Interpretable Job Recommendation from Multiple Sources of Occupational Evidence: Early Results

Eric Keiji Tokuda[1] , Alexandre Cláudio Botazzo Delbem [1] , Anderson da Silva Soares[2]

[1]Institute of Mathematics and Computer Science, University of São Paulo
São Carlos – Brazil

tokudaek@alumni.usp.br,acbd@icmc.usp.br

[2]Institute of Informatics, Federal University of Goiás
Goiânia – Brazil

anderson@inf.ufg.br

Abstract. Skill-based occupation recommenders should show how evidence from different data sources leads to ranked occupations, especially for job seekers with limited literacy or access to guidance. This paper presents a deterministic pipeline for traceable job recommendation that integrates structured occupational descriptors with annotated job-text spans and converts them into yes/no questionnaire evidence. The pipeline builds O*NET occupation profiles, audits SkillSpan skill and knowledge spans, maps them to occupational descriptors, generates traceable questions, links questions back to occupations, and tests whether simulated answers recover the source occupation. The question structure works well under controlled reconstruction: compact task-based questions recover occupations almost perfectly. The multi-source text-integration branch is weaker: lexical extraction predicts 9,375 spans for 2,264 gold spans, and conservative normalization accepts only 662 mappings. The pipeline is therefore a job-recommendation prototype and audit layer for inclusive occupational evidence, showing where evidence remains usable and where extraction or mapping breaks down.

1. Introduction

Recommendation systems increasingly shape how people find jobs, training options, and career-transition pathways [Lawler and Ledford 1992, Ko et al. 2022]. In these settings, rankings often depend on evidence about skills, tasks, tools, and prior work experience, yet the path from that evidence to a recommended occupation is hard to inspect. Dense learned representations, proprietary scoring rules, and multi-stage matching pipelines can rank an occupation highly without showing which pieces of user evidence were preserved, changed, or discarded. This opacity matters in employment-related contexts because recommendations may affect training choices, perceived opportunity, and the kinds of work a person considers reachable.

Skill-based recommendation also faces a multi-source evidence-integration problem. Occupational reference systems describe work through tasks, activities, abilities, skills, and technologies, but job postings and resumes express similar ideas through uncontrolled natural language. A recommender must extract evidence from

text, map it to occupational descriptors, and convert the mapped evidence into a form that supports ranking. Each step can introduce noise: a span extractor may overgenerate, a normalizer may map a specific tool to a vague descriptor, and a ranking model may hide these losses behind a plausible final score [Zhang et al. 2020]. Prior work has studied skill extraction, keyword-based classification of transversal skills, O*NET-based person-occupation modeling, and LLM-based job recommendation interfaces [Gavrilescu et al. 2025, Converse et al. 2004, Ghosh et al. 2023]. This paper focuses on whether occupational and textual evidence can be transformed into inspectable questionnaire signals without losing traceability.

We present early results from a deterministic pipeline that converts occupational descriptors and annotated job-text spans into inspectable artifacts: occupation profiles, extracted spans, normalized mappings, generated questions, occupation-question links, and ranked outputs. Instead of collapsing evidence into a latent vector, descriptors become short yes/no questions whose positive answers provide ranking evidence and a visible chain back to source descriptors. This design is especially relevant for job seekers who may not describe their experience using formal occupational terminology, because the system can ask about concrete tasks rather than requiring polished resumes or abstract skill labels. The study contributes a traceable recommendation pipeline, tests whether questionnaire evidence preserves enough occupational signal for reconstruction, and identifies where evidence is lost when uncontrolled job text is extracted and normalized.

2. Method

We conducted an empirical artifact study of a deterministic pipeline that converts occupational descriptors and annotated job-text spans into auditable recommendation evidence. Each stage emits an intermediate artifact that can be inspected, traced to source content, and rerun from pinned inputs. The pipeline uses three local artifacts: the O*NET text database, which supplies the occupational reference structure (tasks, detailed work activities, skills, abilities, knowledge items, and technology examples); the O*NET Easy Read workbook, which supports lower-barrier occupation framing; and the SkillSpan source archive, which provides annotated job-text spans for extraction and normalization diagnostics.

The operational sequence is fixed: dataset collection, occupation profile construction, scope definition, SkillSpan audit and preparation, span extraction, normalization, question generation, reconstruction, posting diagnostics, and question-quality diagnostics. Occupation profiles are built by grouping O*NET rows by O*NET-SOC Code, retaining occupation title, job zone, task statements, detailed work activities, scored skills, knowledge items, abilities, and technology examples. A lower-barrier branch retains occupations with job_zone <= 2.

The SkillSpan corpus is audited to verify token and tag alignment, then rewritten into a uniform internal format. Span extraction uses a deterministic lexical baseline learned from the SkillSpan training split: it extracts contiguous gold spans, lowercases them into layer-specific lexicons, applies frequency and length filters, and enumerates candidates in the test split through normalized surface-form overlap. The purpose is to create a transparent lower bound, not to compete with semantic

extraction models. Extraction is evaluated using exact, relaxed, skill-span, and knowledge-span F1, along with error-category counts.

Normalization maps extracted job-text spans to the occupation-derived descriptor inventory. Candidate retrieval uses token overlap, with character trigram overlap as a fallback. Candidate scoring combines exact match, token Jaccard overlap, fuzzy string similarity, and a descriptor-type prior. A mapping is accepted only when the top candidate clears a conservative confidence threshold, favoring traceable mappings over broad coverage.

Question generation turns descriptor clusters into short yes/no prompts. Task phrases become "Have you ever …?" questions; skills, knowledge, and technologies use "Do you have experience with …?" prompts. Each question keeps its source cluster and descriptor provenance. Validation flags check sentence length, abstract wording, missing verbs, and duplicate wording. Occupation-question links store the occupation code, cluster identifier, source type, and an inverse-document-frequency-like weight that reduces the influence of generic clusters.

Reconstruction tests whether simulated positive answers can recover the occupation that generated them. Under the evaluator contract, a question is answered positively only when the target occupation is linked to that cluster. Ranking uses a cosine-like overlap score with weighted cluster matches and a coverage bonus. The evaluation compares random answers, random permutation ranking, a descriptor-identity oracle, a compact canonical task-question baseline, and a dense hybrid baseline combining tasks, detailed work activities, skills, knowledge, abilities, and technologies.

To make the artifact auditable across reruns, dataset files are stored locally with SHA-256 hashes, collection is checkpointed, and randomness is constrained by a fixed project seed (0). Reconstruction outputs also record the baseline name, question-bank hash, occupation-link hash, answer-vector hash, occupation-vector hash, and evaluator contract version. These controls make the reported rankings traceable not only to source descriptors but also to the exact intermediate artifacts used to generate them.

3. Results

The results are organized around two questions: whether the integrated evidence representation supports controlled reconstruction, and where evidence is lost when moving from job text into occupational descriptors.

The pipeline constructed profiles for 1,016 occupations; the lower-barrier scope rule, job_zone <= 2, retained 331 occupations. In the canonical task branch, 923 occupations received at least one question link, and all 331 lower-barrier occupations received links. The hybrid branch produced many more questions and links, improving coverage but making the evidence base much heavier.

The deterministic lexical extraction baseline performed weakly on the SkillSpan test set. It predicted 9,375 spans against 2,264 gold spans, showing strong overgeneration. Exact span F1 reached only 0.13, and relaxed span F1 reached 0.23. Skill spans were harder than knowledge spans, with F1 scores of 0.07 and 0.18, re-

spectively. The dominant error category was spurious_prediction, with 7,891 cases. Surface-form matching can find text that looks skill-related, but it does not reliably separate valid skill evidence from nearby occupational language.

Table 1. Lexical extraction performance and main error counts on the SkillSpan test set.

Measure	Value
Exact span F1	0.132
Relaxed span F1	0.231
Skill span F1	0.070
Knowledge span F1	0.179
Predicted spans	9,375
Gold spans	2,264
spurious_prediction	7,891
boundary_mismatch	465
missed_skill_span	425
missed_knowledge_span	328

Normalization showed the same bottleneck from a different angle. Candidate generation succeeded for almost all extracted spans, with a rate of 0.99, but only 662 of 9,375 spans were accepted as mappings, roughly 7% coverage. This indicates that normalization failure is not primarily a retrieval failure: the system can usually produce candidate descriptors, but the accepted evidence lies in a narrow and modest-confidence upper tail, with median confidence 0.431 and p90 confidence only 0.552. Knowledge spans mapped at about three times the rate of skill spans, 0.098 versus 0.031. Accepted mappings concentrated in descriptor types with short lexical anchors: technology (246, 37%) and knowledge (198, 30%) accounted for 444 of the 662 accepted mappings, while detailed work activities (104), skills (62), and tasks (52) contributed only 218. Too few mappings survive the confidence threshold to support dependable recommendation from raw job text.

The generated question bank performed well as a structured representation. The canonical task questions had a mean length below nine tokens, passed the yes/no answerability check, and preserved provenance for all generated items. Error flags for missing verbs, duplicate wording, and unfamiliar technology terms were near zero. This shows that the deterministic generation rules can create short, answerable, traceable prompts from occupational descriptors.

Reconstruction was the strongest result. Random baselines performed poorly, as expected, with MRR below 0.02 across both evaluated scopes. The descriptor-identity oracle achieved perfect reconstruction, confirming that the ranking procedure behaves correctly when evidence is explicitly linked. The compact canonical_task_text baseline achieved near-perfect reconstruction across all profiles, with MRR 0.995, Recall@5 1.000, and mean rank 1.01; in the lower-barrier subset, it reached perfect reconstruction. Figure 1 shows the rank distribution across baselines. The denser task_dwa_skill_hybrid baseline also reached perfect reconstruction, but it required far more positive evidence: 130.39 matched positives on average across

all profiles, compared with 5.00 for the canonical task baseline. The comparison also has a practical interpretation for questionnaire design. Perfect reconstruction is not enough if it requires a dense evidence layer that would be difficult to expose to users. The hybrid branch has 13,212 questions and 206,296 occupation-question links, versus 4,375 questions and 4,615 links for the canonical branch. The compact task branch is therefore more aligned with a conversational recommender because it recovers the target occupation with far fewer positive signals and a much smaller question-link structure.

Table 2. Compact reconstruction results across main baselines.

Baseline	Subset	MRR	R@5	Mean rank	Mean matched
random_answers	all profiles	0.008	0.005	462.00	3.75
random_answers	in scope	0.019	0.015	166.00	3.79
descriptor_identity_oracle	all profiles	1.000	1.000	1.00	5.00
descriptor_identity_oracle	in scope	1.000	1.000	1.00	5.00
canonical_task_text	all profiles	0.995	1.000	1.01	5.00
canonical_task_text	in scope	1.000	1.000	1.00	5.00
task_dwa_skill_hybrid	all profiles	1.000	1.000	1.00	130.39
task_dwa_skill_hybrid	in scope	1.000	1.000	1.00	109.60

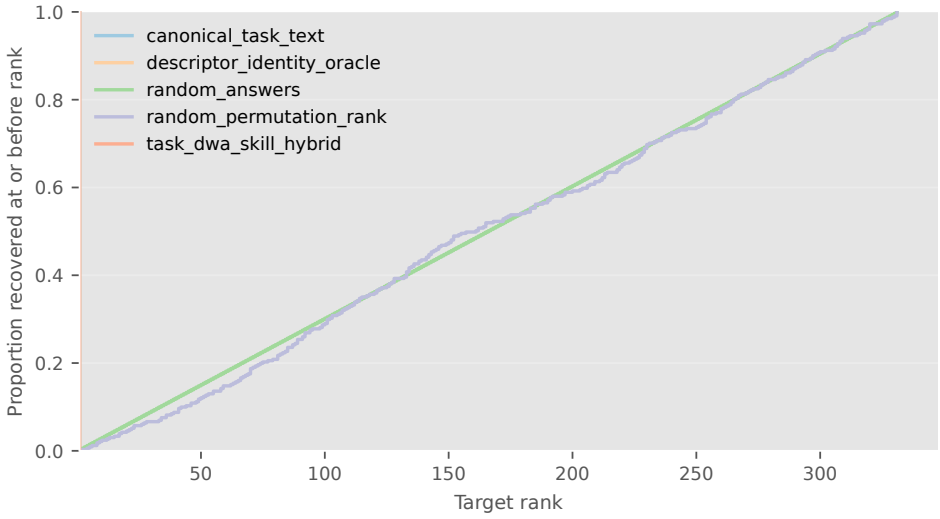

Figure 1. ECDF of target occupation rank by reconstruction baseline. Curves closer to the upper-left indicate faster and more consistent occupation recovery.

4. Conclusion

This paper presented a deterministic pipeline that integrates occupational descriptors and annotated job-text spans into traceable questionnaire evidence for skill-based occupation recommendation. Reconstruction experiments show that compact task-based questions preserve enough occupational signal to recover target occupations under controlled simulation, supporting interfaces that ask about concrete tasks rather than abstract credentials. Extraction and normalization show a different limit: lexical span extraction overgenerates, and conservative mapping accepts

too little evidence from raw job text. The pipeline's main value is diagnostic: it exposes whether failure comes from extraction, descriptor mapping, question generation, or ranking, rather than hiding losses inside a final score. Future work should improve extraction and mapping while keeping provenance and rejection reasons inspectable, so job seekers with limited literacy receive recommendations with visible evidence.

5. Acknowledgements

We acknowledge the support from the São Paulo Research Foundation (FAPESP) (#2025/22069-0, #2024/08485-8, #406774/2022-6, #2025/20876-5); the National Council for Scientific and Technological Development (CNPq) (#385292/2025-2 and #406774/2022-6); the Coordination for the Improvement of Higher Education Personnel (CAPES); the Center for Artificial Intelligence in Health Management (ciaGsaude) (FAPESP grant #2024/08485-8), the FAEPA and FFM foundations, and the Center of Excellence in Artificial Intelligence in Health Management CEPID-CeMEAI/ICMC-USP (CEPID, FAPESP grant #2013/07375-0); the Center of Excellence in Artificial Intelligence (CEIA/UFG); the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR) grant number 408490/2024-1.

References

Converse, P. D., Oswald, F. L., Gillespie, M. A., Field, K. A., and Bizot, E. B. (2004). Matching individuals to occupations using abilities and the o* net: Issues and an application in career guidance. Personnel Psychology, 57(2):451–487.

Gavrilescu, M., Leon, F., and Minea, A.-A. (2025). Techniques for transversal skill classification and relevant keyword extraction from job advertisements. Information, 16(3):167.

Ghosh, S. et al. (2023). JobRecoGPT: LLM-based job recommendation interfaces. arXiv preprint.

Ko, H., Lee, S., Park, Y., and Choi, A. (2022). A survey of recommendation systems: recommendation models, techniques, and application fields. Electronics, 11(1):141.

Lawler, E. and Ledford, G. (1992). A skill-based approach to human resource management. European Management Journal, 10(4):383–391.

Zhang, Y. et al. (2020). Explainable recommendation: a survey and new perspectives. arXiv preprint.