Supplementary Information for "What Do Resume–Job Matching Benchmarks Actually Measure? Keywords, Embeddings and LLMs on Public Data"

Nusrat Jahan Rita

S1 Prompts

LLM judge (Qwen3-4B-Instruct-2507 and Qwen3-1.7B, Section 4.4 of the main paper). The job is cut to its first 450 words and the resume to its first 900 words; "..." marks a cut. Thinking mode is disabled.

```
[system]
You are an experienced technical recruiter. You judge how well a
candidate fits one specific job, based only on the job description
and the resume.

[user]
Job description:
"""
{job}
"""

Candidate resume:
"""
{resume}
"""

How well does this candidate fit THIS job?
A) Good Fit - meets the main requirements (role, key skills,
   experience level)
B) Potential Fit - related background, but missing some key
   requirements
C) No Fit - different field or clearly unqualified

Answer with a single letter: A, B or C.
```

Resume rewrite (Phi-4-mini-instruct, greedy decoding, at most 1,600 new tokens, Section 4.6 of the main paper). No system message is used.

```
Rewrite the resume below so that it reads as polished and
professional, the way a careful career coach or an AI writing
assistant would improve it.

Rules:
- Keep every fact exactly: the same jobs, employers, job titles,
  dates, degrees, skills, tools and numbers.
- Do NOT add any skill, tool, certification, achievement, number
  or responsibility that is not in the original.
```

```
- Do NOT remove information.
- You may improve wording, grammar, structure, headings and bullet
  points, and use stronger action verbs.
- Output only the rewritten resume, with no comments before or
  after it.

Resume:
"""
{resume}
"""
```

S2 Labelling guide

These rules were written after all 150 pairs had been labelled, based on a review of the notes the author made during labelling. While labelling, the author focused only on how well each candidate matched the specific job, not on how good the resume was overall. The original dataset labels were hidden throughout the process.

Good Fit. The candidate has worked in the same role that the job asks for and has almost all of the skills and qualifications it requires. Their experience, and their education where the job asks for it, also line up with what the employer is looking for. Small gaps, like having two years less experience than required, are acceptable if they have done similar work. Missing details that the job does not stress, such as education information, are not a major concern when the candidate has the skills needed for the job.

Potential Fit. The candidate has a background in the right field and meets many of the job requirements, but there is still something important missing. For example, they may have less experience than required, such as 2–3 years when the job asks for 4 or more, or they may lack certain skills or tools like Django, Salesforce, SQL, or C#/.NET. They might also have limited experience in client handling, business analysis, or team leadership. In some cases, the candidate may have a different degree than the one requested, even though their skills and experience are suitable. Another example is someone working in the same profession but in a different area, such as a tax accountant applying for a construction-accountant role. It could also be someone who understands the field but has not worked in it directly, or someone whose experience is far above the level required for the job.

No Fit. The candidate's background is quite different from what the job requires. This usually happens when someone works in a different field, such as a QA tester applying for a software-engineer role, an accountant applying for a developer job or vice versa, an electrical engineer applying for a software position or vice versa, or a data analyst or data engineer applying for a business-analyst role. The same goes for people from customer service, marketing, or product management applying for technical jobs. A candidate may also be a No Fit if they lack the main skills needed for the role, even if a few smaller skills match, or if their experience level is clearly unsuitable and they are missing important skills. Another reason is when several key requirements are missing at the same time, such as having neither SQL knowledge nor a bachelor's degree when the job requires both.

Decision rules.

1. **Role first.** If the candidate's profession did not match the job, the pair was labelled No Fit, even if some skills overlapped. For example, QA and testing candidates applying for software-development jobs were consistently labelled No Fit.

2. **Stated requirements.** The author compared each resume with the requirements listed in the job ad, such as skills, years of experience, and degree. The decision was based on what

the employer asked for, not on a general idea of what makes a strong candidate.

3. **Skills before years.** Years of experience were important, but skills mattered more. If a candidate had the right skills but less experience than required, the pair was labelled Potential Fit rather than No Fit.

4. **Unclear resumes.** When a resume did not clearly explain the candidate's skills or experience, the pair was usually labelled No Fit, although some were labelled Potential Fit. This was the least consistent rule because unclear information made it harder to make a decision.

S3 Additional tables and figures

Table S1 and Table S2 give the agreement between the author's labels and the dataset labels (Section 6.3 of the main paper). Table S3 gives the markdown-injection steps for the methods of Section 6.4 of the main paper. Fig. S1 shows the Good-vs-Potential results of Table 7 in the main paper with 95% confidence intervals. Fig. S2 shows the steps of the polish effect from Table 9 in the main paper.

AI/ML jobs. Under a rule recorded in the analysis notebook before it was run (at least 30 usable jobs), we also report the 35 jobs that mention AI or machine-learning terms. Keyword methods find these jobs harder: TF-IDF scores an NDCG@10 of 0.668 on AI/ML jobs against 0.712 on the others, and BM25 0.607 against 0.681, while E5 does not drop (0.748). E5's advantage over TF-IDF is therefore larger in this subgroup: +0.080 ($p = 0.023$, one exploratory test), possibly because resumes and job texts describe the same AI skills in different words, which embeddings can match and keywords cannot.

Table S1: Agreement between the author's labels and the dataset labels (150 pairs). CIs: bootstrap over pairs. For three labels, the linearly weighted κ is 0.174 [0.057, 0.296].

Decision	Pairs	Raw agreement	κ [95% CI]
Three labels	150	0.413	0.120 [0.015, 0.233]
Fit vs no fit	150	0.667	0.324 [0.177, 0.476]
Good vs Potential (both say fit)	64	0.406	−0.184 [−0.415, 0.063]

Table S2: Confusion matrix of the 150 relabelled pairs: dataset labels (rows) against the author's labels (columns).

	Human: No Fit	Human: Potential	Human: Good
Dataset: No Fit	36	7	7
Dataset: Potential Fit	17	12	21
Dataset: Good Fit	19	17	14

S4 Robustness checks

These checks were added after an external review of the study. Their decision rules were committed to the project's analysis plan before they were run.

Table S3: Score change (in SDs) when markdown is added to the original resume text with no change of wording (181 resumes, 2,076 pairs). Each step is compared with the flat version. All changes are significant (Holm $p < 0.05$) except those marked n.s.; TF-IDF ignores the symbols and does not change.

Method	Lines	Bold	Bullets	Bold + bullets	Full [95% CI]
TF-IDF	0.000	0.000	0.000	0.000	0.000
E5-base	+0.106	+0.718	+0.708	+0.806	+0.912 [0.861, 0.968]
BGE-base	+0.106	+0.091	+0.064	+0.078	+0.184 [0.153, 0.215]
MPNet-base	+0.024	+0.005 (n.s.)	+0.019	−0.012 (n.s.)	+0.012 [−0.048, 0.061]

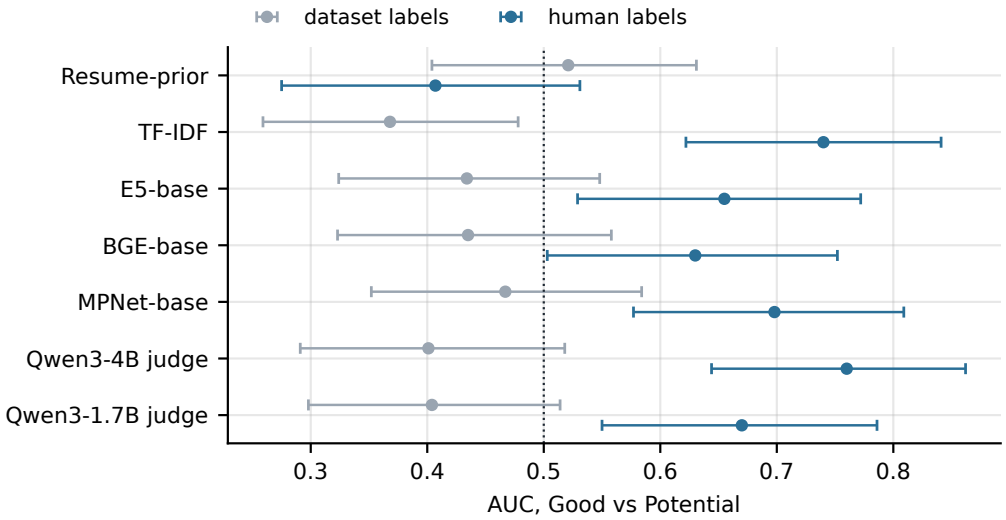


Figure S1: Good-vs-Potential AUC on the 150 relabelled pairs against the dataset labels and the author's labels (bars are 95% CIs; dotted line = chance). Job-aware methods agree more with the human labels; the resume-only prior does the opposite.

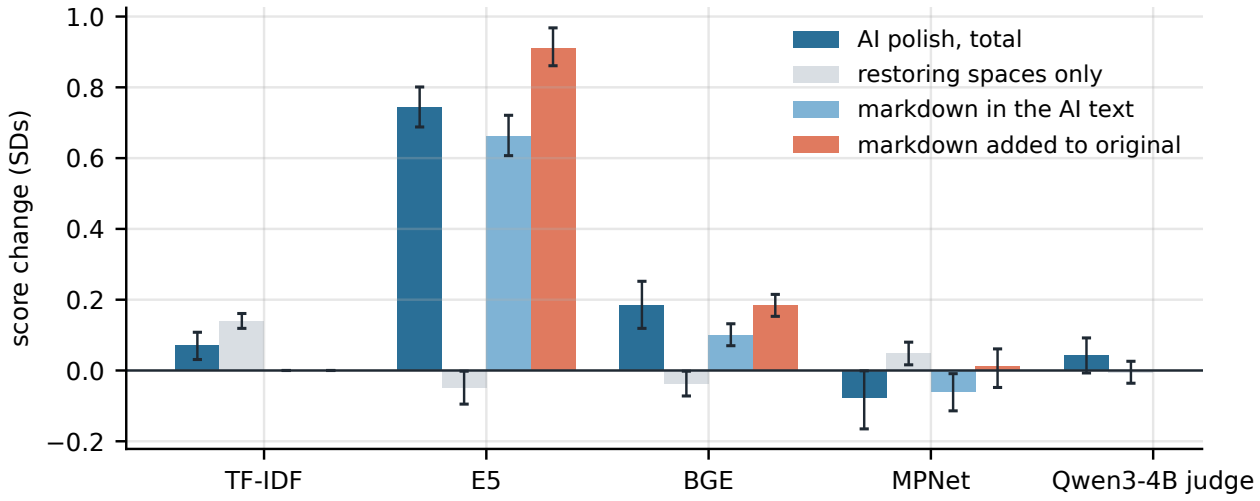


Figure S2: Score change (in SDs) from AI polishing (Table 9 of the main paper) and from adding markdown to the original resumes (Table S3).

Near-copy resumes and jobs

Resume and job IDs are hashes of the exact cleaned text (Section 3.2 of the main paper), so two texts that differ by a single character get different IDs. We call two different texts *near-copies* if their TF-IDF cosine (the settings of Section 4.1 of the main paper) is at least 0.90 or their word 5-shingle Jaccard similarity is at least 0.50. The closest pairs are the same resume with one separator character replaced throughout (for example "–" by "|") or with a date moved, and the same job posting for different cities. Table S4 gives the counts.

A test pair of the 5×5 grid is *affected* if its resume has a near-copy among that fold's strict training resumes, or its job has a near-copy among that fold's training jobs. Table S5 repeats the memorisation analysis without the affected pairs, using the saved scores of the same models (no retraining). The resume-text-only classifier was also re-run with its folds grouped by near-copy group, so that a resume and its near-copies are always in the same fold (Table S6).

Table S4: Near-copies in the dataset and the test pairs they affect. Main rule: cosine ≥ 0.90 or Jaccard ≥ 0.50; the other two rules are a sensitivity check.

	Resumes	Jobs
IDs	643	351
With a near-copy: cos ≥ 0.80 or Jac ≥ 0.30	152	46
With a near-copy: main rule	148	32
With a near-copy: cos ≥ 0.95 or Jac ≥ 0.70	129	15
Near-copy pairs (main rule)	78	18
Groups after merging near-copies (main rule)	568	334
Strict split (Section 5.1 of the main paper)		
Test resumes / jobs with a near-copy in training	25 of 434	5 of 53
Test pairs affected (resume or job)	208 of 1,105	
5×5 *grid*		
Test pairs affected (resume 249, job 155)	384 of 1,630 (23.6%)	

Table S5: Fine-tuned BGE in the 5×5 grid on all test pairs and without the pairs affected by near-copies. Differences: cluster bootstrap over test jobs (2,000 resamples) for AUC; job bootstrap and Wilcoxon signed-rank test for NDCG@5.

	All pairs (1,630; 306 jobs)		Without affected (1,246; 267 jobs)	
Method	AUC fit/no	NDCG@5	AUC fit/no	NDCG@5
FT standard	0.795	0.857	0.797	0.865
FT size-matched	0.787	0.846	0.789	0.850
FT strict	0.758	0.781	0.763	0.793
BGE-base (zero-shot)	0.659	0.734	0.661	0.746
E5-base (zero-shot)	0.644	0.722	0.642	0.745
TF-IDF	0.628	0.736	0.632	0.756
Standard − strict	+0.037 [0.017, 0.057]	+0.076 [0.048, 0.101]	+0.034 [0.009, 0.058]	+0.072 [0.044, 0.099]
Size-matched − strict	+0.029 [0.007, 0.051]	+0.065 [0.039, 0.091]	+0.026 [0.002, 0.051]	+0.057 [0.031, 0.083]
Standard − size-matched	+0.008 [−0.006, 0.022]	+0.010 [−0.005, 0.026]	+0.008 [−0.008, 0.023]	+0.015 [−0.005, 0.035]
Strict − BGE zero-shot	+0.098 [0.067, 0.130]	+0.047 [0.021, 0.072]	+0.102 [0.066, 0.139]	+0.047 [0.017, 0.080]
Strict − TF-IDF	+0.129 [0.098, 0.162]	+0.045 [0.013, 0.076]	+0.131 [0.093, 0.171]	+0.036 [0.002, 0.072]

Table S6: Resume-text-only classifier, per-job Good-vs-Potential AUC in the 91 jobs that use both fit labels, with cross-validation folds grouped by resume ID (as in the main paper's first analysis) or by near-copy group. CIs: bootstrap over jobs.

Classifier	Jobs	Grouped by resume ID	Grouped by near-copy group
Resume text	All mixed (91)	0.635 [0.591, 0.681]	0.598 [0.552, 0.645]
Resume text	Technical (61)	0.666 [0.611, 0.716]	0.623 [0.570, 0.677]
Resume text	Non-technical (30)	0.573 [0.496, 0.654]	0.546 [0.464, 0.631]
Tech words removed	All mixed (91)	0.602 [0.559, 0.645]	0.603 [0.558, 0.650]

Intervals from resampling whole jobs

The 150 relabelled pairs come from only 42 jobs (median 3 pairs per job, at most 11), and the second annotator's 40 pairs from 27 jobs, so pairs from the same job are not independent. Table S7 repeats the human-label intervals with 2,000 bootstrap resamples of whole jobs (seed 42). Point estimates do not change. Three κ intervals change side: H2 vs D moves above 0 for fit vs no fit and to include 0 for three labels, and H1 vs D for Good vs Potential (17 pairs) moves above 0; none changes the reading in the main paper, where the two-human comparison stays inconclusive. Every Good-vs-Potential shift of a job-aware method keeps its interval above 0 and stays significant. As in the analysis plan, E5 is tested on its own (uncorrected) and Holm correction covers the other five job-aware methods.

The two columns of Table 7 in the main paper use different pairs: the 100 pairs the dataset calls a fit and the 78 the author calls a fit. Table S8 repeats the comparison on the 64 pairs that both call a fit, so only the labels differ. The shifts are larger there.

Self-check: the author relabels 30 pairs

To see how stable the author's own judgement is, the author relabelled 30 of the 150 pairs about two weeks after the original labelling, under the same guide and blind conditions and with the old sheet closed. The 30 pairs were drawn before relabelling (10 per original author label, seed 20261019) after excluding the 26 strongest disagreements with the dataset, and were shown with new IDs in a new order. Some memory of the pairs is possible. Table S9 gives the result. Self-agreement on fit vs no fit ($\kappa = 0.426$ [0.146, 0.722]) is well above the agreement between the two humans (0.093), so the low inter-annotator agreement reflects different judges more than random labelling. The second pass was also stricter: one pair moved up and 13 of 30 moved down a level or two, so the author's threshold between levels was not fixed over time.

Table S7: Human-label results with intervals from resampling pairs (as in the main paper) and whole jobs. H1 = author, H2 = second annotator, D = dataset. Holm: over the five job-aware methods other than E5; E5 is the analysis plan's single pre-specified test and is not corrected.

	Estimate	CI, pairs	CI, jobs	p, pairs	p, jobs
κ, H1 vs D (150 pairs, 42 jobs)					
Three labels	0.120	[0.015, 0.233]	[0.020, 0.227]		
Fit vs no fit	0.324	[0.177, 0.476]	[0.185, 0.469]		
Good vs Potential (64 pairs)	−0.184	[−0.415, 0.063]	[−0.375, 0.039]		
κ, second annotator (40 pairs, 27 jobs)					
Three labels, H2 vs H1	0.115	[−0.109, 0.347]	[−0.123, 0.314]		
Three labels, H2 vs D	0.225	[0.006, 0.441]	[−0.002, 0.486]		
Three labels, H1 vs D	0.319	[0.086, 0.546]	[0.106, 0.562]		
Fit vs no fit, H2 vs H1	0.093	[−0.191, 0.395]	[−0.194, 0.374]		
Fit vs no fit, H2 vs D	0.319	[−0.003, 0.595]	[0.021, 0.614]		
Fit vs no fit, H1 vs D	0.340	[0.040, 0.632]	[0.077, 0.609]		
Fit vs no fit, (H2–H1) − (H2–D)	−0.226	[−0.550, 0.119]	[−0.578, 0.115]		
G vs P, H2 vs H1 (13 pairs)	0.270	[−0.250, 0.750]	[−0.250, 0.667]		
G vs P, H2 vs D (18 pairs)	0.143	[−0.286, 0.553]	[−0.250, 0.545]		
G vs P, H1 vs D (17 pairs)	0.422	[−0.031, 0.826]	[0.069, 0.857]		
Good-vs-Potential AUC shift, human − dataset labels (Table 7 of the main paper)					
Resume-prior (primary test)	−0.114	[−0.281, 0.058]	[−0.305, 0.134]	0.211	0.358
E5-base (single test)	+0.221	[0.058, 0.384]	[0.016, 0.399]	0.011	0.037
TF-IDF	+0.372	[0.205, 0.538]	[0.132, 0.573]	<0.001	0.018
BGE-base	+0.195	[0.031, 0.358]	[0.005, 0.391]	0.018	0.044
MPNet-base	+0.232	[0.063, 0.395]	[0.053, 0.384]	0.014	0.028
Qwen3-4B judge	+0.359	[0.194, 0.517]	[0.184, 0.524]	<0.001	<0.001
Qwen3-1.7B judge	+0.266	[0.085, 0.438]	[0.080, 0.428]	0.009	0.016

Table S8: Good-vs-Potential AUC on the 64 pairs that both the dataset and the author call a fit (dataset: 31 Good, 33 Potential; author: 35 Good, 29 Potential). Shift: paired bootstrap over these pairs (2,000 resamples, seed 42), uncorrected p.

Method	AUC, dataset	AUC, author	Shift [95% CI]	p
Resume-prior	0.548	0.387	−0.161 [−0.362, 0.051]	0.138
TF-IDF	0.273	0.705	+0.433 [0.225, 0.629]	<0.001
E5-base	0.328	0.669	+0.341 [0.150, 0.523]	<0.001
BGE-base	0.328	0.631	+0.302 [0.105, 0.489]	0.003
MPNet-base	0.335	0.693	+0.357 [0.159, 0.544]	<0.001
Qwen3-4B judge	0.289	0.736	+0.447 [0.259, 0.623]	<0.001
Qwen3-1.7B judge	0.340	0.633	+0.293 [0.088, 0.495]	0.001

Table S9: Self-check: original author labels (rows) against the same author's labels two weeks later (columns), 30 pairs. κ with 2,000 pair-bootstrap CIs (seed 42): three labels 0.300 [0.061, 0.546], raw agreement 0.533; fit vs no fit 0.426 [0.146, 0.722], raw agreement 0.700.

	Later: No Fit	Later: Potential	Later: Good
Original: No Fit	9	1	0
Original: Potential Fit	6	4	0
Original: Good Fit	2	5	3