A Hybrid Lexicon-LLM Approach to Skill Extraction from Job Advertisements

Natalija Mečkaroska
natalijamechkaroska@gmail.com
University of Ljubljana
Ljubljana, Slovenia

Georgi Trajkov
geotrajkov0@gmail.com
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Abstract

Converting the free-text requirements of job advertisements into structured skill data is a bottleneck for labour-market applications. Two approaches are available without annotated training data: matching against a controlled vocabulary, or prompting a large language model (LLM). Which is preferable has not been established on human-annotated data, since existing comparisons score against benchmarks derived from the taxonomy itself. We compare a lexicon built from the 13,939 skills of the European ESCO taxonomy against three open-weight LLMs, on a manually annotated corpus of 60 advertisements containing 744 skill annotations, scored by concept-level precision, recall and F1 under exact and relaxed matching with McNemar's test. The lexicon is the most precise system (P = .519) while every LLM recalls more (R = .43 to .52), and pairing the lexicon with the best-calibrated LLM outperforms any component alone (F1 = .463). The three LLMs differ sharply in output volume, predicting 13.2, 25.5 and 28.0 skills per advertisement against 12.4 in the reference annotations, so over-generation is a property of a particular LLM rather than of generative extraction. LLMs agree with each other far more than with the lexicon (Jaccard .66–.69 against .21–.22), so the hybrid gain comes from combining paradigms rather than LLMs. The annotated corpus and code are publicly released.

Keywords

skill extraction, ESCO, large language models, information extraction, labour market analysis

1 Introduction

Online job advertisements are among the richest signals about labour demand, but they are unstructured text. An advertisement asking for "solid experience with containerised deployments" states a skill that the European Skills, Competences, Qualifications and Occupations classification (ESCO) [6], the standard European reference vocabulary and the target taxonomy of most recent work [2, 3, 14], records under a different label. Any system reasoning over skills depends on an upstream extraction step whose quality bounds everything downstream. This study arose from one such system: a skill-based job recommender for students, which matches a candidate's skills against those an advertisement requires. It consumes sets of skills rather than spans, which is why we annotate at the level of skill concepts.

Several approaches have been applied, including supervised span labelling [13], retrieval against a taxonomy [3, 14] and generative extraction [2]. We focus on the two available without annotated training data. A lexicon matches surface forms from a controlled vocabulary: transparent, free and auditable, but able to find only what the vocabulary contains. An LLM extractor prompts a large language model (LLM) to report the skills an advertisement states: robust to unanticipated wording, but costly at scale, non-deterministic and prone to reporting implied rather than stated skills. The two fail for different reasons, so if those failures are independent, combining them should recover more than either alone. We test this directly.

Our contributions are: (i) a controlled comparison of lexicon-based, LLM-based and hybrid extraction on a common corpus with significance testing; (ii) a manually annotated corpus of 60 advertisements with 744 concept-level skill annotations, released publicly; and (iii) an analysis of where lexicon false positives originate.

2 Related work

Annotated resources for skill extraction are scarce and sit below the document. Zhang et al. [13] introduced SkillSpan, a span-level dataset without taxonomy links; Decorte et al. [3] later labelled those spans with ESCO concepts and released a sentence-level subset. Senger et al. [12] survey the field and find annotation the binding constraint throughout. Our corpus is smaller but records the ESCO concepts an entire advertisement states, the unit a recommender consumes, and we are not aware of a comparable resource.

A second line treats extraction as taxonomy linking. Decorte et al. [3] show that drawing negatives from related ESCO skills yields the largest gains in distantly supervised extraction, and Zhang et al. [14] formulate the task as entity linking. More recently the taxonomy has been approached generatively: Clavié and Soulié [2] build a zero-shot ESCO matcher without human annotations, and Decorte et al. [4] use LLMs to generate training data for a classifier. These establish that LLM-driven pipelines are strong here, but not how a plain zero-shot prompt compares against the simplest deployable alternative, since both evaluate against benchmarks derived from ESCO itself rather than independent human judgement.

Combining a curated lexicon with a statistical component is a recurring pattern where annotated data is scarce [7], and the same comparison has been made in other extraction domains. Brank et al. [1] compare an expert-curated keyword index against an LLM for detecting olfactory references in Slovenian heritage texts, using the model family we adopt, and report that the keyword method returns higher frequencies at the cost of noise while the LLM is sparser but more context-sensitive, the two agreeing on about a quarter of mentions. Their corpus carries no human annotations, so where the methods disagree neither can be shown correct. Supplying that missing reference point is what this paper contributes.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.84

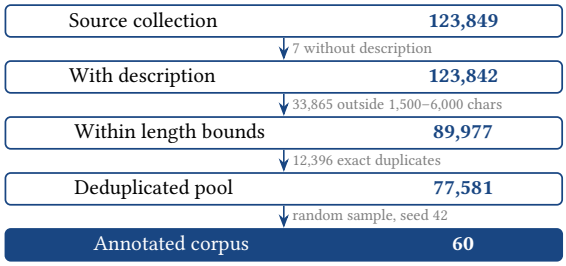


Figure 1: Construction of the evaluation corpus.

3 Data and annotation

3.1 Source corpus

We draw advertisements from a publicly available collection of 123,849 LinkedIn job postings published in 2023–2024 [8]. The collection is predominantly United States-based: 58% of the annotated advertisements contain US-specific markers. ESCO is used as a normalisation vocabulary rather than a jurisdictional standard, supplying a stable label space against which all systems are scored identically. Descriptions contain the complete advertisement text, median length 3,435 characters; we use a static collection rather than a live aggregator API because the two APIs we tested return only truncated descriptions.

3.2 Inclusion and exclusion criteria

No language filter was applied; the sampled advertisements are all English, which reflects the composition of the source collection rather than a selection decision. Four filters were applied, in order. (1) Postings with a missing description were discarded, leaving 123,842. (2) Descriptions were restricted to between 1,500 and 6,000 characters. (3) Exact-duplicate descriptions were removed, retaining the first occurrence. (4) A random sample of 60 postings was drawn from the remaining pool with a fixed seed. Of the 31 fields we retain five (identifier, title, company, location, description). Descriptions were then normalised by unescaping HTML entities and collapsing whitespace; because this ran after filtering, a few fall marginally below the lower bound (minimum 1,493 characters).

3.3 Rationale and effect on the sample

The lower length bound excludes postings that are stubs rather than complete advertisements: the collection contains descriptions as short as 2 characters, consisting of a role summary with no requirements section. The upper bound limits annotation cost, the binding constraint on corpus size, and duplicate removal prevents one posting contributing its skills twice.

Of the 123,842 postings with a description, 89,977 fall within the length bounds and 77,581 remain after deduplication, so the sample was drawn from 63% of the collection. Mean description length is 3,766 characters in the collection against 3,458 in the filtered sample, so the results characterise mid-length postings specifically.

3.4 Annotated subset

The evaluation corpus is the 60 annotated advertisements. They have a mean length of 3,355 characters (median 3,152) and contain 744 skill annotations, a mean of 12.4 per advertisement. Figure 1 summarises the construction of the corpus.

3.5 Annotation scheme

Extraction can be evaluated at the span level, marking character positions of each mention, or at the concept level, recording only the set of skills an advertisement states. We annotate at the concept level, reflecting the downstream use and reducing annotation cost.

We manually annotated each advertisement with the skills it states, following rules fixed before annotation began: a skill is recorded when the advertisement states it as a requirement, preference or responsibility; job titles are not skills; skills a reader would infer but the advertisement does not state are excluded; skills appearing only in equal-opportunity statements, benefits sections or company descriptions are excluded; and mentions with no adequate ESCO counterpart are recorded in the advertisement's own wording.

Annotation was performed by one annotator working from a sheet pre-populated with the lexicon's predictions, which the annotator corrected. This is post-editing rather than independent annotation, and its dependence on the lexicon is measurable: of the 391 proposals shown, 188 (48%) were rejected, and 541 of the 744 final annotations (73%) were added by the annotator. The set is therefore not a restatement of the lexicon's output, though neither is it independent of it; Section 6 discusses the residual bias.

4 Methods

4.1 Task definition

Each system takes an advertisement and returns a set of skills, compared against the reference annotations recorded for that advertisement. All outputs are compared as normalised strings, so every system is scored in the same way.

4.2 Lexicon-based extraction

We construct a lexicon from the ESCO skills classification, taking each concept's preferred label together with its alternative labels as surface forms and discarding forms of two characters or fewer: 13,939 concepts and 100,331 distinct surface forms. Extraction is case-insensitive n-gram matching against the advertisement text, with matched forms mapped back to their concept. The system requires no training data and is deterministic and auditable.

4.3 LLM-based extraction

We evaluate three open-weight instruction-tuned LLMs: Llama 3.3 70B Instruct [10], DeepSeek V3 [5] and Qwen 2.5 72B Instruct [11]. All are open-weight and available through hosted APIs, representative of what a practitioner can deploy without licensing a proprietary LLM; Llama 3.3 70B was chosen additionally for its use in comparable keyword-versus-LLM work [1]. All three were served through the OpenRouter API and prompted identically, zero-shot at temperature 0, requesting a JSON list of skill names, so differences reflect the LLMs rather than the serving configuration. The prompt instructs each LLM to list only skills the advertisement states, exclude the job title, and ignore skills appearing only in equal-opportunity statements, benefits sections or legal disclaimers, mirroring the annotation guideline. No advertisement approaches the LLMs' input limits. Responses failing to parse are retried once. Figure 2 gives the prompt verbatim.

```
You are extracting skills from a job advertisement.
List every skill, tool, technology, programming language, certification,
or professional competence that the advertisement states as a requirement,
preference, or responsibility of the role.
Rules:
- Only list skills the advertisement actually states. Do not infer skills.
- Do not list the job title itself.
- Ignore skills mentioned only in equal-opportunity statements, benefits
sections, or legal disclaimers.
- Use short skill names (1–4 words), lowercase.
Return ONLY a JSON array of strings. No explanation, no markdown.
```

Figure 2: The zero-shot extraction prompt.

	Relaxed			Exact		
System	P	R	F1	P	R	F1
Lexicon (ESCO)	**.519**	.273	.358	**.512**	.269	.352
Llama 3.3 70B	.400	.426	.412	.256	.273	.264
DeepSeek V3	.251	**.516**	.338	.163	.335	.219
Qwen 2.5 72B	.208	.469	.288	.109	.247	.152
Union (lex+Llama)	.384	.583	**.463**	.310	.472	**.374**
Union (lex+DeepSeek)	.261	.653	.373	.211	**.528**	.301
Union (lex+Qwen)	.226	.621	.332	.173	.476	.254

Table 1: Concept-level P, R and F1 on 60 advertisements (744 reference annotations).

4.4 Normalisation and hybrid combination

Because the LLMs emit free-form strings, all outputs, reference annotations included, pass through the same normalisation (lowercasing, punctuation stripping, whitespace collapsing), so no system is advantaged by the wording in which it expresses a skill. We additionally evaluate the union of the lexicon with the best-calibrated LLM, which operationalises the complementarity hypothesis of Section 1: if the two paradigms fail on different skills, their union should recover more than either component.

5 Evaluation and results

5.1 Metrics

We report micro-averaged precision, recall and F1, pooling true positives, false positives and false negatives across advertisements. Because systems express the same skill in different surface forms, we report two matching regimes. Exact matching requires string identity after normalisation. Relaxed matching additionally accepts a prediction when one string contains the other, or when token-level Jaccard similarity with a reference annotation is at least 0.6; for example, the output skill "read drawings" matches the reference annotation "read technical drawings" (Jaccard 2/3). Each reference annotation is matched at most once. Relaxed matching cannot prevent a loose match crediting a prediction a human would judge incorrect, so relaxed figures should be read as an upper bound and exact figures as a lower bound. Significance is assessed with McNemar's test [9], in its chi-square form with continuity correction, over the 744 (advertisement, reference annotation) instances.

5.2 Main results

The two paradigms separate cleanly. Table 1 reports both matching regimes. The lexicon is the most precise system under both; every LLM recalls more once wording differences are tolerated, DeepSeek most (.516 against .273). The union of the lexicon with Llama is the strongest configuration (F1 = .463 relaxed, .374 exact), recovering 58% of the reference annotations against the lexicon's 27%.

Comparison	b	c	χ^2	p
Lexicon vs Llama	113	227	37.6	$< 10^{-9}$
Lexicon vs DeepSeek	99	280	85.5	$< 10^{-15}$
Llama vs DeepSeek	31	98	33.8	$< 10^{-8}$
Llama vs Qwen	53	85	7.0	.008
DeepSeek vs Qwen	86	51	8.4	.004

Table 2: McNemar tests under relaxed matching, over 744 paired instances. b and c are instances retrieved by only the first and only the second system.

The three LLMs differ sharply in how much they output, though not in what they find. The lexicon predicts 6.5 skills per advertisement, Llama 13.2, DeepSeek 25.5 and Qwen 28.0, against 12.4 in the reference annotations. Llama is calibrated close to the annotation density while the other two over-generate and lose precision accordingly, so over-generation is a property of a particular LLM rather than of generative extraction as such. The surplus is systematic: the other two most often add dispositions (*integrity*, *positive attitude*, *adaptability*) and employment conditions that are not competences (*travel*, *lifting*, *background check*). Qwen's phrasing is also longer, 35% of its predictions running to three or more words against 22% of the reference annotations. Pairing the lexicon with either over-generating LLM raises recall (.653 and .621) but costs more precision than it gains.

A second difference concerns vocabulary rather than volume. The lexicon emits ESCO labels by construction, but the LLMs were never told to target the taxonomy: only 12.8% of Llama's predictions, 9.3% of DeepSeek's and 5.6% of Qwen's match an ESCO surface form, an ordering that matches the F1 ranking exactly. Predictions that do match ESCO are roughly twice as likely to be correct (65.7% against 36.1% for Llama, 55.2% against 21.9% for DeepSeek, 53.2% against 18.7% for Qwen). The reference annotations are themselves only 39% ESCO-recognised, so full coverage is not the target; the LLMs fall well short even of that. On 3 of the 60 advertisements the lexicon retrieved nothing.

5.3 Significance

The matching regime matters more for the LLMs than for the lexicon: relaxing the match adds 114, 135 and 165 true positives respectively, against 3 for the lexicon, because the lexicon emits canonical ESCO labels by construction while the LLMs paraphrase. Scoring by string identity therefore understates generative systems.

All five pairwise differences are significant, including those between the three LLMs (Table 2). We control the family-wise error rate with the Holm–Bonferroni procedure; all five remain significant at $\alpha = .05$, the weakest being Llama vs Qwen (adjusted $p = .008$). The union's recall gains are descriptive rather than separate hypothesis tests: under the same one-to-one matching it retrieves 231 more annotations than the lexicon and 117 more than Llama, while losing none, as follows by construction from taking the union. Complementarity is also visible in agreement between the sets of reference annotations retrieved by each system across the corpus: the three LLMs have pairwise Jaccard similarity .66 to .69, compared with .21 to .22 between each LLM and the lexicon. The LLMs retrieve similar annotations, whereas the lexicon retrieves different ones, which is why pairing paradigms helps more than pairing two LLMs.

5.4 Sources of lexicon false positives

Lexicon false positives are concentrated: *plan* (13), *manage time* (12), *think creatively*, *morality* and *history* (9 each) account for 52

of 188 false-positive instances (28%). Inspection shows most are triggered outside the requirement sections: "pregnancy", "morality" and "genetics" match equal-opportunity statements; "history" matches company descriptions and the phrase "work history"; "energy" and "electronics" match benefits copy. In one case an ESCO concept was matched against a query parameter inside an application URL.

5.5 Runtime and cost

Extraction required 4.6 s per advertisement for Llama, 7.2 s for DeepSeek and 7.8 s for Qwen; output ranged from 3,534 tokens for Llama to 10,068 for Qwen, mirroring the over-generation above. All three produced valid JSON on every advertisement, with zero parse failures across 180 extractions. The three runs cost approximately $0.01, $0.03 and $0.02 for Llama, DeepSeek and Qwen respectively, about $0.06 in total for the corpus, while the lexicon runs locally in milliseconds at no cost.

6 Discussion

Neither paradigm is adequate alone. The best single system reaches F1 = .412 and the hybrid .463, which is low in absolute terms and reflects both the difficulty of the task and the deliberately simple configurations we test.

The lexicon's weakness is contextual, not lexical: its errors reflect blindness to where a phrase occurs and in what sense, rather than missing vocabulary. "Manage time" is a genuine requirement in a retail lead role and a false positive in four advertisements where no such requirement appears; "metrology" is correct for a role selling measurement instrumentation and wrong for a payments executive. Restricting matching to requirement-bearing sections, or excluding a stop-list of over-general concepts, would remove many false positives without changing the method.

The LLMs' weakness is calibration to the annotation scheme, and it varies sharply between them. Llama produces approximately the right number of skills per advertisement. DeepSeek and Qwen produce twice as many, splitting single requirements into several skills and reporting attributes from benefits sections. Part of the disagreement is an artefact of our design: the prompt never mentions ESCO, so the LLMs are scored against a vocabulary they were not told to target, and only 5–13% of their output reaches it.

Because the two paradigms fail differently, their union recovers more than either. Where missing a required skill is costlier than carrying a spurious one, as in job recommendation, a hybrid extractor is preferable to either component alone.

6.1 Limitations

The generative side was evaluated under one zero-shot configuration: the prompt did not mention ESCO, give examples, or constrain output to ESCO candidates. Because so little LLM output is expressed as ESCO terms, exact matching has limited interpretive value for them; the large gain under relaxed matching (114, 135 and 165 added true positives against 3 for the lexicon) confirms this. The results therefore characterise these LLMs under one deliberately simple prompt rather than generative extraction in general. The hybrid union's superiority is likewise not guaranteed under stronger configurations that address these limitations; better calibration or ESCO alignment could change the balance between the components.

The corpus is small (60 advertisements), English-language, and all systems were evaluated only on advertisements passing the filters of Section 3.2. Short postings tend to lack the requirements sections all systems depend on, so the lexicon's recall on them is plausibly lower than reported.

Annotation was performed by a single annotator, so no agreement statistic is available, and the reference set was produced by post-editing the lexicon's predictions, so it is not independent of one of the systems it scores. The bias runs in a known direction: a retained false positive inflates the lexicon's precision, and a skill it never proposed and the annotator did not add is absent from the reference set, inflating its recall. The reported lexicon precision of .519 should be read as an upper bound. The same bias understates every LLM, so the finding that all three recall more than the lexicon is conservative, as is the hybrid's advantage over the lexicon; its advantage over Llama is not, because the union contains the lexicon's output and so shares the bias in its favour. Annotating a further subset with no system output visible would quantify this directly.

7 Conclusion and future work

We compared lexicon-based and LLM-based skill extraction on a manually annotated corpus of 60 job advertisements with 744 skill annotations. The lexicon is most precise (.519), all three LLMs recall more (.43 to .52), and pairing the lexicon with the best-calibrated LLM is the strongest configuration (F1 = .463). Five ESCO concepts account for 28% of lexicon false positives, most triggered by text outside the requirement sections.

Three directions follow. Restricting lexicon lookup to passages containing requirements is a minimal change that would remove many false positives. The generative side was left in its simplest form: naming ESCO in the prompt, adding few-shot examples, or re-ranking output against retrieved ESCO candidates are untried, and the gap between ESCO-recognised and unrecognised predictions suggests a validation layer alone would recover much of the lost precision. Finally, the extracted skills feed the graph-based candidate–vacancy matching of the recommender this work was built for.

The annotated corpus, the extraction and evaluation code, and the full result files are available at https://github.com/naatteeee/skill-extraction.

Acknowledgements

The research was supported through the European project Raido (GA 101135800). The authors acknowledge usage of LLMs in grammar refinement and overall manuscript organization and take full responsibility for the final content of this publication.

References

[1] Janez Brank, Inna Novalija, Dunja Mladenić, and Marko Grobelnik. 2025. LLM based approach to extracting smells in Slovenian corpora. In *Proceedings of Data Mining and Data Warehouses (SiKDD 2025)*. DOI: 10.70314/is.2025.sikdd.5.

[2] Benjamin Clavié and Guillaume Soulié. 2023. Large language models as batteries-included zero-shot ESCO skills matchers. In *Proceedings of the 3rd Workshop on Recommender Systems for Human Resources (RecSys-in-HR 2023)* (CEUR Workshop Proceedings). Vol. 3490. arXiv: 2307.03539.

[3] Jens-Joris Decorte, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. 2022. Design of negative sampling strategies for distantly supervised skill extraction. In *Proceedings of the 2nd Workshop on Recommender Systems for Human Resources (RecSys-in-HR 2022)* (CEUR Workshop Proceedings). Vol. 3218.

[4] Jens-Joris Decorte, Severine Verlinden, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Extreme multi-label skill extraction training using large language models. In *International Workshop on AI for Human Resources and Public Employment Services (AI4HR&PES) at ECML-PKDD 2023*. arXiv: 2307.10778.

[5] DeepSeek-AI. 2024. DeepSeek-V3 model card. Hugging Face. (2024).

[6] European Commission. 2025. ESCO v1.2.1. esco.ec.europa.eu. (2025).

[7] Sofija Kochovska, Branko Kavšek, and Jernej Vičič. 2025. A hybrid lexicon-machine learning approach to Macedonian sentiment analysis. In *Proceedings of Data Mining and Data Warehouses (SiKDD 2025)*.

[8] Arsh Koneru. 2024. LinkedIn job postings (2023–2024). Kaggle. CC BY-SA 4.0. (2024).

[9] Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153–157. DOI: 10.1007/BF02295996.

[10] Meta AI. 2024. Llama-3.3-70B-Instruct model card. Hugging Face. (2024).

[11] Qwen Team. 2024. Qwen2.5-72B-Instruct model card. Hugging Face. (2024).

[12] Elena Senger, Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. Deep learning-based computational job market analysis: a survey on skill extraction and classification from job postings. In *Proceedings of NLP4HR 2024*, 1–15. DOI: 10.18653/v1/2024.nlp4hr-1.1.

[13] Mike Zhang, Kristian Nørgaard Jensen, Sif Dam Sonniks, and Barbara Plank. 2022. SkillSpan: hard and soft skill extraction from English job postings. In *Proceedings of NAACL-HLT 2022*, 4962–4984. DOI: 10.18653/v1/2022.naacl-main.366.

[14] Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. Entity linking in the job market domain. In *Findings of EACL 2024*, 410–419. DOI: 10.18653/v1/2024.findings-eacl.28.