

# Large Language Models as Batteries-Included Zero-Shot ESCO Skills Matchers

Benjamin Clavié<sup>1</sup>, Guillaume Soulié<sup>1</sup>

<sup>1</sup>Bright Network Ltd., Edinburgh, UK

## Abstract

Understanding labour market dynamics requires accurately identifying the skills required for and possessed by the workforce. Automation techniques are increasingly being developed to support this effort. However, automatically extracting skills from job postings is challenging due to the vast number of existing skills. The ESCO (European Skills, Competences, Qualifications and Occupations) framework provides a useful reference, listing over 13,000 individual skills. However, skills extraction remains difficult and accurately matching job posts to the ESCO taxonomy is an open problem. In this work, we propose an end-to-end zero-shot system for skills extraction from job descriptions based on large language models (LLMs). We generate synthetic training data for the entirety of ESCO skills and train a classifier to extract skill mentions from job posts. We also employ a similarity retriever to generate skill candidates which are then re-ranked using a second LLM. Using synthetic data achieves an RP@10 score 10 points higher than previous distant supervision approaches. Adding GPT-4 re-ranking improves RP@10 by over 22 points over previous methods. We also show that Framing the task as mock programming when prompting the LLM can lead to better performance than natural language prompts, especially with weaker LLMs. We demonstrate the potential of integrating large language models at both ends of skills matching pipelines. Our approach requires no human annotations and achieve extremely promising results on skills extraction against ESCO.

## Keywords

Skills Matching, ESCO, Large Language Models, Extreme Multi-Label Classification, Synthetic Data Generation

## 1. Introduction

The Job Market is often described as constantly evolving, as technological developments and societal changes result in large changes in its makeup, as has been frequently studied and demonstrated [1, 2]. In recent years, the rapid digitisation of society has led to entirely new categories of skills becoming requirements for many jobs[3]. As demands for skills evolve, there is an increasing need to better understand the skills required by jobs. This has supported the continuous development of skill taxonomies such as the European Union’s European Skills, Competences, Qualifications and Occupations, or **ESCO** [4], framework, developed to improve understanding and efficiency of the wider EU job market.

While useful, such frameworks require **skill extraction** (SE) approaches to understand skills present in job postings and enable automation at scale. Skills Extraction has recently been the subject of an increased amount of interest[5], which has been further compounded by research showing that a large proportion of required skills are implicitly expressed rather than explicitly stated within

postings [6].

However, automated skills extraction faces considerable roadblocks. Most notably, many efforts are limited in scale by the lack of available data for both training and evaluation. Recent efforts have begun to try and alleviate this effort, by publicly releasing manually annotated datasets. These approaches, while promising, suffer from the complexity of the task, as the ESCO taxonomy contains 13 890 individual skills. As a result, the task is often re-framed or simplified, for example by converting the task into a span-extraction task [7, 8], leaving direct matching to the taxonomy for future work. Some recent work has explored this Extreme Multi-Label Classification (XLMC) task with an end-to-end approach to matching and extraction with non-ESCO taxonomies, with encouraging results [6, 9], although sometimes relying on simplifying the taxonomy by using only higher-level labels [10].

In fact, the task of *Skills Extraction against a taxonomy* could be framed as a two-tasks process: ① an **extraction** step, focused on recognising the potential mentions of skills, or groups of skills, from the content of job postings, on which strong progress has been made [7, 11, 6], and ② a **matching** step, akin to extreme multi-label classification, focused on linking these mentions with fine-grained taxonomies, which remains a difficult problem.

The sheer number of existing skills makes it very difficult to obtain sufficient training data for comprehensive coverage. As such, various techniques, such as using the ESCO API as a form of distant supervision [12] and

*RecSys in HR’23: The 3rd Workshop on Recommender Systems for Human Resources, in conjunction with the 17th ACM Conference on Recommender Systems, September 18–22, 2023, Singapore, Singapore.*

Corresponding author: Benjamin Clavié

✉ ben.clavie@brightnetwork.co.uk (B. Clavié);

guillaume.soulie@brightnetwork.co.uk (G. Soulié)

☎ 0000-0002-8346-8703 (B. Clavié)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License

Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

generating training data through such distant supervision techniques [13], have been explored.

Meanwhile, the rapid development and improvement of generative large language models (LLMs) [14] and especially instruction-tuned LLMs [15, 16], models further trained to specifically follow natural language instructions, has resulted in the use of generative LLMs on a large array of applications, often yielding competitive or even state-of-the-art results across many tasks, as highlighted by the release and strong performance of OpenAI’s GPT-4 [17]. Notably, LLMs have shown their ability to improve performance on text retrieval tasks through the generation of synthetic training data based on a handful of real examples [18]. They have also shown considerable problem-solving ability, often anthropomorphised as *reasoning*, which appears to be even stronger when the task is framed as a programming task [19].

Due to the nature of the training data used by Large Language Models, which contains very large volumes of content scraped from across the internet [20], we hypothesise that the knowledge representation embedded within the models makes them particularly suitable for broader job understanding tasks, as many job postings, as well as skills descriptions, were present in this data. As such, we will explore the creation of a **zero-shot skills matching** pipeline through the use of large language models, focusing in this case on the use of GPT-3.5 and GPT-4.

**Contributions** In this paper, we show that: ① Large Language Models can reliably generate zero-shot training data that improves performance in the skill matching task. We then show that using this data to power both similarity-based retrieval approaches and linear classifier models trained on the data can generate good lists of potential skills found within given text extracts, outperforming the previous approaches. ② The **Skill Matching** task can be framed as a two-step problem, with the first stage consisting of generating a list of potential skill matches and the second stage focusing on re-ranking these potential matches. We show that LLMs can be used as zero-shot rerankers for this second step of the extraction pipeline with very strong performance without the need for annotated data. ③ Framing the skills matching task as a mock programming problem provides a further performance boost for both large language models tested. More notably, it improves the performance of the less capable model by over 10 percentage points by enhancing its ability to follow instructions.

## 2. LLM-Generated Data

As discussed, we follow a two-step process to form the **Skills Matching pipeline**. While, as we will demonstrate later, LLMs are strong zero-shot rerankers, it’s impossible

for them to work as standalone classifiers for all ESCO skills, as the 13890 skills listed in ESCO do not fit within the context window of most Language Models. Even if this were possible, the increased compute and processing time required by current Transformer-based models (used by all LLMs) would make this approach unsuitable in many cases, despite promising advancements in more computationally efficient approaches to Transformer models [21].

We thus choose to leverage large generative models for the skills matching task through the generation of **synthetic training data**. We believe that this synthetic training could allow us to perform the skills matching task without having to reframe the task or rely on useful but limited distant supervision techniques[13].

For each of the 13890 skills contained in ESCO, we prompt GPT-3.5<sup>1</sup> to generate forty example sentences that could be used in a job posting in order to refer to the skill. We specifically request that the sentences be phrased in a variety of ways, and be of various lengths (from just a few words directly referring to the skill to a few sentences mentioning it implicitly). We provide slightly different instructions based on the "skill type", for instance requesting explicit mentions in more examples when generating data for a skill contained in the **tech** ESCO skill listing, as programming languages tend to be clearly mentioned in job ads. We also use additional information or skill descriptions present in the ESCO data to enrich our prompt and help the model disambiguate between potentially ambiguous terms.

Once the training data is generated, it is not thoroughly manually reviewed. We performed programmatic checks, showing that a full forty examples were generated for more than 97% of skills, with a few of them having fewer examples due to model context size limits or failure to properly follow instructions, which were not addressed, as the entirety of skills but one<sup>2</sup> had more than 30 generated examples. We sampled a random 100 skills for manual review to ensure that the generated data met our criteria and that the prompts contained no obvious mistakes.

The prompts used for data generation are provided in appendix A.

---

<sup>1</sup>Our limited experiments showed that all recent LLMs, including the open-source Flan-UL2 [22], could potentially perform this task without a strong negative impact on performance. Thoroughly evaluating different LLMs for this data generation step is beyond the scope of this work, but would likely be a valuable future area of research, as we noticed considerable style differences between different models, leading us to believe a combination of different ones could generate a more diverse training data set.

<sup>2</sup>The skill "semen insertion", related to veterinary work, could not have training data generated for it by the GPT model family, as its name triggers OpenAI’s content filters.

### 3. Potential Skills Identification

Following data generation, our next step is the identification of potential skills contained within a given text span. We generate these potential skills through two main approaches: a **linear classifier-based approach**, using linear regression classifiers on frozen embeddings, and **textual similarity approaches**, where we rely on the cosine distance between embeddings to determine whether a skill is potentially present or not.

For both of these approaches, we use the E5-LARGE-V2 text embedding model [23], the current state-of-the-art embedder for similarity-based information retrieval whose generated embeddings have also been shown to reach strong performance for few-shot text classifications, making it particularly suitable for both of our approaches.

#### 3.1. Classifier Candidates

Our first set of candidates is generated by the use of simple logistic regression classifiers. We train one binary classifier per label, in a one-versus-all classification approach. We use no real-world data for these classifiers, instead training only on the synthetic data generated as described in section 3. For any given class, we treat all example sentences generated for the label as positive examples and sample twice as many examples from other labels to use as negative examples.

Following previous work on skill classification [13], we use partial hard negative sampling [24] to ensure our models are better at distinguishing between very similar labels. To do so, we make it so 10% of the negative examples are hard negatives, sampled from labels associated with the example sentences having the highest cosine similarity with the positive examples.

At inference time, we consider every positive classification from the classifiers as a candidate label to be used for re-ranking.

#### 3.2. Similarity-based Candidates

Our second approach to generating candidate is based on **cosine similarity** between embeddings. We generate candidate through two distinct approaches: **label similarity** and **sentence similarity**.

**Label Similarity** This is a simple similarity look-up between a target extract and the full list of existing labels. We do not set a threshold for this step, rather, we treat the 40 most similar labels as candidates to be provided to the re-ranker.

**Sentence Similarity** For this candidate generator, we use the cosine distance between the current extract and

the synthetic example sentences. If two sentences with the same label are part of the 40 most similar sentences, we add the label to the list of candidates. This is in effect similar to a simplified form of k-nearest neighbour classification [25].

### 4. Zero-Shot Potential Skills Ranking

Once the list of potential skills found in a given span is generated, we prompt an LLM to extract and rank the ten most likely skills in order of suitability. This is the **reranking** step, a key component of information retrieval pipelines, which is increasingly performed by fine-tuned language models [26]. In this work, we explore how a zero-shot approach leveraging LLMs' "learned knowledge" performs on our task.

#### 4.1. A Note About the Evaluated LLMs

We report results for both GPT 3.5, an instruction-tuned [16], more powerful version of GPT-3 [27], more specifically GPT-3.5-TURBO-0301 and GPT 4 [17] (GPT-4-0314), which produced the most promising results in our exploratory work. Most of our prompt engineering work was performed for GPT 4, and re-used as-is for GPT 3.5.

While we have not conducted extensive experiments using them, our exploratory work has shown that open-source LLMs, of which FLAN-UL2 [22] was the best performing at the time of this work, failed to produce reliable outputs, frequently "hallucinating" skills, in a way similar to GPT 3.5. However, in the case of GPT 3.5, this was entirely mitigated by the *Python approached* described below.

At the time of conducting our experiments, FALCON [28] had not yet been released and the authors did not have access to LLAMA [29].

#### 4.2. LLMs as Reranker

We use prompting and prompt engineering [30] and describe the task or reranking in the prompt. We use a chat-formatted prompt, through OpenAI's ChatML [31]. We give the model a broad description of its role as its initial prompt, followed by the detailed instructions for the task, and a mocked message from the model acknowledging and summarising the instructions.

We then provide the model with a list of potential skills, generated by the previously described methods. We experiment with both providing the model with information about the score it received as a potential skill, through either classification class probability or textual similarity, depending on the potential skill's source. We found this had no impact on performance, and therefore do not provide this information to the model in our final evaluation

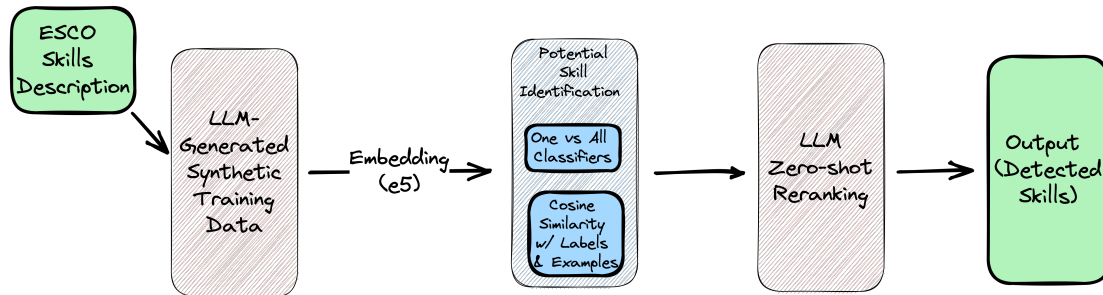


Figure 1: High-level overview of the full process.

to reduce the number of tokens used in our prompts, thus reducing the required compute. We pass all skills identified by the CLASSIFIER approach as well as up to 60 skills identified by the SIMILARITY-BASED APPROACH.

We request an ordered list of the ten most likely skill matches in our prompt. In all cases, we provide the model with the ability to use the *NO\_LABEL* skill to reach 10 skills it identifies fewer or no matches.

All the prompts used for the reranking task are provided in Appendix A.

#### 4.2.1. Mock Python Programming Variant

Recent work has shown that Large Language Models can often perform better on "reasoning" tasks when they are approached as programming exercises [19]. Additionally, anecdotal evidence often states that it is easier to control the output of large language models when requiring programming-like outputs, supposedly due to programming languages' more structured nature. While investigating the full extent of these claims is beyond the scope of this work, we experiment with modifying our re-ranking prompt to include explicit instructions to answer exclusively in Python, in the form of a function returning an ordered list of ranked skills, and outputting the justification for inclusion as a comment. No other modifications to the instructions are made.

We choose Python over other programming languages as it is often a good proportion of the programming data commonly used to train and evaluate LLMs [32, 33] and requires very little adjustments to the existing re-ranking step presented in the previous section, which is itself implemented in Python.

## 5. Experimental Setup

As ESCO-based skills matching is an Extreme Multi-Label Classification (XMLC) task, we choose to frame

it similarly to an information retrieval task and use IR-inspired methods, for which we provide a high-level overview of our architecture in Figure 1. Skills matching against the ESCO taxonomy, due to the very large number and granularity of skill labels, justifies this framing: while we want to assign as many labels present in our test set as possible, it is also highly likely that many potentially relevant labels are not attached to our test examples, either because of oversight or because of subjective judgement in situations where multiple similar labels applied. As such, our aim is to maximise our retrieval of test labels, without harsh penalties for additional labels assigned by the model.

## 5.1. Evaluation

### 5.1.1. Data

We evaluate our approach on the dataset provided by Decorte et al. [13]. Their work built upon the SkillsSpan dataset[7], a publicly available dataset focused on the detection of text spans containing the mention of either *skills* or *knowledge*, which are two sub-categories of *skills* as broadly defined within the ESCO framework. Using the extracted spans, Decorte et al. manually assigned ESCO skills to the extracted spans in order to create a dataset of spans annotated with the matching ESCO skill(s). To the best of our knowledge, this represents the best effort at creating an evaluation dataset using the full extent of ESCO's fine-grained approach rather than approximations or groupings. We use the *validation* set provided to tweak our prompts and evaluate our models on the *test* set.

The data contains two distinct subsets, *TECH*, which focuses on data extracted from jobs within the tech sector, and *HOUSE*, containing more generalist jobs. We report results for each set separately, following the literature.

Both subsets contain a list of skills that were identified within a given span by human annotators.

	House				Tech			
	MRR	RP@1	RP@5	RP@10	MRR	RP@1	RP@5	RP@10
Decorte et al. [13] (best approach)	0.299	N/A	30.82	38.69	0.339	N/A	31.71	39.19
Classifiers (no rerank)	0.326	<b>27.20</b>	<b>37.60</b>	<b>46.47</b>	0.299	27.16	33.41	39.86
Similarity (no rerank)	<b>0.355</b>	26.44	35.22	43.73	<b>0.405</b>	<b>32.84</b>	<b>49.67</b>	<b>58.66</b>
GPT3.5 Re-ranking								
+Classifier	0.232	18.32	24.10	27.94	0.279	21.95	29.30	32.48
+Similarity	0.369	29.39	34.40	38.93	0.413	35.01	43.26	47.15
+Both	0.372	27.02	32.93	38.68	0.369	29.67	37.55	43.24
+Both + Python	<b>0.427</b>	<b>36.92</b>	<b>43.57</b>	<b>51.44</b>	<b>0.488</b>	<b>40.53</b>	<b>52.50</b>	<b>59.75</b>
GPT4 Re-ranking								
+Classifier	0.446	37.16	48.40	53.44	0.442	39.10	46.77	51.70
+Similarity	0.467	36.40	48.35	54.52	0.481	40.82	54.15	62.71
+Both	<b>0.507</b>	<b>42.91</b>	<b>56.67</b>	60.09	0.512	45.67	59.47	64.03
+Both + Python	0.495	40.70	53.34	<b>61.02</b>	<b>0.537</b>	<b>46.52</b>	<b>61.50</b>	<b>68.94</b>

Table 1.: Results for the various skills matching approaches. Best results within a category in *italicised bold*, best overall results in **bold**

### 5.1.2. Metrics

As the authors introducing the dataset note [13], it is highly unlikely for human annotation to be fully exhaustive given the wealth of label, as such, the aim of this task is to extract as many labels as possible, without penalising the model for "near misses", which could still be appropriate labels. We thus follow their lead and report the macro-averaged **R-Precision@k (RP@k)**, which is particularly well-suited to evaluated extreme multilabel classification tasks such as this one [34, 13] as well as the **Mean Reciprocal Rank (MRR)** of the highest ranked correct label as a further indication of ranking quality.

## 6. Results and Discussions

The results of our experiments are presented in in Table 1. We report the performance of the full pipeline, with both GPT 3.5 and GPT 4 re-ranking, as well as the previous state-of-the-art performance obtained by Decorte et al. in the paper introducing the dataset [13]. We also report the results of both our **Classifier** and **Similarity** approaches without the re-ranking step, both to showcase the performance obtained via the use of LLM-generated training data and to serve as a baseline for the re-ranking approaches.

We notice that, on their own, both of these no-reranking approaches achieve competitive performance against previous methods, with the similarity approach marginally outperforming the classifier one on the **House** dataset but performing noticeably worse on the **Tech** one. These results are encouraging, as they require no real-world training data and are extremely fast at inference-time, requiring only simple computations. Their **RP@k** scales particularly well with higher  $k$  values, highlighting their

ability to propose a number of correct labels but not ranking them optimally.

GPT-4 reranking results in considerable improvements over all non-reranked methods, strongly outperforming all other methods in all approaches and the best-performing variant reaching an RP@10 of **61.02** on the **House** dataset and **68.94** on the **Tech**, a respective improvement of **22.33** and **29.75** percentage points over the previous best approach and **14.55** and **10.28** over our best non-reranked methods. We notice that in all cases, the performance obtained by combining potential skills generated by both the classifier and the similarity approaches is noticeably stronger than when using only one method of generating candidates. However, when using a single method of generating potential candidates, we notice that the similarity-based approach tends to outperforms the classifier-based approach on both datasets, especially on the **House** dataset.

The performance of GPT-3.5 re-ranking is more mixed. With natural language prompting, its performance is an overall downgrade over the non-reranked approaches. Unlike GPT-4, we also notice that combining both methods of potential skill generation does not systematically improve performance, especially on the **Tech** dataset where using only similarity-based entries resulted in overall stronger results. When using only the classifiers-based candidates, we notice that the GPT-3.5 ranking actually decreases performance. One of the noticeable reasons for this weaker performance is GPT-3.5's seemingly weaker ability to follow guidelines: despite our experiments in modifying the model prompt, it would frequently "hallucinate" skills whose wording was directly inspired from the target span, and ranking them higher than the skills provided.

For both GPT variants, we notice strong performance



with the **Python** prompt variation, where we explicitly request that the model output is a Python function returning the ranked list of skills. In the case of GPT-4, the Python variant significantly outperforms natural language prompting on the **Tech** dataset, but performs slightly worse on **House** for all metrics but RP@10. For GPT-3.5, however, Python prompting nearly entirely eliminates the problem of hallucinating skills, and greatly improves the performance across all metrics on both datasets. This appears to suggest that framing the problem as a programming problem, which are frequently used to train LLMs, helps ground reasoning and improve performance in re-ranking tasks in a way natural language prompt engineering cannot, although more experiments are needed to confirm this.

Overall, the use of LLM-generated training data outperforms the state-of-the-art distant supervision approaches, and that zero-shot LLM re-ranking further increases performance, considerably outperforming all previous approaches.

## 7. Limitations and Future Work

While our work shows very strong potential for LLMs in both generating training data and improving inference-time predictions for skills matching, we believe that there are three key limitations to our work that should be explored in future work.

**Broader Scope** We focus on a small, focused dataset which has previously been explored in the literature. We believe that our approach is likely to generalise well to both other taxonomies and different datasets relying on ESCO. We believe future work should explore building upon this method to explore more data sources and evaluation approaches.

**Representation Types** Our study explores only the use of e5 [23] embeddings, due to their very strong out-of-the-box performance. However, these embeddings are general domain representations and are only one approach among many. We believe future work exploring different approaches to representation could yield better results and valuable insight. Notably, further exploring techniques common within the field of information retrieval, utilising powerful cross-encoders such as ColBERT, and combining deep-learning based forms of representations with simpler but powerful approaches such as tf-idf capturing different kinds of information could prove very valuable.

**LLMs Used** This work uses the GPT family of model, and more specifically, GPT-4. These models are gated behind APIs and their weights are not publicly available. While they perform well, future work should explore the applicability of open-source LLMs, such as Falcon [28], as well as look for more efficient approaches. Additionally, we intend to explore if using a more diverse set of

generative models, trained on different datasets, could improve our synthetic training data generation by generating more semantically varied examples.

**Domain-Specific Models** Our approach focuses on the use of general domain model, with no further training to adapt them to the language used within job postings specifically. While we believe that this kind of information is present within the training corpora of the large language models we use, we believe that better targeted models could facilitate the development of more efficient approaches as mentioned above. Notably, models such as JobBERT [35] and ESCOXML-R [5] have shown the potential of domain-specific fine-tuning on existing tasks. Meanwhile, the LLM literature highlights how considerably smaller language models, with an order of magnitude fewer parameters than GPT-3, can reach competitive performance through fine-tuning on small but very high quality datasets [36, 37].

**Potential Skills Generation** Our experiments indicate that varying the number of potential skills given to GPT-4 does not have a major impact, if any, on its ranking performance. However, we did not extensively experiment with different ways of generating the potential skills list, and the impact that prompt modifications, such as different ordering or indicating the source or probability given to a label by the initial classifier would have. Additionally, we conducted only very moderate experimentation in optimising the hyper-parameters of our classifier-based candidate generation (as described in Appendix B) or with alternate ways of computing similarity, such as using SVM-based retrievers [38]. We plan to explore these optimisations in future work.

**Impact on Recommender Systems** While out of the scope of this study, one of the key goals of better job/skill matching is facilitating the use of recommender systems to highlight good matches between jobseekers and job postings, thus contributing to alleviating the job/skill mismatch [39]. Our early results in using the output of the pipeline introduced in this paper have shown promising results, and we intend to further explore the best use of this skills extraction pipeline within end-to-end job recommender systems in future work.

## 8. Conclusion

In this work, we have proposed a novel end-to-end zero-shot pipeline for skills matching against the ESCO taxonomy using Large Language Models (LLMs). We have shown that LLMs can generate high-quality synthetic training data to improve candidate generation, outperforming existing approaches without needing any non-synthetic training data. We have also demonstrated that state-of-the-art LLMs can act as strong zero-shot re-rankers as the final step of the skill matching pipeline,

resulting in another large performance improvement.

Our experiments also show that framing the re-ranking task as a mock Python programming problem results in significant performance gains, especially for less capable models. We believe that this framing helps the models better follow the task instructions in re-ranking contexts, especially when working with less powerful models.

Overall, our work highlights the strong potential for Large Language Models for the low-resource context of working with the ESCO taxonomy, through leveraging the limited information present in the taxonomy to guide the generation of targeted synthetic data, as well as through zero-shot application of their capabilities. While our experiments have focused on a single dataset and taxonomy, namely ESCO, we believe that our approach holds potential to support further work in automated understanding of the job market at scale, and we release the prompts we have used in order to support these efforts.

## References

- [1] P. A. Todd, J. D. McKeen, R. B. Gallupe, The evolution of is job skills: A content analysis of is job advertisements from 1970 to 1990, *MIS quarterly* (1995) 1–27.
- [2] V. World Economic Forum, The future of jobs report 2020, *WEF Reports* (2020).
- [3] E. Brumberger, C. Lauer, The evolution of technical communication: An analysis of industry job postings, *Technical Communication* 62 (2015) 224–243.
- [4] M. le Vrang, A. Papantoniou, E. Pauwels, P. Fannes, D. Vandestein, J. De Smedt, *Esco: Boosting job matching in europe with semantic interoperability*, *Computer* 47 (2014) 57–64.
- [5] M. Zhang, R. van der Goot, B. Plank, *Escoxmlr: Multilingual taxonomy-driven pre-training for the job market domain*, *arXiv preprint arXiv:2305.12092* (2023).
- [6] A. Bholra, K. Halder, A. Prasad, M.-Y. Kan, Retrieving skills from job descriptions: A language model based extreme multi-label classification framework, in: *Proceedings of the 28th international conference on computational linguistics*, 2020, pp. 5832–5842.
- [7] M. Zhang, K. Jensen, S. Sonniks, B. 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*, 2022, pp. 4962–4984.
- [8] L. Sayfullina, E. Malmi, J. Kannala, Learning representations for soft skill matching, in: *Analysis of Images, Social Networks and Texts: 7th International Conference, AIST 2018, Moscow, Russia*, July 5–7, 2018, *Revised Selected Papers 7*, Springer, 2018, pp. 141–152.
- [9] N. Goyal, J. Kalra, C. Sharma, R. Mutharaju, N. Sachdeva, P. Kumaraguru, *Jobxmlc: Extreme multi-label classification of job skills with graph neural networks*, in: *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 2136–2146.
- [10] K. F. F. Jiechieu, N. Tsopze, *Skills prediction based on multi-label resume classification using cnn with model predictions explanation*, *Neural Computing and Applications* 33 (2021) 5069–5087.
- [11] D. Beauchemin, J. Laumonier, Y. L. Ster, M. Yasmine, "fijo": a french insurance soft skill detection dataset, *arXiv preprint arXiv:2204.05208* (2022).
- [12] M. Zhang, K. N. Jensen, B. Plank, *Kompetencer: Fine-grained skill classification in danish job postings via distant supervision and transfer learning*, *arXiv preprint arXiv:2205.01381* (2022).
- [13] J.-J. Decorte, J. V. Haute, J. Deleu, C. Develder, T. Demeester, *Design of negative sampling strategies for distantly supervised skill extraction*, 2022. *arXiv:2209.05987*.
- [14] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., *A survey of large language models*, *arXiv preprint arXiv:2303.18223* (2023).
- [15] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al., *Scaling instruction-finetuned language models*, *arXiv preprint arXiv:2210.11416* (2022).
- [16] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., *Training language models to follow instructions with human feedback*, *arXiv preprint arXiv:2203.02155* (2022).
- [17] OpenAI, *Gpt-4 technical report*, 2023. *arXiv:2303.08774*.
- [18] L. Bonifacio, H. Abonizio, M. Fadaee, R. Nogueira, *Inpars: Data augmentation for information retrieval using large language models*, 2022. *arXiv:2202.05144*.
- [19] A. Madaan, S. Zhou, U. Alon, Y. Yang, G. Neubig, *Language models of code are few-shot common-sense learners*, *arXiv preprint arXiv:2210.07128* (2022).
- [20] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, M. Gardner, *Documenting large webtext corpora: A case study on the colossal clean crawled corpus*, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 1286–1305.
- [21] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, C. Ré, *Flashattention: Fast and memory-efficient exact attention*

- with io-awareness, 2022. arXiv:2205.14135.
- [22] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, D. Bahri, T. Schuster, S. Zheng, et al., Ul2: Unifying language learning paradigms, in: The Eleventh International Conference on Learning Representations, 2023.
- [23] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, 2022. arXiv:2212.03533.
- [24] J. Robinson, C.-Y. Chuang, S. Sra, S. Jegelka, Contrastive learning with hard negative samples, 2021. arXiv:2010.04592.
- [25] P. Cunningham, S. J. Delany, k-nearest neighbour classifiers-a tutorial, ACM computing surveys (CSUR) 54 (2021) 1–25.
- [26] R. Nogueira, K. Cho, Passage re-ranking with bert, 2020. arXiv:1901.04085.
- [27] T. B. Brown, et al., Language models are few-shot learners, NeurIPS 2020 (2020).
- [28] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocar, M. Debbah, E. Goffinet, D. Hessel, J. Launay, Q. Malartic, B. Noune, B. Pannier, G. Penedo, Falcon-40B: an open large language model with state-of-the-art performance (2023).
- [29] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
- [30] E. Saravia, Prompt Engineering Guide, <https://github.com/dair-ai/Prompt-Engineering-Guide> (2022).
- [31] Microsoft, Learn how to work with the ChatGPT and GPT-4 models (preview), 2023.
- [32] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al., Evaluating large language models trained on code, arXiv preprint arXiv:2107.03374 (2021).
- [33] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311 (2022).
- [34] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, I. Androutsopoulos, Large-scale multi-label text classification on eu legislation, arXiv preprint arXiv:1906.02192 (2019).
- [35] J.-J. Decorte, J. Van Haute, T. Demeester, C. Develder, Jobbert: Understanding job titles through skills, arXiv preprint arXiv:2109.09605 (2021).
- [36] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, A. Awadallah, Orca: Progressive learning from complex explanation traces of gpt-4, arXiv preprint arXiv:2306.02707 (2023).
- [37] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4, 2023. URL: <https://lmsys.org/blog/2023-03-30-vicuna/>.
- [38] T. Malisiewicz, A. Gupta, A. A. Efros, Ensemble of exemplar-svms for object detection and beyond, in: 2011 International conference on computer vision, IEEE, 2011, pp. 89–96.
- [39] S. McGuinness, K. Pouliakas, P. Redmond, Skills mismatch: Concepts, measurement and policy approaches, Journal of Economic Surveys 32 (2018) 985–1015.
- [40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.

## A. Prompts

### A.1. Training Data Generation

You are the leading AI Writer at a large, multinational HR agency. You are considered as the world’s best expert at expressing required skills and knowledge in a variety of clear ways. You are particularly proficient with the ESCO Occupation and Skills framework. As you are widely lauded for your job posting writing ability, you will assist the user in all job-posting, job requirements and occupational skills related tasks.

You work in collaboration with ESCO to gather rigid standards for job postings. Given a list of ESCO skills and knowledges, you’re asked to provide forty examples that could be found in a job ad and refer to the skill or knowledge component. You may be given a skill family to help you disambiguate if the skill name could refer to multiple things. Ensure that your examples are well written and could be found in real job advertisement.

Write a variety of different sentences and ensure your examples are well diversified. Use a variety of styles. Write examples using both shorter and longer sentences, as well as examples using short paragraphs of a few sentences, where sometimes only one is directly relevant to the skill. You’re trying to provide a representative sample of the many, many ways real job postings would evoke a skill.

At least {FIVE for tech skills, ZERO for language skills, 80% (THIRTY-TWO)} of your examples must not contain an explicit reference to the skill and must thus not contain the given skill string. Extra Information/Alternative Names (you may discard this information if irrelevant): {ALTERNATE NAMES IN THE ESCO DATABASE}



Avoid explicitly using the wording of this extra information in your examples. Skill: {target}""

## A.2. Reranking

**Instructions:** You are given an extract from a job posting. As an AI job and skills expert, you need to assist in whatever task is requested of you. I will give you a sentence referring to a skill extracted from a job posting, as well as a list of potential skill labels. You are asked to extract and rank the likely skills from the candidates list into a ranked list of 10.

It is possible that none match, in which case you will say NO\_LABEL. You must either use one from the list or NO\_LABEL.

You may not use any label not provided in the example list. If you use NO\_LABEL, do not assign any other label.

You will rank the top 10 most likely labels from the candidates, and provide an explanation as to why they are picked and ranked where they are.

That means that if two labels are applicable, but one is much broader, you should pick the less broad one slightly above the broader one. For example, a skill related to specific kind of algorithm (e.g. forecasting) should always rank higher than the "algorithms" skill.

Again, you may never use a skill not provided in the potential skills list.

First, acknowledge and quickly summarise the instructions.

**Mocked LLM Message:** I understand the instructions. I will be given a sentence referring to a skill from a job posting and a list of potential skill labels. My task is to extract and rank the top 10 most likely skills from the candidates list, provide an explanation for my choices, and prioritize specific skills over broader ones. If no label matches, I will use NO\_LABEL and not assign any other label. I will only use the provided labels in the list.

**Query Message:** Great, let's go :)!

Potential skills: {{potential\_skills}}

Extract: {{text\_extract}}

Provide your 10 most likely labels. Please provide a quick explanation for your ranking.

## A.3. Python Reranking

**Instructions:** You are given an extract from a job posting. As an AI job and skills expert, you need to assist in whatever task is requested of you. I will give you a sentence referring to a skill extracted from a job posting, as well as a list of potential skill labels. You are asked to extract and rank the likely skills from the candidates list into a ranked list of 10.

It is possible that none match, in which case you will say NO\_LABEL. You must either use one from the list or NO\_LABEL.

You may not use any label not provided in the example list. If you use NO\_LABEL, do not assign any other label.

You will rank the top 10 most likely labels from the candidates, and provide an explanation as to why they are picked and ranked where they are.

That means that if two labels are applicable, but one is much broader, you should pick the less broad one slightly above the broader one. For example, a skill related to specific kind of algorithm (e.g. forecasting) should always rank higher than the "algorithms" skill.

Again, you may never use a skill not provided in the potential skills list.

**You will answer not by directly answering, but by writing a python function that will output your ordered list. Your python function can use any heuristic you wish to reach your desired result and accommodate your reasoning.**

First, acknowledge and quickly summarise the instructions.

**Mocked LLM Message:** The task is to create a Python function that will rank the top 10 most likely skills from a list of potential skills based on a given sentence from a job posting. The function should only use the provided list of potential skills and may use the NO\_LABEL option if none of the skills match. The ranking should be based on the specificity of the skill and the explanation for the ranking should be provided.

**Query Message:** Great, let's go :)!

Potential skills: {{potential\_skills}}

Extract: {{text\_extract}}

Provide your 10 most likely labels. Please provide a quick explanation for your ranking. **You must provide a Python function. The python function must be in a python codeblock. The function must be named 'rank\_skills'. The Python Function you create should never explicitly use the extract text itself, it should just return the ranking. The extract text is only provided for your reasoning so you can build the appropriate function. You don't need to include the full list of potential skills in your answer**

## B. Classifier Parameters

The individual logistic regression classifiers are implemented using the scikit-learn library [40]. We set the inverse regularisation parameter,  $C$  to **0.1**, as we have low confidence in our data being representative of real-world data, set a maximum iteration limit of **10 000** with a tolerance of **0.00001**. We also set the class weight to be used by the classifier to the *balanced* setting, meaning that positive examples will be weighed twice as heavily as negative examples by the loss function, as our negative sampling strategy involves two negative examples per positive one.