EHR-BENCH: A CHINESE MULTI-TASK BENCHMARK FOR RECRUITMENT AI

Anonymous authors
Paper under double-blind review

ABSTRACT

Recruitment applications span diverse operations, from drafting job descriptions and interpreting queries to organizing profiles, assessing fit, and recommending candidates and jobs. Systematic LLM evaluation therefore requires task coverage that reflects this range of business needs. We introduce EHR-BENCH, an open Chinese benchmark comprising twelve tasks across six capability families and five recruitment scenarios. The tasks cover generation, extraction, classification, retrieval, semantic similarity, and reciprocal ranking within one evaluation framework. An internal recruitment dataset guides task definitions, the occupation taxonomy and sampling targets, output requirements, and scoring protocols. Following these specifications, we construct 8,336 task instances from public materials and controlled augmentation, with source-aware expert review and documented provenance. We evaluate twelve LLMs through task-level, capability-level, and aggregate comparisons. Closely ranked models exhibit distinct task strengths, and performance varies across input sources, reference evidence, and candidate pools. Five shared models retain the same observed Overall ordering on a private operational reference, with variation in capability and task correspondence. EHR-BENCH provides a reproducible foundation for systematic model assessment, task-specific model selection, and further research on LLMs for recruitment. Data, evaluation code, and per-sample scoring records are available at `https://anonymous.4open.science/r/ehr-bench-review`.

1 INTRODUCTION

Recruitment applications use LLMs to draft job descriptions, interpret queries, organize profiles, assess fit, and recommend candidates and jobs (Otani et al., 2025; Mashayekhi et al., 2024). These operations share resumes and job descriptions but require different outputs: drafted content, structured attributes, relationship judgments, and candidate orderings. Comparing models across these operations requires a task framework that makes their distinct requirements explicit.

Existing resources address resume classification (Heakl et al., 2024), skill extraction (Zhang et al., 2022), document parsing (Ling et al., 2025), and person–job retrieval (Wang et al., 2026); WorkRB provides a shared framework for work-domain tasks (De Lange et al., 2026). A complementary need is a public Chinese benchmark that connects diverse recruitment operations to enterprise task requirements. We address this need with shared task specifications, documented inputs and reference evidence, and task-specific scoring for aggregate and task-level model comparison.

We introduce EHR-BENCH (Figure 1), an open Chinese benchmark with 8,336 instances across twelve tasks, six capability families, and five recruitment scenarios. An internal recruitment dataset guides the task definitions, occupation taxonomy and sampling targets, output schemas, and scoring. These specifications guide construction from public materials and controlled augmentation, with source-aware expert review of reference labels. We release the data, provenance records, per-sample scoring exports, and evaluation tools for reproducing scores and inspecting task-level differences.

Our twelve-LLM evaluation identifies distinct task advantages through capability profiles and paired comparisons. The top four systems span 0.52 Overall points; task-level results guide selection for specific recruitment operations. Source and candidate-pool analyses further characterize performance across evaluation conditions. Five shared LLMs retain the same observed Overall order on the private operational reference that guided construction, while capability and task correspondence varies.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

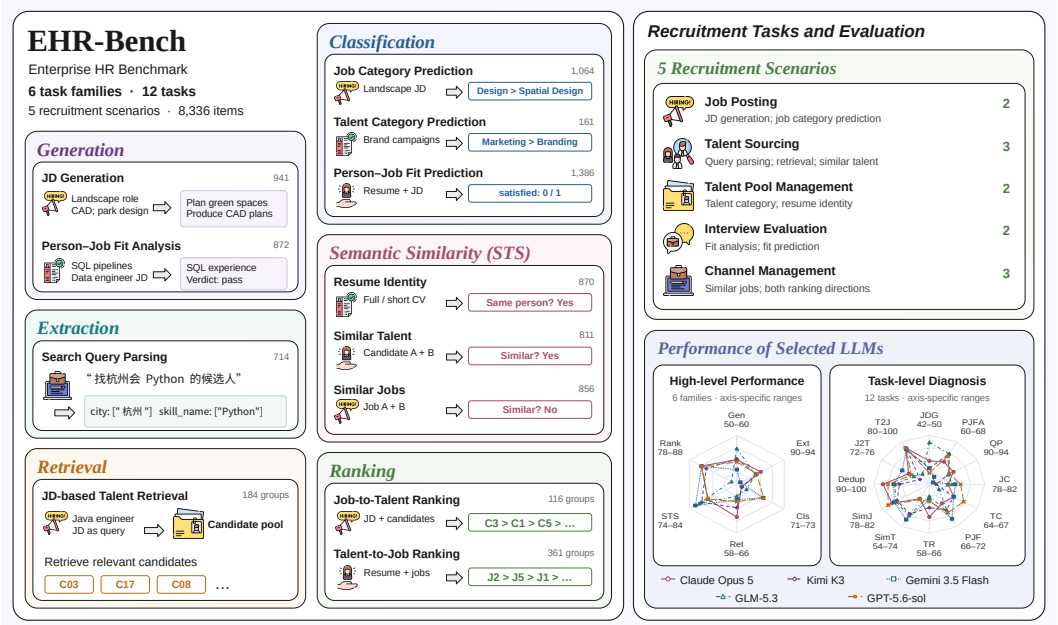

Figure 1: Overview of EHR-BENCH: twelve tasks, six capability families, and five recruitment scenarios. Examples are illustrative; English text is translated. Radars show five leading model series, with scores scaled from center to rim over the range labeled on each axis.

We contribute (1) a framework that translates five recruitment scenarios into twelve tasks and six capability families for cross-operation evaluation; (2) an open, expert-reviewed resource whose construction is guided by an internal recruitment dataset, with documented provenance and reusable evaluation tools; and (3) a twelve-LLM study showing that models with similar Overall scores have distinct task-specific advantages and that performance varies across input sources, reference evidence, and candidate pools, informing task-specific model comparison and improvement.

2 RELATED WORK

Recruitment tasks and public benchmarks. Public resources cover complementary operations: resume classification (Heakl et al., 2024), skill extraction (Zhang et al., 2022; Li et al., 2026), multilingual resume parsing (Ling et al., 2025), and resume–JD questions and explanations (Carrino et al., 2026). Candidate selection builds on requirement–experience representations (Zhu et al., 2018; Qin et al., 2018; Li et al., 2020), interaction evidence (Bian et al., 2020), and retrieval with hard-negative mining (Yu et al., 2025). PJB provides Chinese resume retrieval (Wang et al., 2026), while WorkRB supports ranking and multilabel classification (TechWolf Research Team, 2026). Table 1 contrasts their evaluation interfaces and candidate conditions. EHR-BENCH combines generation, structured interpretation, relationship judgments, and candidate selection across five recruitment scenarios, with enterprise-guided construction and shared public protocols. Appendix H compares languages, task scope, output requirements, and label evidence.

Multi-task evaluation and resource construction. HELM (Liang et al., 2023), MTEB (Muennighoff et al., 2023), BEIR (Thakur et al., 2021), and LegalBench (Guha et al., 2023) motivate diverse task coverage and capability-specific reporting. Synthetic CV resources (Bruera et al., 2022) and JobSkape (Magron et al., 2024) provide construction routes for employment text; data statements (Bender & Friedman, 2018) and datasheets (Gebru et al., 2021) organize provenance and intended use. EHR-BENCH connects task coverage, construction, and evaluation: internal recruitment operations guide task specifications and occupation sampling targets; public instances retain documented reference evidence; and shared protocols support comparisons across operations and between public and operational results for shared models.

Table 1: Evaluation interfaces and candidate conditions. WorkRB coverage includes the ranking and classification interfaces in its accompanying toolkit (TechWolf Research Team, 2026).

Resource	Language / output	Candidate conditions / scope
ResumeBench (Ling et al., 2025)	Five languages; PDF to JSON	Schema-guided resume parsing
PJB (Wang et al., 2026)	Chinese; resume retrieval	297 queries; 197,674-resume collection
WorkRB (De Lange et al., 2026)	Multilingual; ranking and classification	Ontology- and corpus-based candidate spaces
EHR-Bench (ours)	**Chinese; text, spans, paths, relations, orderings**	**12 operations; 8–50-candidate retrieval and 5–10-candidate ranking pools**

3 EHR-BENCH: DESIGN AND CONSTRUCTION

3.1 RECRUITMENT SCENARIOS AND TASK FRAMEWORK

Figure 1 organizes recruitment operations at three levels. Five scenarios provide the business contexts: job posting, talent sourcing, talent pool management, interview evaluation, and channel management. Six capability families group the abilities required across these contexts; twelve independently evaluated tasks define inputs, outputs, references, and metrics (Table 2; examples in Appendix A.6).

Generation (Gen) covers JD drafting and fit analysis; Extraction (Ext) recovers typed query spans; Classification (Cls) covers occupation assignment and binary fit judgments. Retrieval (Ret) and Ranking (Rank) evaluate candidate selection, with ranking in both directions. Semantic Textual Similarity (STS) covers relatedness and identity. Shared resumes and jobs thus support complementary tests of content production, structured interpretation, relationship judgments, and candidate decisions.

Table 2: Task definitions and abbreviations used throughout the paper. n counts records (†: query groups). R, J, V: ROUGE-L, rubric score, verdict accuracy; R@k: Recall@k.

Family	Task	Input → output	n	Main metric
Gen	Job description generation (JDG)	Job constraints → JD	941	$.6R + .4J$
	Person–job fit analysis (PJFA)	Resume + job → verdict, analysis	872	$.5V + .3R + .2J$
Ext	Query parsing (QP)	Query → typed spans	714	Strict span-F1
Cls	Job classification (JC)	Job text → category path	1,064	PathScore
	Talent classification (TC)	Resume → category path	161	PathScore
	Person–job fit (PJF)	Resume + job → binary fit	1,386	Balanced accuracy
Ret	Talent retrieval (TR)	Job + pool → candidate ordering	184†	Mean R@3, R@10, MRR@10
STS	Talent similarity (SimT)	Two profiles → relatedness	811	Macro-F1
	Job similarity (SimJ)	Two jobs → relatedness	856	Macro-F1
	Resume deduplication (Dedup)	Two resumes → identity	870	Macro-F1
Rank	Job-to-talent ranking (J2T)	Job + pool → talent ordering	116†	nDCG@5
	Talent-to-job ranking (T2J)	Resume + pool → job ordering	361†	nDCG@5

3.2 DATA CONSTRUCTION GUIDED BY THE INTERNAL DATASET

The internal dataset guides public task specifications (Appendix A.2): adopted JD drafts inform generation fields and rubrics; HR assessments distinguish fit decisions from verdict-plus-analysis outputs; annotated recruitment queries define entity schemas, while search and recommendation interactions inform candidate interfaces. Occupation labels supply a shared three-level taxonomy with eighteen first-level categories and sampling targets based on internal proportions. Public and constructed examples implement these specifications with source-specific references and expert review. Shared-model ordering is examined after construction (Section 6).

External inputs combine public profiles, JDs, interaction records, and controlled augmentation. Thirteen company job boards are accessed through the Greenhouse public API.

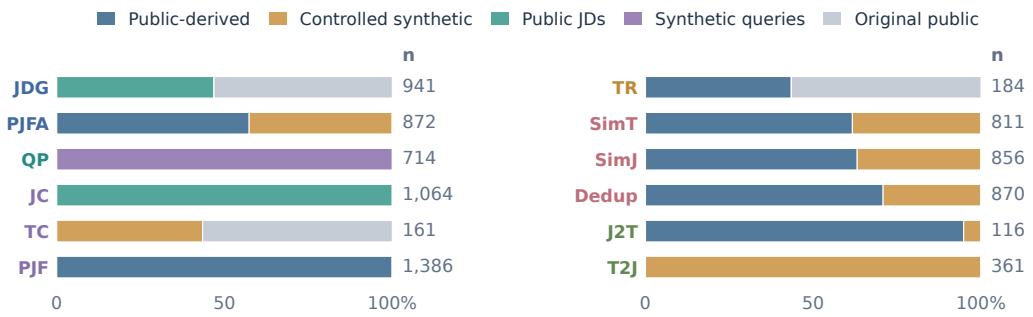

Figure 2: Source composition by task. Colors indicate five categories of record provenance and construction; counts treat each retrieval or ranking group as one evaluation instance.

These postings supply 441 JDG and 339 JC instances: GPT-5.6-sol translates English postings, humans proofread them, and cleaning removes company introductions and contact details. Native Chinese postings follow a separate pathway. Claude-opus-5 and GPT-5.6-sol each generate half of the persona-conditioned resume content, using occupation, experience, city, industry, and education seeds; experts assign their task labels (Appendix A.5). Resume-NER and a small Chinese adaptation of ranking examples provide further augmentation. Each instance retains its source identifier.

Reference labels retain task- and source-specific evidence. Public JDs supply generation references or receive occupation labels. Generated queries and candidate entity outputs undergo review; classification uses model-assisted label proposals. PJF combines 900 behavior-derived pairs, whose satisfaction labels undergo mapping and applicability checks, with 486 added pairs labeled by domain experts with initial suggestions from an internal fit model. SimT and SimJ seed pairs from shared-job and shared-candidate satisfaction, respectively, alongside occupational rules; experts determine final similarity labels from the paired profiles or job descriptions. Dedup identifies the same person under controlled profile transformations.

Candidate tasks specify the anchor, eligible pool, and relevance evidence: TR combines application-based and constructed pools; J2T uses satisfaction-derived relevance with a small constructed supplement; T2J uses constructed relevance. Each ranking group has one positive candidate; synthetic T2J fields named after behavior encode constructed relevance rather than observed events. Appendix A details the operational labels and construction rules.

3.3 Dataset Composition and Quality Assurance

Each retrieval or ranking group counts once among the 8,336 benchmark instances. TR has 184 groups (6,028 candidates), J2T 116 (847), and T2J 361 (2,888).

Figure 2 summarizes record sources and construction types by task. Constructed records may reuse public text; label evidence is specified separately. Within the enterprise reference's eighteen first-level occupations, job classification covers all eighteen and talent classification seventeen. TC sampling retains 161 of 735 records to target internal occupation proportions. Section 6 compares these cohorts with the operational reference.

Schema, range, duplicate, and cross-field checks detect structural problems; Qwen-plus flags potential semantic errors and disagreements. Three recruitment-domain experts confirm labels against guidelines and original evidence, preserving inherited meanings and adjudicating constructed relations. This source-aware workflow shares review principles with CC-OCR v2 and UNIKIE-BENCH (Peng et al., 2026; Ji et al., 2026). Independent double annotation covers 760 records; the sample-size-weighted mean of task-level Cohen's κ is 0.86 for the 474 records from six discrete-label tasks. A separate stratified audit confirms 395/400 records as correct. Appendix A.1 documents the review SOP, expert roles, and overlapping quality-check subsets.

4 EVALUATION PROTOCOL

The protocol combines task-specific metrics, shared model settings, and explicit output-handling rules to support reproducible model comparisons at task and capability levels.

4.1 TASK METRICS AND CAPABILITY AGGREGATION

Task-specific metrics evaluate the outputs in Table 2 on 0–100 scales; aggregation first averages within capabilities, then across them. JDG combines character-level ROUGE-L (R; Lin, 2004) and a rubric judge (J) as $0.6R+0.4J$: the components assess reference overlap and JD quality, respectively. PJFA adds verdict accuracy (V), using $0.5V + 0.3R + 0.2J$; all components use 0–100 scales. PJFA awards verdict credit only for a reference-matching decision extracted by the fixed parser; available analysis text is scored separately. The rubric-based evaluation follows the LLM-as-a-judge approach (Zheng et al., 2023). Component scores distinguish verdict agreement, reference overlap, and rubric quality. Query parsing uses corpus-level strict span micro-F1. Occupation classification divides the correctly matched prefix length by the reference depth. Similarity and identity tasks use macro-F1.

PJF uses balanced accuracy, $(\mathrm{TPR} + \mathrm{TNR})/2$, on parsed binary predictions, with invalid predictions mapped to the negative label (0) before calculation (Appendix B). Retrieval averages Recall@3, Recall@10, and MRR@10 within each group, then averages groups. Reciprocal ranking uses nDCG@5 (Järvelin & Kekäläinen, 2002) to score the single positive's position in each shortlist.

For tasks T_c in category c, the reporting protocol is

$$S_c = |T_c|^{-1} \sum_{t \in T_c} s_t, \qquad S_{\mathrm{Overall}} = \frac{1}{6} \sum_{c} S_c. \tag{1}$$

Task weights $1/(6|T_c|)$ are $1/6$ in Extraction/Retrieval, $1/12$ in Generation/Ranking, and $1/18$ in Classification/STS. Equal capability weights keep each family's contribution independent of its task count; individual metrics retain their task-specific meaning.

4.2 MODELS AND EVALUATION SETTINGS

We evaluate twelve LLMs and tabulate task, capability, and Overall scores to two decimal places. All 144 task scores are reproducible from per-sample exports (Appendix I). Qwen-plus screens candidate labels and Qwen-3.6-Flash applies common task-specific rubrics to generation outputs; neither is evaluated. Some evaluated models assist content construction; experts adjudicate labels.

The evaluation runner uses fixed Chinese task prompts, plain-text JD generation, and JSON for the other tasks. Temperature is zero except for Kimi-K3, which uses its endpoint default. Thinking is disabled where configurable; GLM-5.3 uses low reasoning effort. Occupation prompts contain the category tree and truncate input text at 3,500 characters. Ranking receives a candidate-dependent output budget. Appendix B.1 records system settings, release identifiers, and runner defaults; the artifact record identifies evaluation prompts, output parsers, generation rubrics, and task scorers.

4.3 OUTPUT HANDLING AND PAIRED COMPARISONS

Output handling and scored-set coverage follow task-specific rules. For generation, classification, binary judgments, and query parsing, missing or empty raw outputs are omitted; nonempty malformed outputs are handled by the task parser. Retrieval and ranking iterate over all supplied groups. A ranking must contain every candidate exactly once after the specified ID-repair rules, otherwise its group score is zero. Retrieval evaluates only returned top-ten slots, with invalid or duplicate IDs receiving zero relevance. All model–task sets are complete except two MiniMax-M3 refusals, one each in JC and SimJ. Appendix C reports their identifiers and the effect of counting them as failures.

Paired comparisons align scored sample IDs and use 10,000 bootstrap resamples of connected components of observed entity IDs, preserving complete query groups (Efron & Tibshirani, 1993). We report paired 95% bootstrap intervals for task-level differences on the common scored subsets. Appendix E lists dependence units, fixed seeds, and single-component cases without intervals.

5 Benchmark Results and Analysis

5.1 Overall Performance and Task-Specific Strengths

Table 3: Overall and capability scores. Bold / underline indicate the best / second-best distinct values, including ties. Overall averages the six capabilities.

System	Overall	Gen	Ext	Cls	Ret	STS	Rank
Claude-opus-5	**74.53**	55.03	**92.02**	71.20	**63.37**	80.10	**85.45**
Kimi-K3	<u>74.28</u>	<u>55.16</u>	91.66	71.23	<u>61.84</u>	81.38	84.42
GPT-5.6-sol	74.10	54.56	<u>91.68</u>	**72.14**	60.84	80.05	85.32
Gemini-3.5-flash	74.01	53.00	90.31	<u>72.10</u>	60.54	**82.69**	<u>85.41</u>
Claude-opus-4.8	73.81	54.92	91.36	71.51	60.44	79.84	84.81
GLM-5.3	73.54	**57.30**	91.57	71.41	59.95	81.92	79.11
GPT-5.5	73.28	54.24	89.19	71.89	60.96	79.15	84.25
Qwen3.8-max	72.88	54.14	90.79	71.34	61.66	75.98	83.39
GLM-5.2	72.50	52.35	89.49	71.56	61.08	<u>82.13</u>	78.36
Kimi-2.6	72.15	50.76	87.93	69.95	60.47	81.59	82.18
DeepSeek v4-Pro	70.94	49.84	88.86	65.86	59.47	80.86	80.72
MiniMax-M3	69.03	40.06	86.99	67.00	60.37	78.82	80.96

Overall scores range from 69.03 to 74.53 (Table 3), with a 0.52-point span across the top four. Across all twelve models, capability profiles reveal distinct strengths: Claude-opus-5 leads Extraction, Retrieval, and Ranking; GPT-5.6-sol leads Classification; Gemini-3.5-flash leads STS; and sixth-ranked GLM-5.3 leads Generation (57.30). Figure 3 extends this view to all twelve tasks, showing how strengths vary across individual recruitment operations within the same Overall ordering.

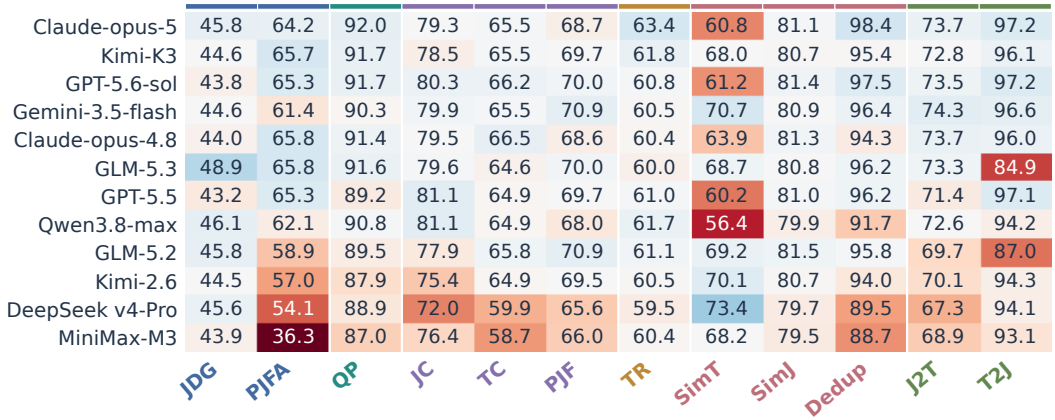

Figure 3: Twelve-task scores in Overall order. Red/blue indicate scores below/above each task's median, clipped at ± 15 points; colors show within-task relative position, not cross-task difficulty.

Paired comparisons identify task-level advantages among closely ranked models. Against Kimi-K3, Claude-opus-5 has a 1.19-point JDG advantage (paired 95% interval $[0.98, 1.40]$) and 2.99 on Dedup ($[1.65, 4.36]$), while Kimi-K3 leads PJFA by 1.43 ($[0.43, 2.41]$). These contrasts support task-specific model selection among closely ranked systems. Omitting Retrieval or Ranking changes the leading system (Appendix E.1); Overall and task profiles support different selection needs.

5.2 Task-Level Diagnostics

We examine taxonomy depth, generation components, and output validity to interpret task scores.

Occupation classification by taxonomy depth. Across all twelve systems, first-level job classification accuracy exceeds accuracy through the second level by 15.13–19.83 percentage points, identifying fine-grained occupation assignment as an improvement target (Appendix D, Figure 7b).

Generation components. GLM-5.3 has a higher PJFA composite than Claude-opus-5 (65.75 versus 64.25) and lower verdict accuracy (79.82% versus 80.96%). Higher ROUGE-L and rubric scores account for the difference in composite scores (Appendix D).

Output validity. MiniMax-M3's JSON parsing success rate on 872 PJFA outputs is 34.17%, compared with 97.82–100% for the other systems, making JSON-format compliance a major source of verdict-score loss. GLM-5.3 fails the complete-permutation requirement on 43 of 361 T2J groups, which receive zero in all-group nDCG@5 (Appendices D and D.2).

5.3 Performance across Evaluation Conditions

We compare performance across candidate pools, reference-label sources, public and synthetic inputs, and relation types under the same fixed task protocols.

Table 4: Candidate-task baselines by source (0–100). Uniform: expected score over permutations. LLM: the best whole-task system, scored within each stratum. Synthetic J2T has six groups.

Task	Source	n	Uniform	TF-IDF	LLM
TR	Public interactions	104	16.68	37.90	45.55
TR	Derived pools	80	53.34	72.22	86.53
J2T	Derived	110	43.10	56.97	74.16
J2T	Synthetic	6	44.75	13.63	76.03
T2J	Synthetic	361	36.86	89.05	97.19

Candidate-pool conditions. Retrieval scores are lower on public-interaction TR pools than on derived pools. The strongest whole-task LLM scores 45.55 and 86.53, respectively. On public-interaction pools, character-bigram TF-IDF scores 37.90 and uniform orderings 16.68; both controls also score higher on derived pools (Table 4). Synthetic T2J reaches 89.05 with TF-IDF alone. These controls characterize the candidate conditions underlying the LLM comparisons.

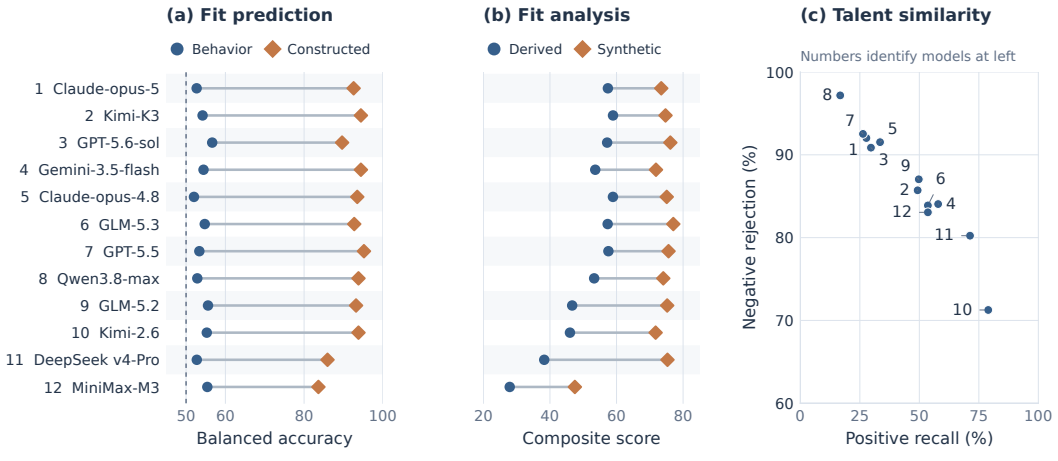

Figure 4: Performance across evaluation conditions (0–100). (a) PJF: behavior-derived ($n = 900$) and constructed (486); dashed line: chance at 50. (b) PJFA: interaction-derived (500) and synthetic (372). (c) SimT recall and negative rejection (811); numbers identify the systems listed at left.

Reference-label sources. PJF performance differs across reference-label sources under the same fit instruction. Across twelve LLMs, balanced accuracy is 52.01–56.64 on 900 behavior-derived pairs versus 83.64–95.22 on 486 constructed pairs (Figure 4a). Resume–job text cosine is inversely associated with satisfaction (AUC 0.400). A supervised TF-IDF probe reaches 60.30–61.01 balanced accuracy with both texts under resume- or job-grouped cross-validation; job-only features reach 62.72–63.28, versus 52.62–54.26 for resume-only features (Appendix F.1). These controls distinguish predictive job-side associations from pairwise text overlap, complementing the comparison of model agreement with the two reference sources.

Public and synthetic inputs. All systems score higher on synthetic PJFA and TC. Synthetic PJFA scores exceed interaction-derived scores by 15.77–37.03 points; TC spans 38.46–44.51 on public records versus 85.00–99.29 on synthetic records. The TC gap coincides with different input conditions: public profiles have a median length of 219 characters, compared with 979 for synthetic resumes. The gap identifies short public profiles as a challenging evaluation condition; length, content, and label depth vary together across the two source strata (Appendix F).

Relation types and decision profiles. Models differ in which related profiles they recover: SimT positive recall ranges from 16.75% to 78.95%, and negative rejection from 71.26% to 97.18% (Figure 4c). On SimJ, all systems label 77.30–95.04% of the same-category pairs with final negative labels as similar, while rejecting 99.50–100% of cross-category negatives. The narrow aggregate range (79.47–81.47) therefore masks low agreement on within-occupation negative pairs under the fixed similarity prompts used for all twelve evaluated systems (Appendix F.2).

6 Comparison with a Private Operational Reference

6.1 Shared Specifications and Population Comparison

After guiding public construction, the internal dataset provides an operational reference under shared task definitions, taxonomy, schemas, and scoring. We compare five shared LLMs: Kimi-K3, Qwen3.8-max, GLM-5.2, Kimi-2.6, and DeepSeek v4-Pro.

The classification cohorts provide a population comparison. External job/talent classification covers 18/18 and 17/18 first-level occupations, respectively, with total-variation distance 0.140 from the corresponding internal proportions. Labels retain their task-specific evidence: recruiter acceptance, HR judgment, public interactions, or constructed relevance. Appendix G.1 reports full occupation distributions together with input lengths, field coverage, and candidate-pool characteristics (Table 32).

6.2 Overall Agreement and Task-Level Variation

Both collections give the same observed Overall ordering for the five shared models (Spearman $\rho = 1.00$, ten of ten pairs in the same direction).

Omitting each capability in turn from both collections and averaging the remaining five gives cross-collection rank correlations of 0.90–1.00 at the reported score precision (Appendix G.3). The five-model aggregate orderings remain closely aligned under all six omissions.

Capability correspondence ranges from $\rho = 0.20$ for STS to 0.90 for Ranking, revealing variation beneath the shared Overall order (Figure 5).

Across twelve tasks, 78/120 pairwise comparisons agree in direction (65.0%), 40 disagree, and two tie externally. Ties remain in the denominator. These task-level comparisons locate differences in model ordering across the two collections (Appendix G).

The production suite further illustrates task-specific strengths: its internal PJF and T2J scores (74.30 and 65.98) exceed the best of the five LLMs (65.76 and 63.34), while its Overall is 63.79.

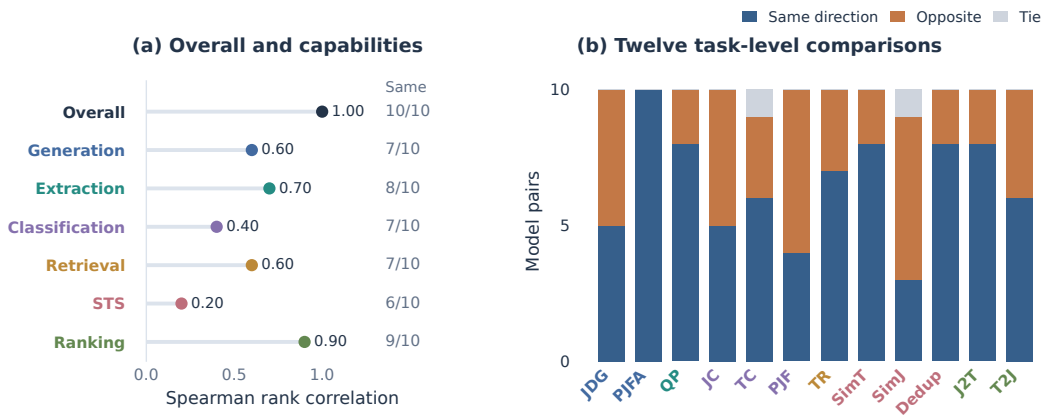

Figure 5: Correspondence on five shared models. (a) Overall and capability Spearman correlations, with same-direction counts out of ten pairs. (b) Direction counts for each task at displayed precision; ties remain in the denominator of ten model pairs per task.

For the five shared models, agreement in Overall ordering connects the public benchmark to the operational reference; capability and task differences guide task-specific model selection.

7 Discussion and Limitations

The released data, scorers, and per-sample scoring records support controlled comparisons of models and prompting strategies across twelve recruitment tasks. Capability and component profiles connect aggregate performance to specific operations, guiding model selection and targeted improvements under shared inputs, references, and candidate pools.

Source analyses compare models under documented input and reference conditions. Historical records supply interaction and satisfaction references, with experts checking source-label mappings and applicability and judging newly constructed relations. Content, length, and reference evidence vary together across source strata. Candidate tasks assess selection within supplied pools, with one positive per ranking group.

Experts assign task labels to model-generated resume content. Source-level comparisons across five multi-source tasks show task-dependent strengths for the two resume generators, with no consistent ranking advantage on synthetic records (Appendix F.3).

Scores measure task execution under the released prompts, specified inference settings, and fixed scoring protocols, including a shared generation rubric judge. These task-level comparisons inform model screening; application-specific assessment addresses hiring utility, exposure, and fairness (see Ethics statement; Raghavan et al., 2020; Kumar et al., 2023; Iso et al., 2025).

8 Conclusion

EHR-Bench establishes an open Chinese evaluation framework spanning twelve tasks, six capability families, and five recruitment scenarios. Guided by an internal business dataset, its construction translates enterprise task requirements into expert-reviewed public data, documented references, and reproducible protocols. The twelve-LLM study identifies task-specific strengths among closely ranked models and characterizes performance across input, reference, and candidate conditions. The task framework, open resource, and evaluation tools provide a shared basis for model comparison, task-specific model selection, and research on models and prompting strategies for recruitment.

ETHICS STATEMENT

The external collection combines public-source materials and constructed examples with source cleaning, annotation, and review. Its intended use is research evaluation of recruitment operations. Historical recruitment signals can reflect exposure and selection patterns; deployment requires assessment in the relevant organizational and fairness setting. Internal applicant and enterprise records remain in the controlled environment, with aggregate statistics used for the comparison.

REPRODUCIBILITY STATEMENT

All 144 model–task scores are reproduced offline from per-sample predictions and component statistics using the released scoring formulas, within 0.011 points on the 0–100 scale. Corpus, scorer, and result identifiers, coverage rules, and reproduction commands appear in Appendix I. The released artifacts support component, source, baseline, and paired analyses using saved predictions, component statistics, and rubric scores. Aggregate private-reference tables support reproduction of the correspondence statistics.

AI USE STATEMENT

AI models supported controlled generation and candidate-label preparation and checking. For persona-conditioned resumes, Claude-opus-5 and GPT-5.6-sol generated equal shares of the content and cross-reviewed each other's outputs. Qwen-plus assisted label-quality screening, and an internal fit model supplied initial judgments for added PJF pairs. A team of three domain experts reviewed final labels, checking inherited source labels and confirming or revising newly assigned judgments. Human reviewers also performed spot checks, independent double annotation, and anchor review. Qwen-3.6-Flash provided the common rubric scores for generated outputs. We distinguish data construction and annotation from output evaluation. The authors led the research and manuscript preparation. Generative AI tools assisted literature search, manuscript organization, paragraph drafting and language editing, and analysis-code implementation. The authors take responsibility for the final text, analyses, and conclusions.

REFERENCES

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041/.

Shuqing Bian, Xu Chen, Wayne Xin Zhao, Kun Zhou, Yupeng Hou, Yang Song, Tao Zhang, and Ji-Rong Wen. Learning to match jobs with resumes from sparse interaction data using multi-view co-teaching network. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pp. 65–74, 2020. doi: 10.1145/3340531.3411929. URL https://doi.org/10.1145/3340531.3411929.

Andrea Bruera, Francesco Alda, and Francesco Di Cerbo. Generating realistic synthetic curricula vitae for machine learning applications under differential privacy. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data*, pp. 53–63. European Language Resources Association, 2022. URL https://aclanthology.org/2022.legal-1.11/.

Casimiro Pio Carrino, Paula Estrella, Rabih Zbib, Carlos Escolano, and José A. R. Fonollosa. JobResQA: Semi-automatic multilingual benchmark creation for LLM machine reading comprehension on résumés and job descriptions. In *Proceedings of the Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026)*, pp. 161–176. ELRA Language Resources Association (ELRA), 2026. doi: 10.63317/4pzwmqxt46xp. URL https://aclanthology.org/2026.resourceful-4.15/.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

Matthias De Lange, Warre Veys, Federico Retyk, Daniel Deniz, Warren Jouanneau, Mike Zhang, Aleksander Bielinski, Emma Jouffroy, Nicole Clobes, Nina Baranowska, David Graus, Marc Palyart, Rabih Zbib, Dimitra Gkatzia, Thomas Demeester, Tijl De Bie, Toine Bogers, Jens-Joris Decorte, and Jeroen Van Hautte. WorkRB: A community-driven evaluation framework for AI in the work domain. *arXiv preprint arXiv:2604.13055*, 2026. URL https://arxiv.org/abs/2604.13055.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, first edition, 1993. URL https://www.routledge.com/An-Introduction-to-the-Bootstrap/Efron-Tibshirani/p/book/9780412042317.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010. URL https://doi.org/10.1016/j.patrec.2005.10.010.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44123–44279, 2023. doi: 10.52202/075280-1915. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/89e44582fd28ddfea1ea4dcb0ebbf4b0-Abstract-Datasets_and_Benchmarks.html.

Ahmed Heakl, Youssef Mohamed, Noran Mohamed, Aly Elsharkawy, and Ahmed Zaky. ResuméAtlas: Revisiting resume classification with large-scale datasets and large language models. *Procedia Computer Science*, 244:158–165, 2024. doi: 10.1016/j.procs.2024.10.189. URL https://www.sciencedirect.com/science/article/pii/S1877050924029909.

Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. Evaluating bias in LLMs for job–resume matching: Gender, race, and education. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 672–683. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-industry.55. URL https://aclanthology.org/2025.naacl-industry.55/.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418. URL https://doi.org/10.1145/582415.582418.

Yifan Ji, Zhipeng Xu, Zhenghao Liu, Zulong Chen, Qian Zhang, Zhibo Yang, Junyang Lin, Yu Gu, Ge Yu, and Maosong Sun. UNIKIE-BENCH: Benchmarking large multimodal models for key information extraction in visual documents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6331–6352, 2026. doi: 10.18653/v1/2026.acl-long.287. URL https://aclanthology.org/2026.acl-long.287/.

Deepak Kumar, Tessa Grosz, Navid Rekabsaz, Elisabeth Greif, and Markus Schedl. Fairness of recommender systems in the recruitment domain: An analysis from technical and legal perspectives. *Frontiers in Big Data*, 6:1245198, 2023. doi: 10.3389/fdata.2023.1245198. URL https://doi.org/10.3389/fdata.2023.1245198.

Changmao Li, Elaine Fisher, Rebecca Thomas, Steve Pittard, Vicki Hertzberg, and Jinho D. Choi. Competence-level prediction and resume & job description matching using context-aware transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8456–8466. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.679. URL https://aclanthology.org/2020.emnlp-main.679/.

Guojing Li, Zichuan Fu, Junyi Li, Wenxia Zhou, Xinyang Wu, Jinning Yang, Jingtong Gao, Feng Huang, and Xiangyu Zhao. Chinese-SkillSpan: A span-level dataset for ESCO-aligned competency extraction from Chinese job ads. *arXiv preprint arXiv:2604.23009*, 2026. URL https://arxiv.org/abs/2604.23009.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Zijian Ling, Han Zhang, Jiahao Cui, Zhequn Wu, Xu Sun, Guohao Li, and Xiangjian He. Beyond human labels: A multi-linguistic auto-generated benchmark for evaluating large language models on resume parsing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31919–31945. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1626. URL https://aclanthology.org/2025.emnlp-main.1626/.

Antoine Magron, Anna Dai, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. JobSkape: A framework for generating synthetic job postings to enhance skill matching. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources*, pp. 43–58. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.nlp4hr-1.4. URL https://aclanthology.org/2024.nlp4hr-1.4/.

Yoosof Mashayekhi, Nan Li, Bo Kang, Jefrey Lijffijt, and Tijl De Bie. A challenge-based survey of e-recruitment recommendation systems. *ACM Computing Surveys*, 56(10):1–33, 2024. doi: 10.1145/3659942. URL https://doi.org/10.1145/3659942.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Naoki Otani, Nikita Bhutani, and Estevam Hruschka. Natural language processing for human resources: A survey. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 583–597. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-industry.47. URL https://aclanthology.org/2025.naacl-industry.47/.

Chunyi Peng, Zhipeng Xu, Yuqi Xiong, Zulong Chen, Junhao Ji, Qing Liu, Zhenghao Liu, Zubao Qin, Bing Zhao, Jianqiang Wan, Jun Tang, Zhibo Yang, Shuai Bai, Dayiheng Liu, and Maosong Sun. CC-OCR V2: Fine-grained attribution of LMM failures in real-world visual document understanding. *arXiv preprint arXiv:2605.03903*, 2026. doi: 10.48550/arXiv.2605.03903. URL https://arxiv.org/abs/2605.03903.

Chuan Qin, Hengshu Zhu, Tong Xu, Chen Zhu, Liang Jiang, Enhong Chen, and Hui Xiong. Enhancing person–job fit for talent recruitment: An ability-aware neural network approach. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 25–34, 2018. doi: 10.1145/3209978.3210025. URL https://doi.org/10.1145/3209978.3210025.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 469–481, 2020. doi: 10.1145/3351095.3372828. URL https://doi.org/10.1145/3351095.3372828.

Eva Sánchez Salido, Roser Morante, Julio Gonzalo, Guillermo Marco, Jorge Carrillo-de Albornoz, Laura Plaza, Enrique Amigo, Andrés Fernandez García, Alejandro Benito-Santos, Adrián Ghajari Espinosa, and Victor Fresno. Bilingual evaluation of language models on general knowledge in university entrance exams with minimal contamination. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6184–6200. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.413/.

TechWolf Research Team. WorkRB: Work research benchmark, 2026. URL https://github.com/techwolf-ai/workrb/tree/c417039140a2c60d87c302d4106554707b90b1a0. Version 0.6.0; software snapshot accessed September 21, 2026.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

Guangzhi Wang, Xiaohui Yang, Kai Li, Jiawen He, Kai Yang, Ruixuan Zhang, and Zhi Liu. PJB: A reasoning-aware benchmark for person-job retrieval. *arXiv preprint arXiv:2603.17386*, 2026. URL https://arxiv.org/abs/2603.17386.

Xiao Yu, Ruize Xu, Chengyuan Xue, Jinzhong Zhang, Xu Ma, and Zhou Yu. ConFit v2: Improving resume–job matching using hypothetical resume embedding and runner-up hard-negative mining. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12775–12790. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.661. URL https://aclanthology.org/2025.findings-acl.661/.

Mike Zhang, Kristian Jensen, Sif Sonniks, and Barbara Plank. SkillSpan: Hard and soft skill extraction from English job postings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4962–4984. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.366. URL https://aclanthology.org/2022.naacl-main.366/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Chen Zhu, Hengshu Zhu, Hui Xiong, Chao Ma, Fang Xie, Pengliang Ding, and Pan Li. Person–job fit: Adapting the right talent for the right job with joint representation learning. *ACM Transactions on Management Information Systems*, 9(3):1–17, 2018. doi: 10.1145/3234465. URL https://doi.org/10.1145/3234465.

The appendices document construction and provenance (Appendix A), scoring and execution (Appendix B), full results and coverage (Appendix C), component and paired diagnostics (Appendices D–E), data conditions (Appendix F), and operational results (Appendix G), followed by resource comparisons and the artifact record.

A Task Interfaces and Construction Record

Each instance includes sample and task identifiers, language, input, and source metadata; reference labels are stored separately. Pair tasks supply two objects, and group tasks an anchor and eligible candidates. Generation returns a JD or a verdict with analysis. Query parsing returns typed entities whose surface forms are located in the query for exact span scoring. Occupation outputs are taxonomy paths; binary outputs express the task-specific relation. Retrieval and reciprocal ranking request complete orderings. Retrieval scores the first ten slots; reciprocal ranking requires a complete permutation. Prompts and schemas accompany the release.

Table 5: Operational label evidence for relation and candidate tasks. Final labels follow task-specific review of the initial construction.

Task	External positive evidence	External negatives / construction	Internal counterpart
PJF	Inherited satisfaction labels in 900 public pairs; expert judgments in 486 added pairs	Applied without satisfaction (215); browsed without application (207); expert-reviewed new pairings	Resume assessment pass / fail
TR	Application in public interaction pools; constructed relevance in derived pools	Other candidates within the supplied pool	Talent selected for assessment after JD-search exposure; sampled negatives
SimT	Shared-job satisfaction seeds; same second-level occupation seeds	Same-category or same-first-level negatives; cross-category random pairs; review sets final labels	Both candidates pass the final interview for the same job
SimJ	Shared-candidate satisfaction seeds; occupation-based pairs	Same-category and cross-category pairs; review sets final labels	Same candidate passes final interviews for both jobs
Dedup	Masked, lossy, or multiple snapshots of one identity	Distinct-person neighbors matched on occupation, education, or seniority; reviewed labels	Expert-labeled same-person / different-person pairs
J2T	Satisfaction-derived relevance; small constructed supplement	Remaining candidates in each single-positive pool	Recommended talent accepted by the user
T2J	Constructed compatibility, not an observed event	Remaining jobs in each single-positive pool	Recommended job applied to by the candidate

Browsing, application, and satisfaction are distinct public interaction signals. The private reference includes recruiter acceptance, candidate applications, and assessment outcomes. These event definitions specify the interpretation of each task's labels. Generation references, typed query annotations, and occupation labels are described in the source and review subsections below.

SimT assesses similarity in professional background, career direction, and seniority; SimJ assesses similarity in job responsibilities and skill requirements. In the behavior-derived portion, satisfaction records linking two talents to the same job, or the same talent to two jobs, seed candidate pairs. Shared interactions and occupation categories guide sampling; HR experts determine the final similarity labels from the paired profiles or job descriptions.

Layered annotation and review. We retain each candidate label's construction chain, distinguishing observed interactions or public annotations from deterministic constructions and model-assisted proposals. Schema, range, duplicate, and cross-field checks identify structural problems. Qwen-plus flags potential semantic errors and disagreements for the expert review described below; its suggestions do not determine the reference labels. CC-OCR v2 describes initial annotation, additional review, disagreement adjudication, and automatic consistency checks (Peng et al., 2026). UNIKIE-BENCH distinguishes converted from re-annotated data, using model-assisted verification and human review (Ji et al., 2026). These approaches motivate staged, source-aware verification; our final label decisions remain with the recruitment-domain expert team. Cross-model checks of generated resume content and rubric scoring of model outputs serve separate purposes (Appendices A.5 and B.2).

A.1 Human Annotation and Review

A team of three native Chinese domain experts with recruitment/HR and NLP/annotation experience reviews reference labels using shared task guidelines and original evidence. The standard operating procedure (SOP) in Figure 6 separates source-label verification from judgment of newly constructed relations. Qwen-plus supplies screening suggestions; experts retain final authority over label acceptance and revision.

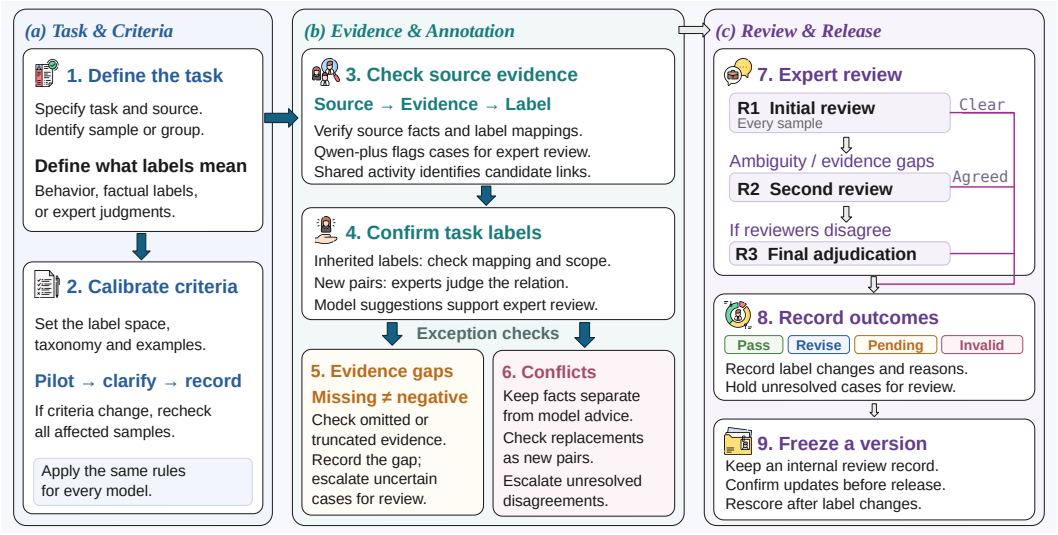

Figure 6: Reference-label review SOP. R1 reviews all samples; R2 handles cases requiring further review; R3 adjudicates disagreements.

Source-dependent label handling. For the initial 900 public behavior-derived PJF pairs, labels pass from the `satisfied` field in `table3_action.csv`, through the original label, to benchmark `reference.satisfied`. The satisfaction signal records HR judgments based on the same resume and job information supplied to the models. Review checks the mapping, conversion, and applicability of that label. The behavioral references comprise 478 satisfaction positives, 215 application-without-satisfaction negatives, and 207 browsing-without-application negatives.

The collection was subsequently expanded with 486 expert-labeled resume–job pairs. An internal person–job fit model provides an initial judgment as a review aid; domain experts determine the final label from the new pair's evidence. The added pairs include 162 positives and 324 negatives. The full collection retains public-data-derived input provenance (Appendix F.1).

For SimT and SimJ, experts judge similarity from the paired resumes or job descriptions. Shared interactions and occupation-based rules provide candidate pairs; expert review confirms or revises their proposed labels.

Three-expert collaboration. The team first calibrates task scope, label spaces, and boundary examples. R1 conducts the initial expert review of every sample. Cases with unclear provenance, missing evidence, or difficult judgments pass to R2 for additional review; R3 adjudicates disagreements using the original sources and task criteria. These roles form an escalation process, not three independent annotations of every record.

The SOP distinguishes an unmet requirement from an unstated or undetermined one. Insufficient evidence is flagged for review rather than automatically treated as a negative judgment. Recorded behavior, textual evidence, and model suggestions are kept distinct; verified mapping or conversion errors can be corrected without reinterpreting the underlying event. Review outcomes are recorded as pass, revise, pending, or invalid. Unresolved cases remain pending rather than being finalized as definite labels. Confirmed changes retain their rationale and version history; changes to task criteria require review of affected records under the updated rules. The review procedure does not use an evaluated model's score to select labels.

Business-specific procedures and item-level expert-review records are not part of the public release. The SOP documents the general workflow; aggregate quality checks are reported below.

Independent quality checks. Additional quality checks comprise source-stratified spot checks on 1,250 records, independent double annotation on 760, and review of 300 anchor examples. These subsets overlap. Double annotation starts with two independent judgments; a senior reviewer resolves disagreements. For PJF, both annotators independently assess fit from the public resume and job inputs, without access to source labels. These independent judgments are collected separately from the escalation workflow above. Ambiguous cases inform guideline revisions.

Table 6 summarizes agreement on the six discrete-label tasks. The sample-size-weighted mean of task-level Cohen's κ (Cohen, 1960) is 0.86 across 474 records. The remaining double-annotated records cover generation, extraction, and candidate tasks. A final stratified audit finds 395 of 400 records correct (98.75%), providing a separate check of the resulting labels.

Table 6: Independent annotation agreement before adjudication.

Task	Double-annotated records	Cohen's κ
JC	98	0.89
TC	20	0.83
PJF	125	0.86
SimT	74	0.80
SimJ	78	0.82
Dedup	79	0.92

A.2 Operational Basis of Task Design

Table 7 connects the operations represented in the enterprise reference to the twelve external tasks. Shared specifications guide external construction using public and constructed materials, with task-specific labels established through the evidence and review processes documented in this appendix.

Table 7: Business operations, corresponding tasks, and external construction guidance.

Operational evidence	**Tasks**	**External design and construction guidance**
Vacancy drafting: JD outputs adopted by platform users.	JDG	Required job fields, the JD-generation operation, and a shared quality rubric.
Candidate assessment: screening pass/reject decisions and HR-written interview assessments.	PJF, PJFA	Distinct fit-decision and verdict-plus-analysis outputs; shared analysis rubric.
Search understanding: recruitment queries with annotated entity spans.	QP	Typed entities, span-based extraction, and the query-output schema.
Profile organization: operationally maintained occupation labels for jobs and talents.	JC, TC	A shared hierarchical occupation space, category-path outputs, and occupation-based sampling targets.
Talent discovery: candidates scheduled for assessment after JD-search exposure.	TR	Job anchors, eligible candidate pools, and relevance-based retrieval evaluation.
Related-profile discovery: talents passing final interviews for the same job, or jobs for which the same talent passes.	SimT, SimJ	Recruitment-domain similarity criteria; shared interactions seed pairs for review, alongside random and occupation-related constructions.
Identity resolution: expert-labeled resume versions and different-person pairs.	Dedup	Same-person identity judgments, controlled profile variations, and negative-pair construction.
Reciprocal recommendation: accepted talent recommendations and jobs applied to by talents.	J2T, T2J	Direction-specific recommendation inputs, candidate-list interfaces, and relevance-based ordering.

A.3 Source Resources and Transformations

The source tags in Figure 2 describe record sources and construction types. Table 9 maps source materials to their uses. Source text, task transformation, and label evidence are separate fields: an interaction label and a constructed relation can use similar profile text while relying on different forms of reference-label evidence.

Table 8: Record-level source counts, including query groups. Percentages use 8,336 records and may not sum to 100 through rounding.

Source tag	Records	Percent
public_dataset	695	8.3
public_job_descriptions	1,505	18.1
public_data_derived	3,732	44.8
controlled_synthetic	1,690	20.3
synthetic	714	8.6

Table 9: Source resources, construction uses, and label evidence.

Resource / material	Construction use	Label evidence
Tianchi–Zhaopin competition data	Recruitment text and interaction-derived pairs or pools	Observed interaction fields for the original relation; task rules or reviewed judgments for constructed relations
`iic/resume_ner`	Resume text for talent classification and identity examples	Occupation labeling, same-person transformations, and negative pairs
Avera	Small Chinese adaptation used in reciprocal-ranking augmentation	Task-specific candidate relevance
Greenhouse company job boards	Cleaned JD references and occupation classification	Reference text or reviewed category labels
Controlled synthetic materials	Persona-based resumes and other task-specific constructions	Expert labels for persona-generated resumes; task-specific rules and review elsewhere

Original resources and license references. The recruitment profiles, job descriptions, and interaction records originate from the Tianchi–Zhaopin person–job matching competition and its preliminary dataset. The competition-data license reference is CC BY-SA 4.0. The `iic/resume_ner` resource is distributed by ModelScope's `iic` account; its copyright notice specifies CC BY 4.0. These references apply to the corresponding source materials and their derivatives.

Task preparation and review. DeepSeek v4-Pro assists the initial cleaning and interface standardization of public JDs. Task records follow the interfaces and occupation taxonomy in Section 3. Query references record the entities expressed in each query; relation labels use interaction evidence, construction rules, and expert judgments as appropriate to each task. Qwen-plus assists screening; experts verify source-label mappings and confirm labels for newly constructed examples (Appendix A.1).

The resource contains 8,336 records, counting each query group once. Expanding groups into candidate rows gives 17,438 rows, a separate measure of candidate exposure.

A.4 Public Job-Description Collection

The additional JD collection comprises over 1,000 postings from thirteen company careers boards, accessed through the Greenhouse Job Board API. The public JSON endpoint includes the posting text when `content=true`. Table 10 records the source boards; the collection count refers to retrieved

postings, while Table 2 counts final task instances. English posting content is translated into Chinese by GPT-5.6-sol and manually proofread. Cleaning also removes company introductions and contact details before task construction. The final collection contains 441 translated JDG references and 339 translated JC inputs. The other 725 JC inputs are native Chinese public postings; these share the public-JD provenance tag but not the translation pathway. The shared endpoint pattern is:

```
https://boards-api.greenhouse.io/v1/boards/{board}/jobs?content=true
```

Table 10: Company boards used for public JD collection; tokens link to JSON endpoints.

Company	Board token	Company	Board token
Stripe	`stripe`	Planet Labs	`planetlabs`
Anthropic	`anthropic`	Esri	`esri`
Coinbase	`coinbase`	Gotion	`gotion`
Cloudflare	`cloudflare`	Flexport	`flexport`
Roblox	`roblox`	project44	`project44`
Taboola	`taboola`	ShipMonk	`shipmonk`
Wolt	`wolt`		

A.5 Persona-Conditioned Resume Synthesis

Persona design. Seeds combine a public three-level occupation tree (18 first-level groups) with experience bands (new graduate, 1–3, 3–5, 5–10, and 10+ years), city, industry, and education. An occupation-specific skill dictionary constrains skill selection. These public persona categories specify the content-generation conditions.

Content generation and expert labels. Claude-opus-5 and GPT-5.6-sol each generate half of the synthetic resume content in this stage. The structured JSON includes basic information, education, work experience with quantified achievements, projects, and skills. Generation targets 800–2,000 Chinese characters and two to four work-experience sections. Human experts assign the associated task labels. The two generators also appear in the evaluated system set; their content-generation role is recorded separately from expert labeling and the common rubric judge used for output scoring.

Quality control. Rule checks cover schema completeness, timeline consistency, and required skills. Each generator's outputs are reviewed by the other model: Claude reviews GPT-generated resumes and GPT reviews Claude-generated resumes. Realism, internal consistency, and occupation fit are scored on a 1–5 scale, with a quality threshold of 4; outputs below the threshold are regenerated with feedback. MinHash deduplication completes the pipeline. Generator and evaluator comparisons appear in Appendix F.3.

A.6 Illustrative Task Instances

Table 11 uses author-written, English-language illustrations of the Chinese task interfaces. The snippets are schematic inputs and target forms, not dataset records or model predictions; identifiers and people are fictitious. For a resume and job, a fit decision returns a binary label, fit analysis pairs a verdict with an explanation, and candidate selection introduces a supplied comparison pool. The same objects therefore support distinct evaluation operations.

Table 11: Illustrative task inputs and outputs; field names follow evaluation prompts.

Task	Illustrative input	Target form and task meaning
JDG	Data engineer; SQL; maintain data pipelines.	A reference job description stating the role's duties and requirements.
PJFA	Resume mentions SQL pipelines; job requires SQL and pipeline maintenance.	`verdict: pass`; `analysis_text`: explanation of the stated experience and requirements.

Task	Illustrative input	Target form and task meaning
QP	"Find SQL engineers in City A."	`skill_name: [SQL]`, `jobname: [engineers]`, `city: [City A]`; surface strings are mapped back to exact query spans.
JC	A job description focused on data-pipeline development.	`cate_labels`: an ordered occupation path selected from the supplied taxonomy.
TC	A resume describing data-pipeline development experience.	`cate_labels`: the talent's occupation path in that taxonomy, rather than a label for a job vacancy.
PJF	A supplied resume–job pair.	`satisfied: 1` or `0`; the reference relation follows the pair's behavioral or constructed label evidence.
TR	A job and candidates with IDs `r1`, `r2`, ...	`ranking: [r2, r1, ...]`; the prompt requests all IDs, while relevance is evaluated in the first ten slots.
SimT	Profile A: mid-level Java backend developer. Profile B: Java engineer with comparable backend experience.	`is_similar: 1`; similar professional direction and seniority.
SimJ	Job A: build Java backend services. Job B: develop Java APIs and maintain backend services.	`is_similar: 1`; similar responsibilities and skill requirements.
Dedup	Two versions of a fictitious person's resume, with one shortened.	`is_same_person: 1`; the target is identity under the specified transformations.
J2T	A job and a five-person eligible pool, `r1`–`r5`.	`ranking: [r3,r1,r5,r2,r4]`; all five IDs must occur exactly once, and the labeled positive determines nDCG@5.
T2J	A resume and an eight-job pool, `j1`–`j8`.	`ranking: [j2,j5,j1,j3, j4,j6,j7,j8]`; a full permutation in the reciprocal direction.

B Metric and Execution Details

The benchmark first averages tasks within each capability, then averages the six capabilities. Each Extraction or Retrieval task contributes $1/6$ of Overall; each Generation or Ranking task $1/12$; each Classification or STS task $1/18$. The weights depend only on the task design, so adding a system does not change an existing system's score. Ranks and score ranges are descriptive views of the reported results. Scores are not standardized or weighted by observed dispersion.

All metric descriptions here refer to the scoring implementation. ROUGE-L is computed over Chinese characters. JD generation uses $0.6R + 0.4J$; PJF analysis uses $0.5V + 0.3R + 0.2J$, with each component on a 0–100 scale. The judge is Qwen-3.6-Flash under the task-specific rubric. Occupation PathScore is the correctly predicted prefix length divided by the reference path depth, with the first level serving as the gate for subsequent levels. PJF balanced accuracy consumes parsed binary outputs. The scorer names this metric AUC: for binary scores, ROC-AUC (Fawcett, 2006) equals balanced accuracy when both classes are present, using average ranks for ties. Invalid predictions map to the negative label (0) before this calculation; macro-F1 retains an invalid prediction as an error for its true class.

Retrieval evaluates the explicitly returned first ten slots. Duplicate and invalid identifiers occupy a slot with zero relevance. Recall uses the number of positives in the full supplied pool as its denominator, and MRR is truncated at those ten slots. Reciprocal ranking requires a complete permutation of the supplied candidates; a missing, duplicate, or unrepairable identifier makes the group score zero. The shared ID-repair routine accepts exact IDs, unique valid prefixes, or a unique fuzzy match above

similarity 0.9 with a margin greater than 0.05. No missing candidate is appended from the input order. The two directions use nDCG@5; MRR and other cutoffs are diagnostic outputs.

Missing/empty raw predictions are skipped by generation, query, occupation, and binary scorers. A nonempty malformed query produces an empty entity set; malformed occupation output yields no matched prefix. Generation uses the raw text fallback for text comparison, and PJFA verdict comparison applies to the parsed field. Ranking and retrieval iterate over the full task records, retaining empty-output groups as zero. These rules define the scored-set denominators.

B.1 Execution Interfaces and System Identifiers

Table 12 summarizes decoding settings and the evaluation runner's input and output budgets. Temperature is 0.0 for all systems except Kimi-K3, whose endpoint uses its default temperature. Thinking is disabled where configurable; GLM-5.3 uses `reasoning_effort=low`. DeepSeek v4-Pro was accessed through the vendor's official API using release `0813`. Qwen-3.6-Flash scores both collections with the same script, formulas, and rubrics.

Table 12: Decoding settings and input/output budgets.

Setting	Implementation
Prompt / output	Chinese task instructions; plain-text JD generation, task-specific JSON otherwise. Retrieval and ranking request all candidate IDs.
Temperature (runs)	0.0; Kimi-K3 omits this parameter and uses the endpoint default.
Reasoning (runs)	Thinking disabled where configurable; GLM-5.3 uses low reasoning effort.
Output budget (runner)	Default 4,096 tokens outside ranking; ranking uses $\max(16{,}384,\ 60N + 8{,}192)$ for N candidates.
Input truncation	Occupation: 3,500 characters; PJF/PJFA: 3,000 per text; retrieval candidate experience: default 300; ranking candidate profile: 400.

The result manifest maps leaderboard names to per-sample export folders for reproducing the diagnostic panels. It records the judge separately from evaluated models.

B.2 Generation Rubrics and Component Interpretation

The two generation tasks combine complementary measurements. Character-level ROUGE-L compares wording with a reference. PJFA additionally evaluates the structured verdict against its task label. Qwen-3.6-Flash scores generated content using the operational rubrics in Table 13. These components retain separate values in the result exports and in Table 16.

Table 13: Rubric dimensions and their scoring scales in the rubric judge.

Task	Dimensions	Score formation
JDG	Structure; alignment with job requirements; professionalism; actionable requirements; concision; consistency; attractiveness	Each dimension 0–3; weights 3, 3, 2, 2, 1, 2, 2; normalized to 0–100.
PJFA	Reasonableness of strengths/weaknesses; supporting evidence; clarity of structure; professionalism	Four dimensions of 0–25; total 0–100.

For PJFA, the normal parsed path passes `analysis_text` to the rubric judge; the structured `verdict` is scored separately. Verdict credit requires both successful extraction under the shared parsing protocol and agreement with the reference. Failed extraction or a nonmatching verdict gives $V = 0$; even maximal text components then yield at most 50 composite points. A raw-text fallback applies when the analysis field is empty or unavailable. The explanation itself can contain a conclusion, but text credit cannot restore the lost verdict component or establish JSON compliance. Accordingly, reference-verdict agreement and rubric quality characterize different aspects of an output; their conjunction is visible in the component and conditional-score tables (Appendix D).

The judge receives the generated text (up to 2,500 characters). PJFA also includes up to 2,500 reference-text characters and source context assembled from the first 800 characters of each resume and job, then truncated jointly to 1,200 characters. JDG receives up to 1,200 characters of its structured job constraints. These are judge-side budgets: source context is shortened for 400/872 PJFA records (45.9%). The reference explanation is visible to the judge, so its assessment is reference-conditioned. The evaluated-system input budgets are given in Table 12. The supplied scorer fixes these interfaces across systems, while retaining verdict accuracy as a separate decision measure.

C Full External Results and Scored-Set Coverage

Table 14: Complete external task scores. Bold / underline mark the best / second-best distinct displayed values, including ties. Abbreviations follow Table 2.

	Gen		Ext	Cls			Ret	STS			Rank	
System	JDG	PJFA	QP	JC	TC	PJF	TR	SimT	SimJ	Dedup	J2T	T2J
Claude-opus-5	45.81	64.25	**92.02**	79.35	65.53	68.71	**63.37**	60.80	81.11	**98.39**	73.72	**97.19**
Kimi-K3	44.63	65.68	91.66	78.48	65.53	69.68	61.84	68.04	80.70	95.40	72.78	96.06
GPT-5.6-sol	43.84	65.28	91.68	80.26	66.15	70.01	60.84	61.23	81.44	97.47	73.49	97.16
Gemini-3.5-flash	44.62	61.39	90.31	79.87	65.53	70.91	60.54	70.71	80.93	96.44	**74.26**	96.56
Claude-opus-4.8	44.03	**65.82**	91.36	79.51	**66.46**	68.56	60.44	63.86	81.32	94.35	73.66	95.96
GLM-5.3	**48.86**	65.75	91.57	79.64	64.60	69.98	59.95	68.74	80.81	96.21	73.30	84.92
GPT-5.5	43.21	65.27	89.19	**81.09**	64.91	69.66	60.96	60.22	81.02	96.20	71.36	97.13
Qwen3.8-max	46.13	62.14	90.79	81.08	64.91	68.02	61.66	56.39	79.86	91.68	72.60	94.18
GLM-5.2	45.84	58.86	89.49	77.93	65.84	**70.92**	61.08	69.17	**81.47**	95.75	69.74	86.98
Kimi-2.6	44.52	56.99	87.93	75.39	64.91	69.54	60.47	70.07	80.70	94.01	70.07	94.29
DeepSeek v4-Pro	45.61	54.08	88.86	72.02	59.94	65.62	59.47	**73.42**	79.68	89.48	67.32	94.13
MiniMax-M3	43.85	36.26	86.99	76.36	58.70	65.95	60.37	68.24	79.47	88.74	68.86	93.07

All twelve systems have the task counts in Table 2, with two MiniMax-M3 refusals documented below. Ranking validity is assessed separately in Appendix D.

Refusals and scored-set coverage. MiniMax-M3 declines two inputs: `post_class::000488` in JC and `sim_post::000067` in SimJ. Its scored counts are therefore 1,063/1,064 and 855/856, respectively. All other systems return outputs for these two inputs, and every other model–task set is complete. The task scorers omit these two absent predictions from the main scores.

Table 15 also counts both refusals as failures: zero PathScore for JC and a true-class error for SimJ, with macro-F1 recomputed from aggregate counts. The change is less than 0.1 point per affected task and leaves the Overall ordering unchanged.

Table 15: MiniMax-M3 scores with refusals omitted or counted as failures.

Task	Scored / total	Main score	Refusals as failures
JC	1063/1064	76.36	76.28
SimJ	855/856	79.47	79.41

D Component and Execution Diagnostics

Figure 7 summarizes generation components, occupation depth, and ranking validity; the following tables give their full task-level breakdowns.

GLM-5.3 has a higher PJFA composite than Claude-opus-5 (65.75 versus 64.25) and lower verdict accuracy (79.82 versus 80.96). Higher ROUGE-L and rubric scores account for the difference in composite scores.

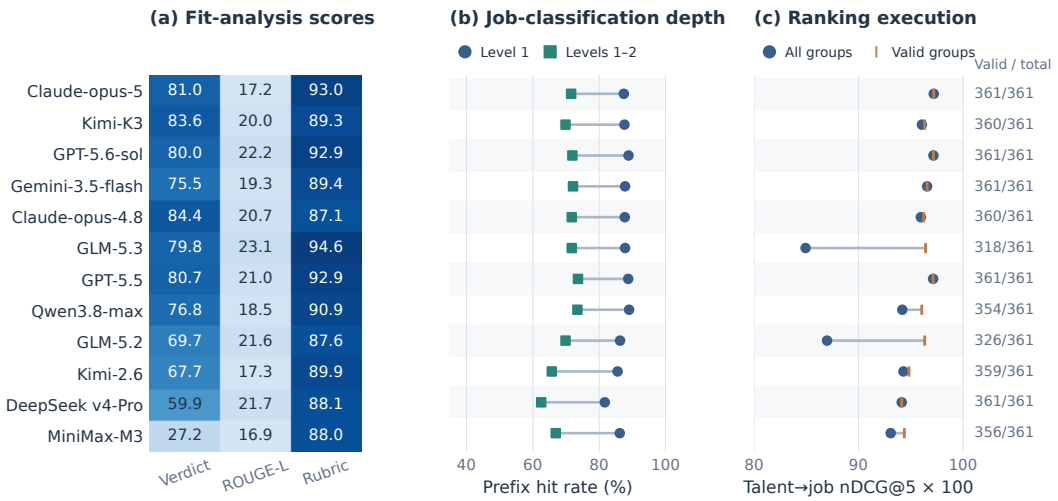

Figure 7: Twelve-system diagnostics in Overall order. (a) PJFA components (0–100; $n = 872$). (b) JC prefix hit rates ($n = 1{,}064$; MiniMax-M3: $1{,}063$). (c) T2J nDCG@5 on all 361 groups versus valid groups, with valid/total counts.

For GLM-5.3 relative to Claude-opus-5, the six category differences divided by six sum to -0.985 Overall points from displayed scores: Generation contributes $+0.378$ and Ranking -1.057. This is a decomposition of the fixed aggregation rule.

Table 16: Generation components (0–100). JDG weights ROUGE-L/rubric by 0.6/0.4; PJFA weights verdict/ROUGE-L/rubric by 0.5/0.3/0.2.

	Job description generation			Person–job fit analysis			
System	ROUGE-L	Rubric	Score	Verdict	ROUGE-L	Rubric	Score
Claude-opus-5	17.01	89.01	45.81	80.96	17.24	92.96	64.25
Kimi-K3	20.01	81.54	44.63	83.60	20.03	89.34	65.68
GPT-5.6-sol	19.70	80.04	43.84	80.05	22.25	92.93	65.28
Gemini-3.5-flash	16.54	86.74	44.62	75.46	19.26	89.43	61.39
Claude-opus-4.8	20.43	79.43	44.03	84.40	20.67	87.09	65.82
GLM-5.3	20.69	91.12	48.86	79.82	23.05	94.64	65.75
GPT-5.5	19.17	79.29	43.21	80.73	21.05	92.94	65.27
Qwen3.8-max	19.71	85.75	46.13	76.83	18.47	90.91	62.14
GLM-5.2	19.97	84.65	45.84	69.72	21.56	87.63	58.86
Kimi-2.6	19.48	82.09	44.52	67.66	17.25	89.93	56.99
DeepSeek v4-Pro	19.64	84.57	45.61	59.86	21.74	88.13	54.08
MiniMax-M3	19.79	79.93	43.85	27.18	16.91	88.01	36.26

At level k, the eligible set contains records with reference depth at least k. All systems correctly match the six depth-three TC records; Table 17 reports the first two TC levels and all three JC levels.

GLM-5.3 fails the complete-list requirement on 43 of 361 T2J groups, which receive zero credit. Its official all-group nDCG@5 is 84.92; Figure 7c reports valid-group scores.

Table 17: Occupation prefix accuracy (percent), with eligible counts in parentheses. A hit requires the full prefix through that level.

	Job classification			Talent classification	
System	L1	L1–2	L1–3	L1	L1–2
Claude-opus-5	87.4 (1064)	71.5 (1064)	87.5 (32)	71.4 (161)	59.6 (161)
Kimi-K3	87.6 (1064)	69.8 (1064)	75.0 (32)	73.3 (161)	57.8 (161)
GPT-5.6-sol	88.8 (1064)	71.9 (1064)	90.6 (32)	72.7 (161)	59.6 (161)
Gemini-3.5-flash	87.8 (1064)	72.1 (1064)	93.8 (32)	71.4 (161)	59.6 (161)
Claude-opus-4.8	87.7 (1064)	71.7 (1064)	78.1 (32)	72.0 (161)	60.9 (161)
GLM-5.3	87.8 (1064)	71.7 (1064)	78.1 (32)	70.8 (161)	58.4 (161)
GPT-5.5	88.7 (1064)	73.6 (1064)	93.8 (32)	72.0 (161)	57.8 (161)
Qwen3.8-max	89.0 (1064)	73.4 (1064)	87.5 (32)	71.4 (161)	58.4 (161)
GLM-5.2	86.3 (1064)	69.8 (1064)	81.2 (32)	73.3 (161)	58.4 (161)
Kimi-2.6	85.5 (1064)	65.7 (1064)	71.9 (32)	72.7 (161)	57.1 (161)
DeepSeek v4-Pro	81.7 (1064)	62.5 (1064)	84.4 (32)	68.3 (161)	51.6 (161)
MiniMax-M3	86.2 (1063)	66.9 (1063)	75.0 (32)	67.1 (161)	50.3 (161)

Table 18: Groups satisfying the complete-permutation ranking protocol. Main nDCG@5 averages over all groups, with zero for groups failing the protocol.

System	Job→talent	Talent→job
Claude-opus-5	116/116	361/361
Kimi-K3	116/116	360/361
GPT-5.6-sol	116/116	361/361
Gemini-3.5-flash	116/116	361/361
Claude-opus-4.8	116/116	360/361
GLM-5.3	115/116	318/361
GPT-5.5	116/116	361/361
Qwen3.8-max	116/116	354/361
GLM-5.2	113/116	326/361
Kimi-2.6	116/116	359/361
DeepSeek v4-Pro	116/116	361/361
MiniMax-M3	116/116	356/361

D.1 Joint Reading of Verdict and Explanation Scores

For each system, Table 19 groups the 872 PJFA outputs by the scored verdict: reference-matching ($V = 100$) and reference-nonmatching ($V = 0$). It reports the rubric mean for both subsets. The rubric evaluates the explanation text under the criteria in Appendix B.2; the verdict component measures the structured decision. The composite assigns half its weight to verdict agreement, ensuring that the decision contributes directly. Nonmatches include verdicts unavailable under the required JSON interface and extracted verdicts that disagree with the reference.

Table 19: PJFA rubric conditional on parsed verdict agreement. Each system's two subsets sum to 872 outputs; rubric scores are on a 0–100 scale.

	Reference-matching		Reference-nonmatching	
System	n	Mean rubric	n	Mean rubric
Claude-opus-5	706	93.34	166	91.34
Kimi-K3	729	90.14	143	85.24
GPT-5.6-sol	698	93.18	174	91.93
Gemini-3.5-flash	658	90.32	214	86.70
Claude-opus-4.8	736	88.45	136	79.74
GLM-5.3	696	95.16	176	92.56
GPT-5.5	704	93.32	168	91.35
Qwen3.8-max	670	91.71	202	88.27
GLM-5.2	608	88.87	264	84.77
Kimi-2.6	590	90.92	282	87.85
DeepSeek v4-Pro	522	90.39	350	84.75
MiniMax-M3	237	91.41	635	86.74

Across all systems, the mean rubric on reference-nonmatching verdicts is 79.74–92.56. Joint component reporting distinguishes explanation ratings from successful delivery of a matching verdict.

D.2 PJFA JSON Output Compliance

Table 20 audits all 872 nonempty PJFA responses per system using the same released JSON parser. It attempts code-fenced JSON, the full response, and an embedded object from the first opening to the last closing brace, each requiring successful JSON decoding. The verdict-availability column measures extraction through this parser. A parsing failure receives zero for the verdict component; explanation scoring can still use the raw-text fallback.

Table 20: PJFA output parsing under the shared evaluation protocol. Verdict availability measures successful extraction through the released JSON parser.

System	Nonempty outputs	JSON parsed	Parse rate (%)	Verdict available (%)
Claude-opus-5	872	858	98.39	98.39
Kimi-K3	872	870	99.77	99.77
GPT-5.6-sol	872	872	100.00	100.00
Gemini-3.5-flash	872	856	98.17	98.17
Claude-opus-4.8	872	869	99.66	99.66
GLM-5.3	872	871	99.89	99.89
GPT-5.5	872	872	100.00	100.00
Qwen3.8-max	872	872	100.00	100.00
GLM-5.2	872	871	99.89	99.89
Kimi-2.6	872	853	97.82	97.82
DeepSeek v4-Pro	872	872	100.00	100.00
MiniMax-M3	872	298	34.17	34.17

MiniMax-M3 yields decodable JSON for 298/872 responses (34.17%), compared with 97.82–100% for the other systems. An audit using regular-expression extraction finds one unique verdict in all 872 responses without requiring valid enclosing JSON; 611 verdicts match the reference (70.07%). These findings identify JSON-format compliance as a major source of verdict-score loss. Table 19 groups outputs by the shared scoring protocol, under which available explanation text can still receive rubric credit when verdict extraction fails.

E COMMON-SUBSET PAIRED DIAGNOSTICS

GLM-5.3, DeepSeek v4-Pro, and MiniMax-M3 are compared with Claude-opus-5 to examine generation, binary-relation judgments, and ranking performance. All twelve tasks are retained for each pair. Differences are comparator minus Claude-opus-5; scores and intervals use the intersection of scored sample IDs.

Records sharing observed job/person IDs, including pool candidates, form connected components on the full task before restriction to the common subset. Records without entity IDs form singletons. We resample represented components with replacement, preserving record multiplicities and complete query groups. The analysis supplement records identification coverage and component sizes. Reused candidates connect the talent-to-job graph into one component; its paired difference is descriptive.

Percentile 95% intervals use 10,000 replicates with seed 20260916. Each replicate recomputes F1 from class/span counts, balanced accuracy from class recalls with the scorer's invalid-output mapping, and other scores from record/group totals. Single-class balanced-accuracy replicates are omitted and counted. These exploratory intervals describe the common subset and observed identity links.

Table 21: Paired differences against Claude-opus-5. n: common records/groups; C: observed-entity components. A dash denotes a single component.

Comparator	Task	n	C	Difference	95% interval
GLM-5.3	JDG	941	941	+3.05	[+2.85, +3.25]
GLM-5.3	PJFA	872	872	+1.50	[+0.42, +2.59]
GLM-5.3	QP	714	714	-0.45	[-1.66, +0.75]
GLM-5.3	JC	1064	1064	+0.28	[-1.41, +2.02]
GLM-5.3	TC	161	161	-0.93	[-6.21, +4.04]
GLM-5.3	PJF	1386	874	+1.27	[-0.32, +2.85]
GLM-5.3	TR	184	18	-3.41	[-4.46, +0.24]
GLM-5.3	SimT	811	274	+7.94	[-0.86, +12.03]
GLM-5.3	SimJ	856	618	-0.30	[-1.72, +1.16]
GLM-5.3	Dedup	870	762	-2.18	[-3.47, -0.95]
GLM-5.3	J2T	116	24	-0.42	[-2.90, +2.59]
GLM-5.3	T2J	361	1	-12.27	–
DeepSeek	JDG	941	941	-0.20	[-0.39, -0.01]
DeepSeek	PJFA	872	872	-10.17	[-12.04, -8.30]
DeepSeek	QP	714	714	-3.16	[-4.59, -1.79]
DeepSeek	JC	1064	1064	-7.33	[-9.45, -5.33]
DeepSeek	TC	161	161	-5.59	[-10.87, -0.62]
DeepSeek	PJF	1386	874	-3.09	[-5.20, -0.92]
DeepSeek	TR	184	18	-3.90	[-5.18, +0.44]
DeepSeek	SimT	811	274	+12.62	[+1.83, +17.18]
DeepSeek	SimJ	856	618	-1.43	[-3.36, +0.40]
DeepSeek	Dedup	870	762	-8.91	[-11.01, -6.88]
DeepSeek	J2T	116	24	-6.40	[-9.88, -3.01]
DeepSeek	T2J	361	1	-3.06	–
MiniMax	JDG	941	941	-1.96	[-2.17, -1.75]
MiniMax	PJFA	872	872	-27.98	[-29.76, -26.24]
MiniMax	QP	714	714	-5.03	[-6.52, -3.55]
MiniMax	JC	1063	1063	-3.07	[-5.06, -1.05]
MiniMax	TC	161	161	-6.83	[-12.42, -1.24]
MiniMax	PJF	1386	874	-2.76	[-5.21, -0.30]
MiniMax	TR	184	18	-3.00	[-4.09, +0.80]
MiniMax	SimT	811	274	+7.44	[-3.52, +11.70]
MiniMax	SimJ	855	617	-1.60	[-3.11, -0.13]
MiniMax	Dedup	870	762	-9.65	[-11.87, -7.47]
MiniMax	J2T	116	24	-4.86	[-8.96, -0.80]
MiniMax	T2J	361	1	-4.12	–

E.1 Comparisons among the Leading Overall Systems

Table 22 compares Claude-opus-5 with the next three systems in Overall order, retaining every task for each pair. Differences are Claude minus the named comparator. These task-specific comparisons use the same component-resampling procedure as above, with 10,000 replicates and seed 20260921. Intervals are exploratory, pointwise 95% percentile intervals; their interpretation concerns each task and its observed entity links. The complete collection of comparisons is reported together.

Table 22: Leading-system paired task differences. n: common records/groups; C: observed-entity components. A dash denotes one connected component.

Comparator	Task	n	C	Difference	95% interval
Kimi-K3	JDG	941	941	+1.19	[+0.98, +1.40]
Kimi-K3	PJFA	872	872	-1.43	[-2.41, -0.43]
Kimi-K3	QP	714	714	+0.36	[-0.66, +1.39]
Kimi-K3	JC	1064	1064	+0.88	[-0.69, +2.43]
Kimi-K3	TC	161	161	+0.00	[-3.42, +3.73]
Kimi-K3	PJF	1386	874	-0.97	[-2.69, +0.69]
Kimi-K3	TR	184	18	+1.53	[-1.58, +2.40]
Kimi-K3	SimT	811	274	-7.24	[-13.43, +7.90]
Kimi-K3	SimJ	856	618	+0.42	[-1.09, +1.86]
Kimi-K3	Dedup	870	762	+2.99	[+1.65, +4.36]
Kimi-K3	J2T	116	24	+0.94	[-1.44, +2.70]
Kimi-K3	T2J	361	1	+1.13	–
GPT-5.6-sol	JDG	941	941	+1.98	[+1.80, +2.16]
GPT-5.6-sol	PJFA	872	872	-1.04	[-2.13, +0.06]
GPT-5.6-sol	QP	714	714	+0.34	[-0.61, +1.30]
GPT-5.6-sol	JC	1064	1064	-0.91	[-2.47, +0.64]
GPT-5.6-sol	TC	161	161	-0.62	[-4.35, +3.11]
GPT-5.6-sol	PJF	1386	874	-1.30	[-3.19, +0.49]
GPT-5.6-sol	TR	184	18	+2.53	[+1.87, +3.16]
GPT-5.6-sol	SimT	811	274	-0.43	[-2.77, +4.14]
GPT-5.6-sol	SimJ	856	618	-0.33	[-1.92, +1.16]
GPT-5.6-sol	Dedup	870	762	+0.92	[-0.11, +1.95]
GPT-5.6-sol	J2T	116	24	+0.23	[-2.06, +2.87]
GPT-5.6-sol	T2J	361	1	+0.03	–
Gemini-3.5-flash	JDG	941	941	+1.19	[+1.03, +1.34]
Gemini-3.5-flash	PJFA	872	872	+2.86	[+1.66, +4.10]
Gemini-3.5-flash	QP	714	714	+1.71	[+0.49, +2.91]
Gemini-3.5-flash	JC	1064	1064	-0.52	[-1.96, +0.88]
Gemini-3.5-flash	TC	161	161	+0.00	[-2.80, +2.80]
Gemini-3.5-flash	PJF	1386	874	-2.20	[-3.85, -0.59]
Gemini-3.5-flash	TR	184	18	+2.83	[-0.50, +3.87]
Gemini-3.5-flash	SimT	811	274	-9.91	[-14.38, -0.80]
Gemini-3.5-flash	SimJ	856	618	+0.19	[-1.30, +1.74]
Gemini-3.5-flash	Dedup	870	762	+1.95	[+0.72, +3.24]
Gemini-3.5-flash	J2T	116	24	-0.54	[-3.71, +2.63]
Gemini-3.5-flash	T2J	361	1	+0.63	–

Overall retains the six-family equal-weight definition. A separate leave-one-family-out view in the analysis supplement illustrates the contribution of capability coverage: Claude-opus-5 ranks first for four of the six five-family means; omitting Retrieval puts Kimi-K3 first, while omitting Ranking puts GLM-5.3 first. These operational emphases support reporting the full capability profile.

Table 23: Leading five-family mean after omitting each capability family.

Omitted family	Leading system	Five-family mean
Generation	Claude-opus-5	78.428
Extraction	Claude-opus-5	71.030
Classification	Claude-opus-5	75.194
Retrieval	Kimi-K3	76.770
STS	Claude-opus-5	73.414
Ranking	GLM-5.3	72.430

F Source and Candidate Conditions

The analysis retains within-task source, occupation, reference-depth, query-style, construction-type, and candidate-condition slices wherever those metadata occur. Missing metadata are recorded through per-field coverage rather than inferred. Candidate-size bins are 1–8, 9–16, and 17 or more; positive-count bins are one versus multiple; the positive-rate split is 0.125, and available same-second-level occupation ratios are split at 0.5. These fixed descriptive bins are applied to the released data. Both labels must occur for reporting slice balanced accuracy or binary macro-F1; single-class slices report positive recall or negative rejection instead. All denominators are retained. Complete results, including all occupation strata, are in the machine-readable analysis supplement; Tables 24–25 give every multi-source task's source breakdown for all twelve systems.

The candidate census finds 104 retrieval pools of size 50, 47 of size 12, and 33 of size eight. The two ranking directions have one positive per group, whereas retrieval has between one and 28. Job-to-talent positive input positions 1–10 have counts 13, 16, 20, 10, 22, 13, 7, 4, 6, and 5; talent-to-job positions 1–8 have counts 52, 48, 37, 52, 53, 35, 34, and 50. These counts determine the input-order reference scores in Table 30. The strict permutation and top-ten protocols likewise determine how each candidate ordering is scored.

Source panels in Tables 24–25 retain every system and denominator. "Public" denotes original public-dataset records, "Public JD" public job descriptions, "Derived" public-data-derived, and "Synthetic" controlled synthetic records.

Talent classification makes the associated input conditions concrete: the 91 public records have median input length 219 characters and depth-two labels throughout; the 70 synthetic records have median length 979, with 64 depth-two and six depth-three labels. Across twelve systems, their score ranges are 38.46–44.51 and 85.00–99.29, respectively. These source profiles reflect jointly varying content and label conditions; the tables report every system's scores.

Table 24: Source scores for JDG, PJFA, TC, TR. Header counts give stratum sizes.

	JDG		**PJFA**		**TC**		**TR**	
System	Public (500)	Public JD (441)	Derived (500)	Synthetic (372)	Public (91)	Synthetic (70)	Public (104)	Derived (80)
Claude-opus-5	43.46	48.48	57.41	73.44	40.66	97.86	45.55	86.53
Kimi-K3	42.65	46.87	58.95	74.72	41.21	97.14	42.86	86.53
GPT-5.6-sol	41.14	46.89	57.18	76.17	40.66	99.29	43.08	83.92
Gemini-3.5-flash	42.05	47.54	53.61	71.85	39.56	99.29	41.15	85.74
Claude-opus-4.8	41.89	46.45	58.93	75.09	42.86	97.14	40.09	86.91
GLM-5.3	47.60	50.29	57.34	77.06	41.76	94.29	40.61	85.10
GPT-5.5	40.52	46.27	57.56	75.63	39.01	98.57	42.70	84.69
Qwen3.8-max	43.94	48.60	53.28	74.05	40.11	97.14	42.98	85.95
GLM-5.2	43.73	48.24	46.68	75.22	43.96	94.29	42.02	85.86
Kimi-2.6	42.34	47.00	46.01	71.75	44.51	91.43	41.20	85.53
DeepSeek v4-Pro	43.10	48.46	38.28	75.31	40.11	85.71	40.47	84.17
MiniMax-M3	41.92	46.03	27.91	47.49	38.46	85.00	41.10	85.41

Table 25: Source scores for SimT, SimJ, Dedup, J2T. Header counts give stratum sizes.

	SimT		SimJ		Dedup		J2T	
System	Derived (500)	Synthetic (311)	Derived (540)	Synthetic (316)	Derived (616)	Synthetic (254)	Derived (110)	Synthetic (6)
Claude-opus-5	65.85	53.89	74.09	94.64	99.51	95.67	73.65	74.87
Kimi-K3	58.33	72.78	71.92	98.15	97.89	89.32	72.42	79.36
GPT-5.6-sol	64.14	56.95	72.77	99.07	99.03	93.69	73.05	81.55
Gemini-3.5-flash	67.90	72.81	74.95	91.97	97.08	94.88	74.16	76.03
Claude-opus-4.8	59.44	65.49	73.75	96.67	98.05	85.29	73.59	74.87
GLM-5.3	65.81	70.05	72.53	96.77	97.73	92.52	73.37	72.06
GPT-5.5	63.72	55.25	72.34	98.13	98.54	90.50	70.93	79.36
Qwen3.8-max	53.74	57.11	74.41	89.71	95.77	81.63	72.78	69.24
GLM-5.2	67.88	69.57	75.03	94.12	96.92	92.89	70.34	58.73
Kimi-2.6	67.27	70.28	72.74	95.55	96.43	88.06	70.14	68.72
DeepSeek v4-Pro	69.09	75.30	73.98	90.29	96.26	72.32	67.00	73.21
MiniMax-M3	62.76	69.57	71.86	93.81	96.75	67.56	68.53	74.87

F.1 PJF Construction Strata and a Supervised Lexical Probe

PJF contains 900 behavior-derived pairs (478 satisfaction positives, 207 browsing-without-application negatives, and 215 application-without-satisfaction negatives) and 486 reviewed constructed pairs (162 positives and 324 negatives). Table 26 reports balanced accuracy from final labels and saved predictions within each stratum. Full-task balanced accuracy pools the class counts across strata before averaging positive and negative recall.

Table 26: PJF performance by construction stratum (percent). Behavior TPR/TNR use 478 positive/422 negative final labels.

System	Behavior BA	Constructed BA	Behavior TPR	Behavior TNR
Claude-opus-5	52.69	92.59	72.0	33.4
Kimi-K3	54.20	94.44	74.3	34.1
GPT-5.6-sol	56.64	89.66	71.3	41.9
Gemini-3.5-flash	54.43	94.44	81.4	27.5
Claude-opus-4.8	52.01	93.52	75.1	28.9
GLM-5.3	54.75	92.75	74.9	34.6
GPT-5.5	53.37	95.22	72.4	34.4
Qwen3.8-max	52.85	93.83	62.3	43.4
GLM-5.2	55.58	93.21	78.5	32.7
Kimi-2.6	55.27	93.83	65.3	45.3
DeepSeek v4-Pro	52.73	85.96	57.1	48.3
MiniMax-M3	55.40	83.64	49.0	61.8

The probe combines character 1–2-gram TF-IDF (sublinear term frequency, minimum document frequency 2) with logistic regression ($C = 1$, balanced class weights, at most 3,000 iterations).

Vocabulary and IDF are fitted within each training fold. Five-fold cross-validation groups by resume ID or job ID, with three deterministic group relabelings (seeds 0, 1, 2); each feature configuration uses the same folds within a grouping. Scores pool out-of-fold predictions per seed and then average seeds. A fixed probability threshold of 0.5 gives balanced accuracy. Seed standard deviations describe split variation, not confidence intervals. No metadata, IDs, or construction tags are features.

There are 764 unique resumes and 886 jobs; median text lengths are 154.5 and 363 characters. Job-only features exceed both-text features under either grouping, indicating predictive associations in job text. Resume-only features provide a complementary control. Each split holds out its named entity type; the probe characterizes associations within the released behavioral collection.

A separate unsupervised character-bigram cosine check fits IDF on all 1,800 unlabeled texts. Its AUC is 0.400 overall, 0.469 against browsing negatives, and 0.334 against application-without-

Table 27: Supervised PJF probe ($n = 900$): AUC from held-out probabilities; BA at a 0.5 threshold.

Held-out entity	Text features	AUC	AUC SD	BA (%)
Resume	Resume + job	0.661	0.002	61.01
Resume	Resume only	0.546	0.005	54.26
Resume	Job only	0.671	0.006	63.28
Post	Resume + job	0.651	0.005	60.30
Post	Resume only	0.539	0.008	52.62
Post	Job only	0.666	0.005	62.72

satisfaction negatives. The supervised and unsupervised controls together distinguish learnable behavioral associations from surface overlap. The two PJF strata therefore support evaluation against both interaction evidence and reviewed constructed judgments under a common task interface.

F.2 Relation Coverage, Selectivity, and Construction Difficulty

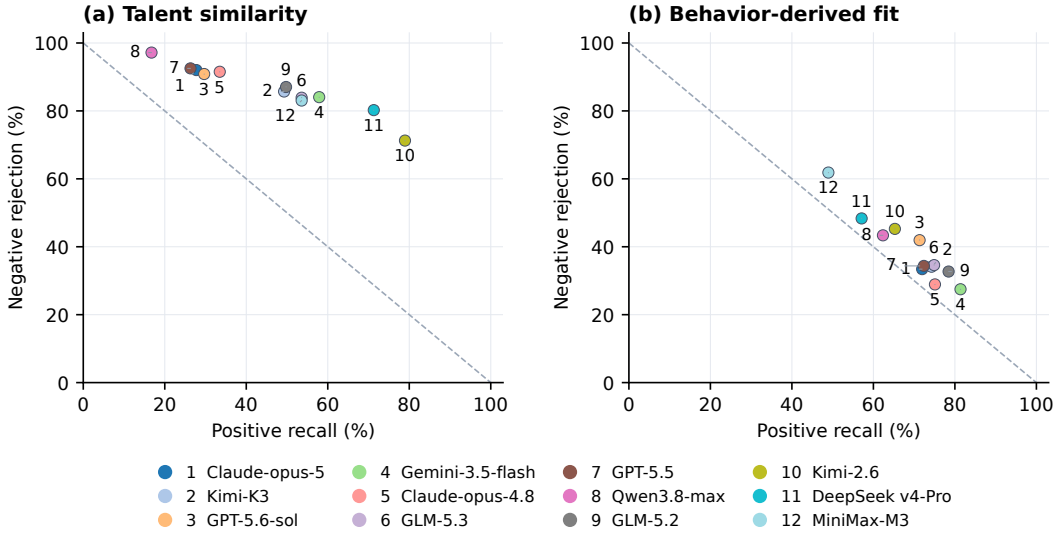

Figure 8: Positive recall and negative rejection for (a) SimT ($n = 811$) and (b) behavior-derived PJF ($n = 900$). Dashed lines mark 50% balanced accuracy; numbers identify systems.

SimT evaluates related-profile discovery using shared-job and occupation-based construction followed by label review. Positive recall ranges from 16.75% to 78.95%, while negative rejection ranges from 71.26% to 97.18%. Figure 8 shows that systems with greater positive recall also tend to accept more negative pairs. On SimJ, rejection of the 141 same-category pairs with final negative labels ranges from 4.96% to 22.70%, versus 99.50–100% on the 199 cross-category negatives. Thus the narrow SimJ range masks large differences in rejection between the two negative strata.

HR-reviewed labels determine scoring; construction tags record sampling intent. SimT's 167 shared-job positive seeds yield 105 positive and 62 negative final labels, which also define the class-conditional diagnostics (Table 28).

F.3 Construction-Model and Judge-Family Overlap

Table 29 compares the two resume generators with other systems across all five multi-source tasks containing controlled-synthetic records. Claude-opus-5 and GPT-5.6-sol rank twelfth and tenth on synthetic SimT, while GPT leads synthetic SimJ and ties for first on TC, and Claude leads Dedup. The generators therefore show task-dependent strengths, without a consistent advantage across synthetic strata. These are source-level comparisons; the metadata do not attribute each record to a generator.

Table 28: Score ranges across twelve systems by construction condition. Binary rows use final-label recall/rejection; QP uses span F1.

Task	Stratum and statistic	Total n	System range
SimT	Same-category negatives: TNR	167	58.08–98.20
SimT	Cross-category negatives: TNR	166	99.40–100.00
SimJ	Same-category final negatives: TNR	141	4.96–22.70
SimJ	Cross-category negatives: TNR	199	99.50–100.00
Dedup	Masked identity positives: TPR	250	93.60–100.00
Dedup	Identity snapshots: TPR	130	43.08–96.15
Dedup	Matched-attribute negatives: TNR	124	79.84–97.58
QP	single	325	79.82–87.95
QP	multi constraint	268	92.28–96.18
QP	noisy	121	77.27–88.55

Table 29: Controlled-synthetic scores of the two resume generators versus the other ten systems. Parentheses give ranks among all twelve systems, with ties.

Task	Claude-opus-5	GPT-5.6-sol	Others' mean	Others' range
PJFA	73.44 (9)	76.17 (2)	71.82	47.49–77.06
TC	97.86 (4)	99.29 (1)	94.00	85.00–99.29
SimT	53.89 (12)	56.95 (10)	67.82	55.25–75.30
SimJ	94.64 (7)	99.07 (1)	94.52	89.71–98.15
Dedup	95.67 (1)	93.69 (3)	85.50	67.56–94.88

The rubric judge is Qwen-3.6-Flash. The evaluated Qwen3.8-max has JDG/PJFA rubric means of 85.75/90.91, below GLM-5.3's 91.12/94.64. The evaluated judge-family model leads neither rubric.

F.4 Candidate-Task Baselines and Paired Gains

Three offline reference methods characterize candidate conditions and contextualize LLM scores using the fixed task labels and pools. *Input order* returns the supplied list unchanged. *Uniform permutation* reports the exact expected score over all candidate permutations, conditional on each pool's size and relevant-item count. *Character-bigram TF-IDF* ranks candidates by cosine similarity to the anchor, with lowercasing, collapsed whitespace, raw term frequency, smoothed IDF $\log((1+N)/(1+\mathrm{df}))+1$, and L2 normalization. IDF is fitted to the unique prompt-visible texts of each task, an unlabeled transductive setting; neither labels, reference answers, identifiers, nor source tags are features. Similarity ties retain input order. The method uses fixed settings, without score-based tuning, and always returns a complete permutation.

Text fields follow the evaluation runner: ranking anchors use up to 1,500 characters and candidates 400; retrieval uses job-title/category/requirement fields, a job description of up to 1,500 characters, and candidate experience of up to 300 characters. Labels enter only the final metric computation. TR averages Recall@3, Recall@10, and MRR@10; both ranking tasks use nDCG@5. Single-positive ranking measures placement of the relevant candidate, while the complete-permutation requirement measures list execution.

Table 30: Baseline and model scores within each source. The model is the overall task leader: Claude-opus-5 for TR/T2J and Gemini-3.5-flash for J2T. Δ: model minus TF-IDF.

Task	Source	n	Uniform	Input	TF-IDF	Model	Δ
TR	Public	104	16.68	19.65	37.90	45.55	+7.66
TR	Derived	80	53.34	54.45	72.22	86.53	+14.31
J2T	Derived	110	43.10	39.92	56.97	74.16	+17.19
J2T	Synthetic	6	44.75	33.33	13.63	76.03	+62.40
T2J	Synthetic	361	36.86	39.80	89.05	97.19	+8.14

Table 31: Paired gains over TF-IDF on all groups. W/T/L counts higher/equal/lower group scores at four decimals (0–1 scale). Intervals use observed-entity components; T2J has one component.

Task	System	W/T/L	Δ	95% interval
TR	Claude-opus-5	101/45/38	+10.55	[+7.71, +17.86]
J2T	Gemini-3.5-flash	63/39/14	+19.53	[+10.80, +28.60]
T2J	Claude-opus-5	75/276/10	+8.14	–

The gains use common complete group sets and the component-resampling procedure from Appendix E, with 10,000 replicates and seed 20260921. The model for each task is selected by its reported task score; these are exploratory comparisons of the observed task leaders. Both sources in TR and J2T show positive mean model gains. T2J's 75 improved, 276 tied, and ten lower-scoring groups show how model gains coexist with strong lexical performance. The six synthetic J2T groups are reported with their own denominator. All methods use the same supplied pools; source comparisons describe these input and label conditions.

G Operational Results and Correspondence Details

G.1 Population Statistics and Measurement Units

Figure 9 compares the two classification cohorts using the shared eighteen-category first level. The 1,064 external job records cover all eighteen categories, and the 161 talent records cover seventeen; the internal cohorts cover all eighteen. The external proportions are recomputed from released reference labels, and the internal proportions come from the construction team's aggregate plotting supplement. Both comparisons yield TV = 0.140 to three decimal places.

Table 32: External and private-reference populations. TV measures occupation distribution distance; ranking pool means are J2T/T2J.

Property	**External**	**Internal**	**Comparison**
Job-classification records	1,064	10,599	18/18 classes; TV = 0.140
Talent-classification records	161	7,716	17/18 classes; TV = 0.140
Job-text mean / median chars	526 / 460	528 / 423	Mean difference 0.4%
PJF city coverage (%)	92.1	100	7.9 pp difference
PJF education coverage (%)	99.9	100	0.1 pp difference
PJF experience coverage (%)	97.9	100	2.1 pp difference
Ranking pool means	7.3 / 8.0	8.1 / 6.0	Task-specific pools

Table 33: Measurement definitions for population comparisons. Sample counts are external/internal; ranking units are complete groups.

Statistic	Measurement	Sample count
Occupation proportions	First-level reference category, `cate_labels[0]`; TV is half the summed absolute proportion differences.	Job: 1,064/10,599 Talent: 161/7,716
Job-text length	Character count of `input.post_text` in job classification.	1,064/10,599
PJF field coverage	Same regex/rule detector applied to resume input text (external: `input.resume_info`) for city, education, and work-experience content.	1,386/880
Ranking pool size	Number of entries in `input.candidates`, summarized per group.	J2T: 116/123 T2J: 361/82

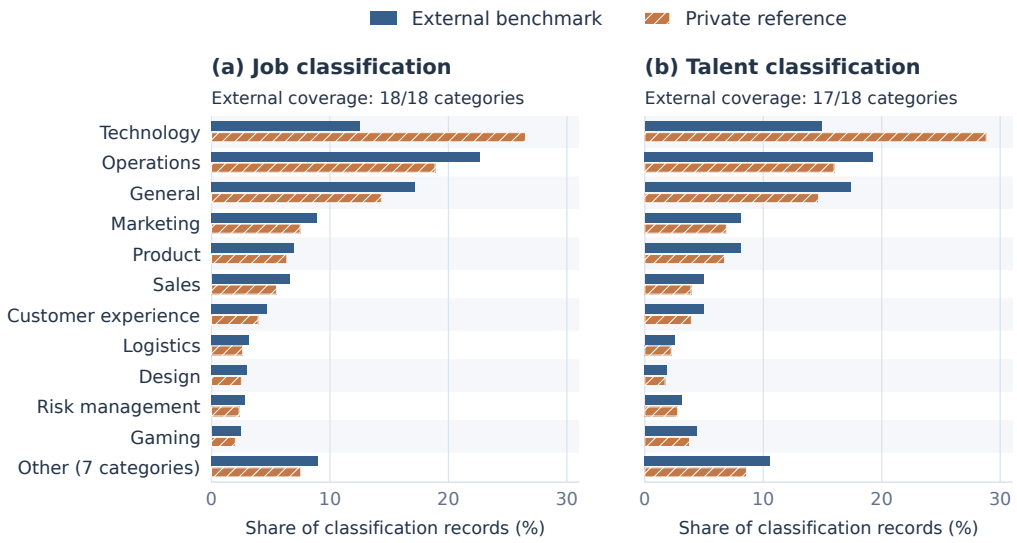

Figure 9: Occupation shares across eighteen categories. Categories below 3% in all four cohorts form Other. Tables 32 and 34 give cohort sizes and counts.

The external job-length mean and median are 525.56 and 460 characters, respectively; Table 32 reports rounded values. PJF field coverage measures detected information in resume text, and candidate groups remain the unit for pool comparisons.

Table 34: Complete first-level occupation counts and shares. Counts are external records; percentages use each cohort's denominator. † marks the categories grouped as Other in Figure 9: each is below 3% in all four cohorts.

	Job classification			Talent classification		
Occupation	**Ext. n**	**Ext. (%)**	**Int. (%)**	**Ext. n**	**Ext. (%)**	**Int. (%)**
Technology	133	12.50	26.50	24	14.91	28.82
Operations	241	22.65	18.87	31	19.25	15.98
General	183	17.20	14.33	28	17.39	14.67
Marketing	95	8.93	7.48	13	8.07	6.88
Product	74	6.95	6.36	13	8.07	6.74
Sales	70	6.58	5.49	8	4.97	3.90
Customer experience	50	4.70	3.93	8	4.97	3.93
Logistics	34	3.20	2.63	4	2.48	2.23
Design	32	3.01	2.52	3	1.86	1.71
Risk management	30	2.82	2.33	5	3.11	2.77
Gaming	26	2.44	2.01	7	4.35	3.77
Data†	25	2.35	1.93	4	2.48	1.92
Supply chain†	23	2.16	1.83	3	1.86	1.65
Finance†	14	1.32	1.11	3	1.86	1.37
Project management†	13	1.22	1.05	4	2.48	1.87
Content†	12	1.13	0.91	2	1.24	1.21
Advertising†	5	0.47	0.43	1	0.62	0.51
Map data†	4	0.38	0.29	0	0.00	0.08

G.2 Scores and Correspondence Statistics

Spearman correlations use average ranks within the five shared systems on each side. Direction agreement counts a pair only when both displayed score differences have the same nonzero sign.

Table 35: Private-reference capability scores; bold marks column maxima. The task-specific internal suite is excluded from shared-model correspondence.

System	Overall	Gen	Ext	Cls	Ret	STS	Rank
Kimi-K3	**70.48**	51.34	92.51	70.05	76.32	**67.56**	**65.12**
Qwen3.8-max	70.18	51.55	**92.95**	**70.51**	77.36	64.49	64.21
GLM-5.2	70.14	**52.51**	92.71	68.53	**77.79**	66.19	63.13
Kimi-2.6	69.68	51.00	90.54	69.83	75.77	65.97	64.99
DeepSeek v4-Pro	68.55	47.75	91.94	67.51	72.88	67.34	63.89
Internal solution	63.79	47.53	84.24	69.28	69.86	48.14	63.70

Table 36: Task scores on the private operational reference.

	Gen		Ext	Cls			Ret	STS			Rank	
System	JDG	PJFA	QP	JC	TC	PJF	TR	SimT	SimJ	Dedup	J2T	T2J
Kimi-K3	48.96	53.73	92.51	72.83	72.98	64.34	76.32	49.82	63.55	89.31	67.05	63.20
Qwen3.8-max	49.80	53.30	92.95	72.98	72.80	65.76	77.36	39.82	64.96	88.68	65.88	62.54
GLM-5.2	52.09	52.93	92.71	73.36	69.61	62.63	77.79	43.32	65.86	89.40	65.21	61.06
Kimi-2.6	51.47	50.52	90.54	74.27	70.36	64.87	75.77	59.06	57.63	81.23	67.53	62.45
DeepSeek v4-Pro	50.99	44.50	91.94	72.79	68.48	61.26	72.88	51.91	67.25	82.85	64.43	63.34
Internal solution	45.70	49.37	84.24	63.42	70.12	74.30	69.86	47.27	40.81	56.34	61.43	65.98

All ten pairs remain in each view's denominator, including a tie on either side. Table 37 retains all capability views and Table 38 all tasks.

For Overall, under an exploratory exchangeable-order permutation null, exact positive concordance has one-sided $p = 1/120 = 0.0083$ (two-sided $p = 0.0167$). This describes the ordering under that null, not uncertainty from sampling records or repeated model calls.

Table 37: Overall and category correspondence for the five shared systems.

View	Spearman ρ	Same / all	Opposite	Tied
Overall	1.00	10/10	0	0
Generation	0.60	7/10	3	0
Extraction	0.70	8/10	2	0
Classification	0.40	7/10	3	0
Retrieval	0.60	7/10	3	0
STS	0.20	6/10	4	0
Ranking	0.90	9/10	1	0

Table 38: Task correspondence. Each task contributes ten comparisons.

Task	Spearman ρ	Same	Opposite	Tied
JDG	0.00	5	5	0
PJFA	1.00	10	0	0
QP	0.70	8	2	0
JC	0.10	5	5	0
TC	0.36	6	3	1
PJF	0.00	4	6	0
TR	0.60	7	3	0
SimT	0.80	8	2	0
SimJ	-0.36	3	6	1
Dedup	0.70	8	2	0
J2T	0.70	8	2	0
T2J	0.30	6	4	0

The pooled task comparison has 78 same-direction, 40 opposite-direction, and two tied pairs at displayed precision, giving 78/120 = 65.0% with ties retained in the denominator. These pairs are dependent, and pooling weights tasks equally, whereas Overall weights capabilities equally. The displayed ties are Qwen3.8-max/Kimi-2.6 on talent classification and Kimi-K3/Kimi-2.6 on job similarity. For standardized task scores, each task is centered and scaled within the five models separately on each side; pooling the twelve standardized pairs of vectors gives $r = 0.494$. Overall pairwise score differences give $r = 0.713$. These are descriptive correspondence statistics. Their interpretation concerns patterns and ordering rather than equality of numerical score gaps. The internal Overall difference between Qwen3.8-max and GLM-5.2 is 0.04 points.

G.3 Public–Private Agreement under Capability Omission

We assess how the five shared models' public–private ordering changes when one capability family is removed from both collections. For each of the six omission conditions, we average the other five reported capability scores with equal weights and compute Spearman correlation between the two model rankings. Both sides use the two-decimal capability scores in `data/text_results.json`; we do not substitute higher-precision public recomputations or infer scores from the reported Overall. Scores are converted to integer hundredths before summation, and equal sums receive average ranks.

Agreement is $\rho = 1.00$ after omitting generation, retrieval, or ranking, and $\rho = 0.90$ after omitting extraction, classification, or STS (Table 39). Table 40 gives every model's mean and rank in all six conditions. Together, these results show closely aligned aggregate orderings across all six omissions. This descriptive analysis assesses aggregation sensitivity for the five shared models at the reported precision; task validity and sampling uncertainty require separate evidence. It compares both collections, complementing the capability-omission analysis of the public twelve-model leaderboard. The official Overall scores and rankings are unchanged. The calculation and its input hash are released in `scripts/operational_sensitivity.py` and `data/operational_sensitivity.json`.

Table 39: Public–private ranking agreement among five shared models after omitting the same capability family from both collections. Each remaining family has equal weight.

Omitted capability	Spearman ρ
Generation	1.00
Extraction	0.90
Classification	0.90
Retrieval	1.00
STS	0.90
Ranking	1.00

Table 40: Five-capability means and model ranks underlying Table 39. Means average two-decimal reported inputs and are displayed to three decimals to preserve differences introduced by averaging. Rank 1 is highest; equal sums receive average ranks.

		Public		Private	
Omitted	Model	Mean	Rank	Mean	Rank
Generation	Kimi-K3	78.106	1	74.312	1
	Qwen3.8-max	76.632	2	73.904	2
	GLM-5.2	76.524	3	73.670	3
	Kimi-2.6	76.424	4	73.420	4
	DeepSeek v4-Pro	75.154	5	72.712	5

Continued on next page

Table 40 continued

Omitted	Model	Public mean	Rank	Private mean	Rank
Extraction	Kimi-K3	70.806	1	66.078	1
	Qwen3.8-max	69.302	2	65.624	3
	GLM-5.2	69.096	3	65.630	2
	Kimi-2.6	68.990	4	65.512	4
	DeepSeek v4-Pro	67.350	5	63.874	5
Classification	Kimi-K3	74.892	1	70.570	1
	Qwen3.8-max	73.192	2	70.112	3
	GLM-5.2	72.682	3	70.466	2
	Kimi-2.6	72.586	4	69.654	4
	DeepSeek v4-Pro	71.950	5	68.760	5
Retrieval	Kimi-K3	76.770	1	69.316	1
	Qwen3.8-max	75.128	2	68.742	2
	GLM-5.2	74.778	3	68.614	3
	Kimi-2.6	74.482	4	68.466	4
	DeepSeek v4-Pro	73.228	5	67.686	5
STS	Kimi-K3	72.862	1	71.068	2
	Qwen3.8-max	72.264	2	71.316	1
	GLM-5.2	70.568	3	70.934	3
	Kimi-2.6	70.258	4	70.426	4
	DeepSeek v4-Pro	68.950	5	68.794	5
Ranking	Kimi-K3	72.254	1	71.556	1
	Qwen3.8-max	70.782	3	71.372	3
	GLM-5.2	71.322	2	71.546	2
	Kimi-2.6	70.140	4	70.622	4
	DeepSeek v4-Pro	68.978	5	69.484	5

H Positioning Against Recruitment Resources

Table 41: Recruitment resources by language, outputs, candidate conditions, and label evidence. WorkRB coverage includes the accompanying toolkit (TechWolf Research Team, 2026).

Resource	Language / focus	Outputs / candidate conditions	Label evidence
ResumeBench (Ling et al., 2025)	Five languages, including Chinese	Resume PDF to schema-defined JSON	Curated templates; GPT-4o-generated JSON aligned with PDF layouts
JobResQA (Carrino et al., 2026)	Five languages, including Chinese	Short answers and explanations over synthetic resume–JD pairs	Linguist-authored QA; human-assisted translation
PJB (Wang et al., 2026)	Chinese retrieval	297 JD queries; fixed collection of 197,674 resumes	Doubao screening and Kimi relevance assessment; human checks on about 20% of queries
WorkRB (De Lange et al., 2026)	Multilingual work domain	Ranking and job-to-skill multilabel classification; ontology- and corpus-based candidate spaces	Task-specific datasets and occupational ontologies; extensible framework
EHR-Bench	Chinese recruitment	12 operations; generated text, typed spans, category paths, binary relations, retrieval, complete permutations	Source-aware label verification; assisted checks and expert judgments for constructed relations

EHR-Bench makes twelve enterprise-informed Chinese recruitment operations available for public LLM evaluation. Task-specific inputs, expert-reviewed references, and output requirements support comparison of decisions, explanations, relations, and candidate selection. Researchers can use these shared protocols to test whether model and prompting improvements carry across tasks and sources. Supplied pools define the candidate conditions for shortlist selection.

ResumeAtlas (Heakl et al., 2024) supplies resume classification data; SkillSpan (Zhang et al., 2022) and Chinese-SkillSpan (Li et al., 2026) provide English skill spans and Chinese ESCO-aligned competency spans, respectively. ConFit v2 (Yu et al., 2025) studies hard-negative retrieval training.

The WorkRB paper reports thirteen ranking tasks in seven groups. Its pinned toolkit (TechWolf Research Team, 2026) also supports ESCO job-to-skill multilabel classification and graded-relevance variants; our comparison includes these interfaces.

Public/private comparisons also occur outside recruitment. UNED-ACCESS (Sánchez Salido et al., 2025) compares previously unreleased entrance-exam questions with a subject-matched subset of MMLU. EHR-Bench uses its enterprise reference both to inform resource construction and to contextualize the observed scores on a fixed shared-model set.

I Reproducibility and Artifact Record

The corpus, prompts, and scorer are identified by commit `496964d`; saved per-sample predictions and component statistics by `e8905f2`; and population summary definitions by `b85bb3a`. The manifest records complete identifiers and SHA-256 digests. All 144 scores for the twelve evaluated systems are recomputed from the saved per-sample artifacts with the released scoring formulas; every model–task difference is at most 0.011 on the 0–100 scale. This check includes binary labels, generation-component weights, strict-span aggregation, and candidate metrics. It reuses saved rubric scores and does not independently rerun the judge. Exported fields vary by task; generation exports provide component scores rather than the complete raw response texts. Business-specific review procedures and item-level expert records are not part of the public release.

From the manuscript directory, with the result archive available at `/path/to/results`, the score check and table generation require:

```
python -m pip install -r scripts/requirements.txt
python scripts/validate_v21.py --release-root \
/path/to/results
python scripts/generate_results.py
```

The validator checks saved scorer and output artifacts; generated tables retain six-family equal-weight Overall. The README gives commands for figures, paired comparisons, candidate baselines, and the PJF probe. Auxiliary results record seeds, features, entity groupings, software versions, and input hashes; additional PJF dependencies are listed separately. Private-reference reproducibility uses aggregate tables; raw records remain in the operational environment.