly issue an unverified recruitment decision; it supplied

bounded semantic evidence within a deterministic, job-specific, and auditable ranking procedure.

2.8 Data Sources and Reference Standards

Institutional recruitment records contained 5,419 applications for 38 positions during the 2025 recruitment cycle. Application materials included structured registration fields, PDF resumes, and, for some candidates applying to research positions, submitted publication lists and retrieved abstracts. Ranking evaluation included only the nine positions with complete archived reference shortlists and records meeting the model-processing requirements, comprising 4,339 applications and 150 reference candidates, as summarized in Table 2.

Table 2: Positions included in the ranking evaluation

Cat.	Pos.	N	R
Administrative	Medical Affairs Department Officer	127	25
Administrative	Finance Department Officer	34	7
Administrative	Information Center Officer	30	6
Health Technical	Nurses for Clinical Nursing Units	3,559	58
Health Technical	Radiology Technician	409	18
Health Technical	Thoracic Surgery Physician	83	11
Research	Central Laboratory Technician	65	12
Research	Thoracic Oncology Institute Laboratory Technician	11	6
Research	Cardiovascular Basic Research Center Laboratory Technician	21	7
Total	—	4,339	150

Abbreviations: Cat., category; Pos., position; N , number of applications; R , number of reference candidates.

For the information-extraction evaluation, researchers manually cross-checked PDF resumes, structured registration fields, and archived materials to establish field-level reference standards. Archived shortlists were used as the operational reference for

ranking analyses. These labels reflect historical expert decision-making processes and should not be interpreted as immutable or completely unbiased definitions of candidate quality.

2.9 Experimental Comparisons

Four groups of experiments were conducted. First, field-extraction results were compared with manually reviewed reference standards. Second, under the same foundation model and deterministic inference settings, the pipeline-based model was compared with the Raw Prompt baseline. The Raw Prompt baseline did not perform prior field extraction or modular decomposition; instead, the full resume text and scoring rules were entered in a single model call. Third, performance was evaluated using manual-review pool sizes of R , approximately $1.25R$, $1.5R$, $2R$, and $3R$, where R denotes the number of reference candidates for the corresponding position. Fourth, the abstract-relevance contribution was removed from the three research positions while all other settings were held constant.

2.10 Evaluation Metrics and Statistical Analysis

Information-extraction performance was evaluated using accuracy, sensitivity, specificity, precision, and F1 score. Ranking performance was evaluated using R-Precision, Recall@ K , F1@ K , average precision, and normalized discounted cumulative gain (NDCG). When $K = R$, the predicted and reference sets have the same size; therefore, Precision@ R , Recall@ R , and F1@ R are numerically identical.

Paired position-level differences between the pipeline-based model and the Raw Prompt method were assessed using exact one-sided Wilcoxon signed-rank tests. A total of 20,000 bootstrap resamples across the nine positions were used to estimate mean differences and their 95% CIs. The resampling and hypothesis-testing unit was the position, not the individual application, because applications were clustered within positions and the comparison concerned macro-average performance across

positions. For the interpretability analysis, Spearman correlation coefficients between the final or module scores and binary archived-shortlist labels were calculated within each position. Macro-averages of the position-specific coefficients were reported, and position-level bootstrapping was used to estimate 95% CIs. Exact one-sided Wilcoxon tests were used to determine whether the position-specific correlations were greater than 0, and paired Wilcoxon tests were used to compare correlations across modules. Depending on the objective of each test, statistical analyses used either two-sided tests or one-sided tests with a prespecified direction; the significance level was $\alpha = 0.05$.

3 Results

3.1 Key-Information Extraction Performance

The key-information extraction results are presented in Table 3.

Table 3: Information-extraction performance against manually reviewed reference standards

Field	Acc.	Sens.	Spec.	Prec.	F1
Educational Background	98.40%	98.41%	98.76%	98.40%	98.41%
Research Experience	99.35%	97.25%	99.29%	96.99%	97.12%
Skills	98.14%	98.18%	97.02%	95.55%	96.85%
Work Experience	99.92%	99.84%	99.87%	99.86%	99.85%
Publications	99.66%	99.66%	99.66%	99.66%	99.66%
Overall	98.85%	98.68%	97.01%	98.22%	98.45%

Abbreviations: Acc., accuracy; Sens., sensitivity; Spec., specificity; Prec., precision; F1, F1 score.

3.2 Comparison With the Raw Prompt Baseline

Table 4 compares the performance of the pipeline-based model and the Raw Prompt baseline at $K = R$. Across the nine positions, the pipeline-based model achieved

higher R-Precision for seven positions and tied the baseline for two. Macro-average R-Precision increased from 61.18% to 82.64%. The paired mean improvement was 21.46 percentage points (position-level bootstrap 95% CI, 8.73–37.51; exact Wilcoxon test, $P = 0.0078$). Macro-average NDCG@ R increased from 0.6632 to 0.8702, corresponding to a paired mean improvement of 0.2070 (position-level bootstrap 95% CI, 0.0771–0.3751; $P = 0.0039$).

Table 4: Person–job matching performance of the pipeline-based model and Raw Prompt baseline at $K = R$

Cat.	Pos.	N	R	RP R-Prec.	PBM R-Prec.	RP NDCG	PBM NDCG
Administrative	Medical Affairs Department Officer	127	25	68.00%	80.00%	0.7461	0.8409
Administrative	Finance Department Officer	34	7	71.43%	85.71%	0.6067	0.9084
Administrative	Information Center Officer	30	6	66.67%	83.33%	0.7409	0.8922
Health Technical	Nurses for Clinical Nursing Units	3,559	58	36.21%	82.76%	0.5060	0.8700
Health Technical	Radiology Technician	409	18	77.78%	77.78%	0.8410	0.8359
Health Technical	Thoracic Surgery Physician	83	11	9.09%	81.82%	0.1037	0.8822
Research	Central Laboratory Technician	65	12	66.67%	83.33%	0.7655	0.8173
Research	Thoracic Oncology Institute Laboratory Technician	11	6	83.33%	83.33%	0.8487	0.8829
Research	Cardiovascular Basic Research Center Laboratory Technician	21	7	71.43%	85.71%	0.8105	0.9021
Macro-average	—	—	—	61.18%	82.64%	0.6632	0.8702

Abbreviations: Cat., category; Pos., position; N , number of candidates; R , number of reference candidates; RP, Raw Prompt; PBM, pipeline-based model; R-Prec., R-Precision; NDCG, normalized discounted cumulative gain.

The largest improvement was observed for the Thoracic Surgery Physician position, for which R-Precision increased from 9.09% to 81.82% and $\text{NDCG}@R$ increased from 0.1037 to 0.8822. Despite the substantially larger number of candidates for the Nurses for Clinical Nursing Units position, its R-Precision increased from 36.21% to 82.76%. For the Radiology Technician position, R-Precision was identical under the two methods, and $\text{NDCG}@R$ differed by only 0.0051.

3.3 Performance Across Manual-Review Pool Sizes

The pipeline-based model maintained higher macro-average $\text{Recall}@K$ across all pool sizes (Table 5). Recall increased from 82.64% at R to 88.47% at approximately $1.25R$, 92.75% at $1.5R$, and 97.83% at $2R$. Further expansion of the pool from $2R$ to $3R$ increased recall by only 0.83 percentage points, whereas F1 decreased as more nonreference candidates entered the review pool.

Table 5: Macro-average performance across manual-review pool sizes

Pool	RP Recall	PBM Recall	RP F1	PBM F1	RP NDCG	PBM NDCG
R	61.18%	82.64%	61.18%	82.64%	0.6632	0.8702
Approx. $1.25R$	66.58%	88.47%	57.35%	76.44%	0.6959	0.9050
Approx. $1.5R$	73.15%	92.75%	57.28%	72.57%	0.7328	0.9287
Approx. $2R$	83.35%	97.83%	55.03%	64.19%	0.7852	0.9548
Approx. $3R$	86.39%	98.66%	44.11%	50.05%	0.7993	0.9592

Abbreviations: Pool, manual-review pool size; RP, Raw Prompt; PBM, pipeline-based model; F1, F1 score; NDCG, normalized discounted cumulative gain.

3.4 Abstract-Module Ablation Study

After removal of the publication-abstract relevance module, R-Precision decreased for the Central Laboratory Technician and Cardiovascular Basic Research Center Laboratory Technician positions, whereas the result for the Thoracic Oncology Institute Laboratory Technician position remained unchanged (Table 6). Across the three research positions, macro-average R-Precision decreased from 84.13% to 73.81%, and macro-average NDCG@ R decreased from 0.8674 to 0.8022.

Table 6: Results of the abstract-module ablation study for research positions; R-Precision Δ values are percentage-point differences

Pos.	R	+Abs. R-Prec.	-Abs. R-Prec.	Δ R-Prec. (pp)	+Abs. NDCG	-Abs. NDCG	Δ NDCG
Central Laboratory Technician	12	83.33%	66.67%	-16.67	0.8173	0.7132	-0.1041
Thoracic Oncology Institute Laboratory Technician	6	83.33%	83.33%	0.00	0.8829	0.8829	0.0000
Cardiovascular Basic Research Center Laboratory Technician	7	85.71%	71.43%	-14.29	0.9021	0.8105	-0.0916
Macro-average	—	84.13%	73.81%	-10.32	0.8674	0.8022	-0.0652

Abbreviations: Pos., position; R , number of reference candidates; +Abs., with publication abstracts; -Abs., without publication abstracts; R-Prec., R-Precision; pp, percentage points; NDCG, normalized discounted cumulative gain; Δ , difference calculated as without abstracts minus with abstracts.

For the Central Laboratory Technician position, eight candidates received valid abstract contributions, accounting for 12.31% of the candidates for that position; the maximum valid contribution was 0.30 points. After ablation, the number of reference candidates in the top- R pool decreased by two. Thus, the effect of this module was concentrated primarily among candidates with similar base scores near the shortlist boundary.

1059 **3.5 Interpretability and Agreement With Archived Expert**

1060

1061 **Shortlists**

1062

1063 As shown in Table 3, the model’s field-extraction results showed high agreement with
1064 the manually reviewed reference standards. To evaluate which model components were
1065 most closely associated with historical expert decisions, associations between module
1066 scores and binary archived-shortlist labels were further calculated within each position;
1067 the results are summarized in Table 7 and visualized in Figure 4.

1071

1072 **Table 7:** Within-position correlations between model scores and archived expert
1073 shortlists

1074

Score/Mod.	Mean ρ	95% CI	P
Final Score	0.583	0.473–0.670	0.002
Research/Experience Module	0.587	0.521–0.642	0.002
Basic Information Module	0.276	0.116–0.423	0.014
Education/Fit Module	0.134	0.029–0.253	0.049

1082

1083

1084 *Abbreviations:* Mod., module; CI, confidence interval; P , exact probability value; ρ , Spearman rank-
1085 correlation coefficient.

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

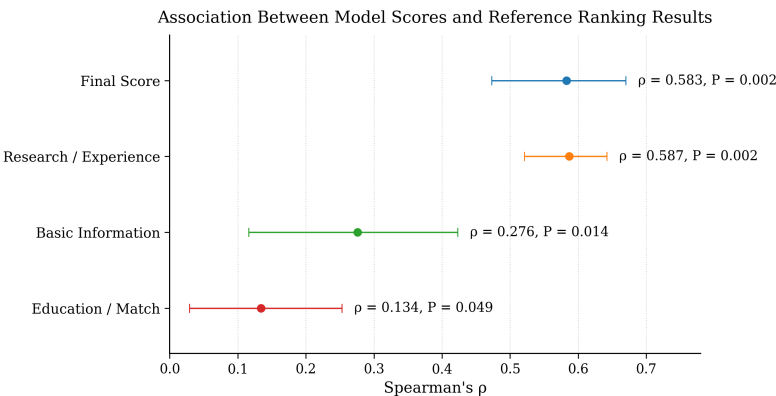
1098

1099

1100

1101 **Fig. 4 Agreement between model scores and archived expert shortlists.** Macro-average
1102 Spearman correlation coefficients are shown across positions. Error bars indicate position-level boot-
1103 strap 95% confidence intervals.

1104



The Research/Experience module showed the strongest association with the archived shortlists (mean $\rho = 0.587$). Its within-position correlations were significantly higher than those of the Basic Information module (paired exact Wilcoxon test, $P = 0.0098$) and the Education/Fit module ($P = 0.0039$). The final score showed a similar association strength (mean $\rho = 0.583$), indicating that weighted fusion preserved the principal signal from the Research/Experience module while incorporating qualification criteria and person–job fit information. These correlations reflect the degree of agreement between the model and the historical expert process.

4 Discussion

4.1 Principal Findings

The pipeline-based framework improved candidate identification and ranking performance across different types of hospital positions. Compared with the Raw Prompt baseline using the same foundation model, modular processing increased macro-average R-Precision by 21.46 percentage points and NDCG@ R by 0.2070. Improvements were particularly pronounced for the Thoracic Surgery Physician and nursing positions, indicating that explicit decomposition is especially valuable when a position combines clearly defined mandatory qualifications with semantic evaluation of professional experience.

The results also support the use of an expanded manual-review pool in practice. At a pool size of approximately $2R$, the system had already retrieved 97.83% of the archived reference candidates, with limited additional gains from further expansion. This finding provides an operational and explainable threshold for manual review but does not imply that automated ranking can replace expert judgment.

1151 4.2 Methodological Contribution and Domain Transfer

1152

1153 The methodological contribution of this study lies in integrating three types of
1154 knowledge: general linguistic knowledge encoded in the frozen LLM, task-specific
1155 recruitment knowledge learned in a compact soft prompt and constrained by expert
1156 scoring schemes, and publication evidence obtained through restricted external
1157 retrieval. Unlike full-parameter fine-tuning, only the continuous prompt embeddings
1158 were optimized, allowing the Qwen backbone to remain unchanged. Unlike a mono-
1159 lithic Raw Prompt approach, the framework separated evidence extraction, semantic
1160 judgment, deterministic rules, and final score calculation. The performance improve-
1161 ments observed in the comparative experiments provide empirical support for this
1162 combined parameter-efficient and system-level design.

1163 The internet-retrieval branch should also be distinguished from unrestricted web
1164 browsing. Queries in this study were generated only from publication titles self-
1165 reported by candidates, and retrieved content was used as supporting evidence rather
1166 than as an independent source of personal information. This restriction improved
1167 traceability and reduced the risk of introducing irrelevant or sensitive information.

1168

1169 4.3 Interpretability and Human Oversight

1170

1171 The interpretability analysis showed that the Research/Experience module was the
1172 component most strongly associated with archived expert shortlists. The abstract
1173 ablation study further demonstrated that even small semantic contributions could
1174 alter the relative ordering of candidates near the shortlist boundary. These findings
1175 indicate that explanations should include both global module importance and local
1176 evidence contributions. In practical deployment, recruiters should be able to inspect
1177 extracted text, triggered rules, module scores, and their effects on the final ranking.

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

However, agreement with historical shortlists is not equivalent to normative correctness. Historical decisions may contain unobserved preferences or biases. Accordingly, the system should be used as a decision-support and prioritization tool, protected demographic attributes should be excluded from scoring, and final decisions should remain subject to accountable human review [16–19].

4.4 Limitations

This study has several limitations. First, it was conducted at a single hospital using retrospective recruitment records; external and prospective validation are therefore required. Second, only nine positions had complete reference shortlists available for ranking comparisons, limiting the sample size for position-level statistical tests. Third, the reference shortlists represented an operational gold standard rather than independently adjudicated continuous ratings. Future studies should collect blinded, multi-rater assessments and report inter-rater agreement, intraclass correlation coefficients, and calibration results. Fourth, the soft prompt was initialized from a manually designed recruitment instruction and learned from historical labels; prompt length, initialization, optimization settings, and checkpoint versions therefore require controlled reporting. Fifth, publication retrieval may be affected by title ambiguity and incomplete metadata. Finally, only the prompt embeddings were tuned; future work should compare this approach with LoRA and full-parameter fine-tuning on strictly independent external test sets.

5 Conclusions

This study demonstrates the value of an internet search-augmented soft prompt tuning framework as a scalable and auditable decision-support system for health-care recruitment. By combining structured evidence extraction, restricted publication

1243 retrieval, parameter-efficient semantic adaptation, job-specific rules, and transpar-
1244 ent score fusion, the system can standardize heterogeneous resume review, identify
1245 qualified candidates more reliably than a monolithic Raw Prompt approach, and con-
1246 centrate human attention on a manageable priority pool while retaining an inspectable
1247 evidence trail for every ranking. The framework is intended to reduce repetitive
1248 screening workload and improve consistency without replacing accountable human
1249 judgment. Multicenter prospective studies should evaluate generalizability, fairness,
1250 and operational impact before broader deployment.

1251

1252

1253 **List of abbreviations**

1254

1255 AI: artificial intelligence; API: application programming interface; CI: confidence inter-
1256 val; DOC: Microsoft Word document format; JSON: JavaScript Object Notation;
1257 LLM: large language model; NDCG: normalized discounted cumulative gain; NLP:
1258 natural language processing; PDF: Portable Document Format; TXT: plain-text file.

1259

1260 **Declarations**

1261

1262

1263 **Ethics approval**

1264

1265 This retrospective study was approved by the Ethics Committee of Shanghai Chest
1266 Hospital (approval no. KS26065). The study was conducted in accordance with the
1267 Declaration of Helsinki and relevant guidelines and regulations. The requirement for
1268 informed consent was waived by the Ethics Committee of Shanghai Chest Hospital
1269 due to the retrospective nature of the study.

1270

1271

1272 **Consent for publication**

1273

1274 Not applicable.

1275

1276

1277

Data availability	1289
	1290
The datasets generated during and/or analysed during the current study are available	1291
from the corresponding author on reasonable request.	1292
	1293
	1294
	1295
Competing interests	1296
	1297
The authors declare no competing interests.	1298
	1299
	1300
Funding	1301
	1302
This study was supported by the China Hospital Development Institute, Shanghai	1303
Jiao Tong University (Grant No. CHDI-2026-Z-069).	1304
	1305
	1306
	1307
References	1308
	1309
[1] Baum M, Kabst R. The Effectiveness of Recruitment Advertisements and Recruit-	1310
ment Websites: Indirect and Interactive Effects on Applicant Attraction. Human	1311
Resource Management. 2014;53(3):353–378.	1312
	1313
	1314
	1315
[2] Zhu C, Zhu H, Xiong H, Ma C, Xie F, Ding P, et al. Person-Job Fit: Adapting	1316
the Right Talent for the Right Job with Joint Representation Learning. ACM	1317
Transactions on Management Information Systems (TMIS). 2018;9(3):1–17.	1318
	1319
	1320
	1321
[3] Patterson F, Roberts C, Hanson MD, Hampe W, Eva K, Ponnampereuma G, et al.	1322
2018 Ottawa Consensus Statement: Selection and Recruitment to the Healthcare	1323
Professions. Medical teacher. 2018;40(11):1091–1101.	1324
	1325
	1326
	1327
[4] Sinha AK, Amir Khusru Akhtar M, Kumar A. Resume Screening Using Natural	1328
Language Processing and Machine Learning: A Systematic Review. Machine	1329
Learning and Information Processing: Proceedings of ICMLIP 2020. 2021;p. 207–	1330
214.	1331
	1332
	1333
	1334

1335 [5] Pal R, Shaikh S, Satpute S, Bhagwat S. Resume Classification Using Vari-
1336 ous Machine Learning Algorithms. In: ITM Web of Conferences. vol. 44. EDP
1337 Sciences; 2022. p. 03011.

1340 [6] Deshmukh A, Raut A. Applying Bert-Based Nlp for Automated Resume
1341 Screening and Candidate Ranking. *Annals of Data Science*. 2025;12(2):591–603.

1344 [7] Vaishampayan S, Leary H, Alebachew YB, Hickman L, Stevenor B, Beck W,
1345 et al. Human and LLM-based Resume Matching: An Observational Study. In:
1346 Findings of the Association for Computational Linguistics: NAACL 2025; 2025.
1347 p. 4823–4838.

1350 [8] Gan C, Zhang Q, Mori T. Application of LLM Agents in Recruitment: A Novel
1351 Framework for Automated Resume Screening. *Journal of Information Processing*.
1352 2024;32:881–893.

1355 [9] Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-Train, Prompt,
1356 and Predict: A Systematic Survey of Prompting Methods in Natural Language
1357 Processing. *ACM computing surveys*. 2023;55(9):1–35.

1360 [10] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the
1361 Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:210807258*.
1362 2021;[arXiv:2108.07258](https://arxiv.org/abs/2108.07258).

1365 [11] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large Language
1366 Models Encode Clinical Knowledge. *Nature*. 2023;620(7972):172–180.

1369 [12] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-
1370 Augmented Generation for Knowledge-Intensive Nlp Tasks. *Advances in neural*
1371 *information processing systems*. 2020;33:9459–9474.

[13] Wang D, Zhang S. Large Language Models in Medical and Healthcare Fields: Applications, Advances, and Challenges. Artificial intelligence review. 2024;57(11):299.	1381 1382 1383 1384 1385 1386
[14] Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, et al. React: Synergizing Reasoning and Acting in Language Models. arXiv preprint arXiv:221003629. 2022;arXiv:2210.03629.	1387 1388 1389 1390 1391 1392
[15] Liu X, Lai H, Yu H, Xu Y, Zeng A, Du Z, et al. WebGLM: Towards an Efficient Web-Enhanced Question Answering System with Human Preferences. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2023. p. 4549–4560.	1393 1394 1395 1396 1397 1398 1399
[16] Lavanchy M, Reichert P, Narayanan J, Savani K. Applicants’ Fairness Perceptions of Algorithm-Driven Hiring Procedures. Journal of Business Ethics. 2023;188(1):125–150.	1400 1401 1402 1403 1404 1405
[17] Alon-Barkat S, Busuioc M. Human–AI Interactions in Public Sector Decision Making: “Automation Bias” and “Selective Adherence” to Algorithmic Advice. Journal of Public Administration Research and Theory. 2023;33(1):153–169.	1406 1407 1408 1409 1410 1411
[18] Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR. Discrimination in the Age of Algorithms. Journal of legal analysis. 2018;10:113–174.	1412 1413 1414 1415
[19] Patterson F, Zibarras L, Ashworth V. Situational Judgement Tests in Medical Education and Training: Research, Theory and Practice: AMEE Guide No. 100. Medical teacher. 2016;38(1):3–17.	1416 1417 1418 1419 1420 1421 1422 1423 1424 1425 1426