Improving Occupational ISCO Classification of Multilingual Swiss Job Postings with LLM-Refined Training Data

Anonymous ACL submission

Abstract

Classifying occupations in multilingual job postings is challenging due to label noise, language variation, and domain-specific terminology. We propose an approach that refines existing silver-standard job labels using large language model (LLM) assessments and integrates them into Multiple Negatives Ranking (MNR) training for SBERT-based ISCO classification. Our method improves classification accuracy across languages while retaining partial ontology alignment. Experimental results show that LLM-assisted curation enhances training data quality, increasing Top-1 accuracy by over 20 percentage points on job postings. Additionally, multilingual performance benefits from positive cross-lingual transfer, with substantial gains in French and Italian. While fine-tuning leads to a slight drop in ontology-specific accuracy, the overall alignment between job ads and occupational classifications improves. Our findings highlight the potential of LLM-guided refinement for enhancing occupation classification in multilingual labor market data.

1 Introduction

A job is “a set of tasks and duties performed, or meant to be performed, by one person”, and an occupation is “a set of jobs whose main tasks and duties are characterized by a high degree of similarity” (International Labour Organization, 2023). The job title in job postings often reveals the occupation (e.g., “Tax Advisor”), but vague titles (e.g., “Associate”) require additional context.

Educational requirements shape how job ads convey occupations, from broad (e.g., “a humanities degree”) to specific, regulated qualifications (e.g., “GrafikerIn EFZ (Graphic Designer with Swiss Federal Diploma),” “Rechtsanwalt (Lawyer with Bar Admission)”). In Switzerland’s skill-focused market, these formal qualifications often map directly to distinct roles. Some ads allow multiple backgrounds, such as “Fein-, Mikro-, oder Polymechaniker” (Precision, Micro, or Polymechanic), underscoring how tasks and qualifications together define an occupation.

Our web-scraped dataset of 4.7 million Swiss job ads since 2001 shows 80% are German, 11% French, 8% English, and under 1% Italian. Many postings intermix these languages, especially in job titles, and machine translation struggles with terms (e.g., “Compliance Officer”) lacking local equivalents. Classification must thus handle these multilingual code-switching phenomena.

Occupation is vital in labor market research, informing most comparative studies and statistical analyses of job ads. Our aim is to extract occupation-relevant content from these ads and map it to a system suited to the Swiss labor market, while preserving international comparability.

CH-ISCO-19 is a Swiss adaptation of the International Standard Classification of Occupations (ISCO) that organizes roles into five levels with 670 detailed classes. While it aligns with ISCO at the first four levels, the fifth level extends coverage for Switzerland’s labor market. Official labels in German, French, and Italian exist at all levels, but English coverage remains limited, especially for finer distinctions. This structure ensures both Swiss-specific granularity and international compatibility, making CH-ISCO-19 a natural target for classifying job ads in Switzerland.

Mapping job ads to the CH-ISCO-19 taxonomy is challenging due to noise in existing labeled data, complex role descriptions, and multilingual content. Our approach combines multilingual embedding adaptation with suitable multilingual ontologies, then employs large language models (LLMs) to refine noisy, silver-standard training data. We first learn semantic similarities between in-domain texts and standardized occupation labels via Masked Language Modeling (MLM) and Multiple Negatives Ranking (MNR). This alignment caters to

German CH-ISCO-19 classes but supports multiple languages. Next, we leverage LLM-based validation to consolidate high-confidence examples into a better dataset. This pipeline preserves Swiss-specific detail and broader ISCO compatibility, while handling the multilingual complexity of job ads.

Contributions (1) We demonstrate an LLM-assisted approach for rating ISCO occupation candidates, exploring prompt variations and in-context examples, (2) we refine MNR-based training by incorporating GPT-rated suggestions as positives or negatives, and (3) we provide a systematic analysis of data-selection strategies and update schemes, examining model retention, multilingual performance, and the trade-offs of incremental refinement.

2 Related Work

Most occupation-classification pipelines rely on *titles* or *rule-based* heuristics, offering little in the way of robust handling for noisy, large-scale datasets (Swiss Federal Statistical Office (FSO), 2017). Conventional systems often prioritize well-structured ontologies like the ISCO (International Labour Organization, 2012) but do not typically address misaligned or imperfect labels. By contrast, we build on CH-ISCO-19 (Swiss Federal Statistical Office (FSO), 2022), augmenting its standard taxonomy with semantic embeddings and multilingual adaptations. Using the textual richness and expressiveness of large ontologies is key for automatic semantic indexing (Pâslaru-Bontaş, 2007).

Recent advances in *large language models* (LLMs) have expanded annotation capabilities: GPT-style models can generate synthetic data (Magron et al., 2024; Decorte et al., 2023) or re-rank classification candidates (Clavié and Soulié, 2023). We follow this trend by using GPT-based assessments to refine the "silver-standard" SJMM labels, specifically focusing on rating ISCO occupation candidates through in-context examples. This distinguishes our method from purely synthetic data approaches by leveraging existing—albeit noisy—occupation mappings and merging them with LLM-evaluated consistency checks.

For *multilingual model updates*, sentence-level embeddings derived from BERT-style encoders (Reimers and Gurevych, 2019) have proven effective in information retrieval tasks. However, they often require *domain adaptation* (Gururangan et al., 2020), particularly in specialized labor market contexts. Researchers have also introduced knowledge-distillation schemes to extend pretrained models across languages with fewer resources (Reimers and Gurevych, 2020). Building on such efforts, we propose a domain-adapted, multilingual SBERT that retains CH-ISCO-19 alignment while addressing job-ad diversity and language imbalance. We then incorporate an LLM-based curation step, which systematically filters and reweights training data—unlike purely automated or purely rule-based pipelines—thus mitigating label noise and enhancing multilingual performance.

Level	Code Prefix: English Examples
1 (n=10)	2: Professionals; 5: Service and Sales Workers
2 (n=43)	25: Information and Communications Technology Professionals; 52: Sales Workers
3 (n=130)	251: Software and Applications Developers and Analysts; 522: Shop Salespersons
4 (n=582)	2512: Software Developers; 5223: Shop Sales Assistants
5 (n=670)	25122: *Software Developer, Business Informatics*; 52231: *Druggist*

Table 1: Overview of the four ISCO occupation classification levels and our fifth CH-ISCO-19 classification target level (Examples in italics are our translations).

3 Ontologies and Datasets

CH-ISCO-19, provided by the Swiss Federal Statistical Office (FSO)[1], extends the ISCO[2] by adding a fifth level tailored to Switzerland's labor market. Table 1 shows five hierarchical levels spanning 670 classes. Although official CH-ISCO-19 labels exist in German, French, Italian, and partly in English, coverage at the fifth level is limited in English.

The FSO maps over **23k Swiss occupations** to CH-ISCO-19, providing around 45k job titles in German, 41k in French, 39k in Italian, and 6.5k in English. Fewer English titles partly stem from its gender-neutral usage versus other languages' male-female distinctions, though for newer IT or managerial roles (e.g., "DevOps Engineer"), English names are often adopted across languages. Despite this, English remains underrepresented, complicating multilingual tasks and prompting us

[1] Official CH-ISCO-19 documentation is available at https://www.bfs.admin.ch/bfs/en/home/statistics/work-income/nomenclatures/ch-isco-19.html

[2] https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/

to focus primarily on German labels while still accommodating multilingual data in our classification pipeline.

ESCO (European Commission, 2017) extends coverage by providing around 15.5k German/French titles, 16k Italian titles, and 32k English titles, plus English descriptions for all 1–4-digit ISCO classes. Table 6 illustrates examples of the 350-character summaries for 3k occupations, describing typical tasks and responsibilities.

The **Swiss Job Market Monitor (SJMM) 1950-2023 dataset**[3] provides 5-digit CH-ISCO-19 labels for approximately 65k job ads, but these were derived from the older manual SBN2000 classifications (Swiss Federal Statistical Office (FSO), 2017) rather than directly assigned. Unlike the skill-based CH-ISCO-19, SBN2000 groups occupations by economic fields, leading to inconsistencies—for example, classifying doctors and nurses together. Since no direct crosswalk, that is correspondence table, exists, mappings rely on occupational titles, often prioritizing approximate rather than precise matches. This lossy conversion introduces potential misalignment, which is difficult to quantify. Despite these limitations, the dataset serves as a valuable "silver standard" for training and testing, helping to refine classification strategies.

3.1 Data Preprocessing

We extract occupation-relevant information from **job ads** using multilingual text zoning. The primary job title, a key classification feature, is marked with special tags (`[BJT]`, `[EJT]`) for standardization. We also extract key job details—tasks, duties, education, and experience—concatenating scattered fragments with ellipses. Extracted text length varies widely, with English texts averaging 1,800 characters, compared to 600–900 in German, French, and Italian. Table 7 illustrates these differences with examples from English and German job ads.

To align representations, job titles in **ontological data** are marked with the same boundary tags as in job ads. Preprocessing is minimal, limited to removing definitional remarks such as "Not Elsewhere Classified."

4 Domain-Specific SBERT Models

For multilingual occupation classification, we require models that handle job-specific terminology while supporting both German and multilingual inputs. German is central as CH-ISCO-19 class labels, ontological data, and most job ads are in German. However, multilingual capability is essential. To assess the best practice, we test two domain-adapted language models: a German-specific model and a multilingual model.

Multilingual Anchors
[BJT] Kommunikationsplanerin [EJT]
[BJT] Communications planner [EJT]
[BJT] Planificateur en communication [EJT]
[BJT] Pianificatrice di comunicazione [EJT]
You analyse and plan the way a brand is positioned on the market.
Sie analysieren und planen die Positionierung einer Marke auf dem Markt.
Ils analysent et planifient la mise en place d'une marque sur le marché.
Analizzano e pianificano le modalità di immissione sul mercato di un marchio.
[BJT] Medienmanager [EJT]
Positive
[BJT] Fachkräfte in Marketing und Werbung [EJT]

Table 2: Examples of anchor-positive pairs, illustrating how occupation titles and descriptions (anchors) in various languages are mapped to their corresponding German ISCO classes (positive) during the ontological pre-fine-tuning.

We reuse the German *jobGBERT*[4], further pretrained on 2 million spans from job ads, including task descriptions, work activities, and translated O*NET occupation class titles. We refer to this model as *jobLMde*. For multilingual adaptation, we continue MLM training *xlm-roberta-base* (Conneau et al., 2020) using a balanced dataset of 7.8 million job ads, upsampling less frequent languages to ensure robust cross-lingual performance. Training follows best practices from Gururangan et al. (2020). [5] The resulting domain-adapted multilingual model is referred to as *jobLMmulti*.

We fine-tune both models with Multiple Negatives Ranking (MNR) loss (Henderson et al., 2017) to align job titles with their corresponding ISCO classes, simulating the classification task. Given our multilingual focus, we train on job titles and descriptions (CH-ISCO and ESCO data) in four languages, linking them to German CH-ISCO labels. Table 2 shows examples of positive pairs used in training. This directly tunes the mapping of multilingual occupation data (titles and descriptions) to 670 German fifth level classes (*Occ2ISCOde* set-

[3] https://doi.org/10.48573/17e3-0t73

[4] Available on Hugging Face: jobGBERT

[5] 25 epochs, a batch size of 2048, and the Adam optimizer with tuned hyperparameters.

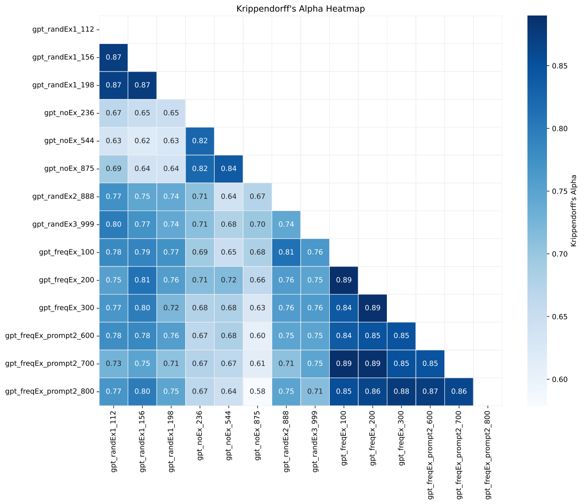

Figure 1: Heatmap of pairwise Krippendorff's α values between GPT runs for rating ISCO classes for job ads in the full test set (n=400).

ting). We also ran MNR adaptation experiments with cross-lingual translation pairs of any available ontological data, but although much more training data could be generated this way, the target task did not improve.

4.1 Results of Ontological Adaptation

We evaluate MNR SBERT adaptation by predicting 5-digit CH-ISCO-19 classes for occupation titles in our ontology. The best-performing *Occ2ISCOde* setting achieves Top-1 accuracy of 91.4% for the German model and 90.9% for the multilingual model. In 99% of cases, the correct class appears among the Top-3 predictions, confirming that MNR effectively aligns multilingual occupation titles with 670 German CH-ISCO labels.

However, when evaluated on SJMM CH-ISCO-19 silver-standard data, all models show disappointingly modest Top-1 performance of around 37%, despite strong ontology alignment and a Top-3 accuracy of 57%. After reviewing discrepancies between Top-1 SBERT similarity suggestions and SJMM silver-standard data, we observed cases where SBERT provided more specific and accurate classifications than the SJMM labels (see Table 8 for examples). To clarify the reasons for this low agreement and to improve the SBERT models in an automated way, we resort to LLM-assisted data validation and refinement.

Condition	**n**	**randEx1**	**randExMix**	**noEx**	**freqEx**	**freqEx_prompt2**
SJMM = SBERT R 1	35	**0.83**	**0.83**	0.74	**0.83**	**0.83**
SJMM Occupation	100	0.65	0.64	0.55	**0.67**	0.66
SBERT R 1	100	**0.51**	0.49	0.46	0.47	0.48
SBERT R 2	100	**0.25**	0.24	**0.25**	**0.25**	**0.25**
SBERT R 3	100	0.16	0.17	**0.20**	0.15	0.16
All Cases	400	**0.30**	0.29	0.28	0.28	0.28

Table 3: Average GPT rating scores for SJMM occupations and top-ranked (R) SBERT candidates across different experimental settings for the cross-lingual test set. The highest scores per condition are indicated in bold. Scores are based on majority votes from three runs per setting. The majority vote for RandExMix is calculated using the runs RandEx2_888, RandEx3_999, and RandEx1_198.

5 LLM-Assisted Data Curation Approach

While our pre-fine-tuned SBERT models perform well on ontology-based tasks, applying them to job ads reveals substantial challenges. A key factor is the silver-standard quality of SJMM data—its occupation labels are uncertain due to conversion issues. To address this, we refine training data using LLM-assisted validation.

5.1 Prompt Engineering for LLM Ratings

How accurate and consistent are LLM ratings across experimental settings? We conducted a prompt engineering feasibility study to assess GPT 4o's[6] ability to rate the quality of CH-ISCO-19 classification candidates. GPT receives SJMM labels and top SBERT suggestions, evaluating their relevance based on structured prompts. Existing research has indicated that LLMs can be effective for re-ranking classification candidates (Magron et al., 2024), particularly when provided with structured input. We therefore experiment with different prompt formats and in-context learning (ICL) examples, as suggested by Dong et al. (2024). Figure 5 shows the ICL system prompt.

Setup. We randomly sample 100 SJMM-labeled job ads (25 per language) and compare SJMM silver-standard occupations with top 1-3 SBERT suggestions from *jobLMde-MNR*. If the SJMM label is among the top-3 SBERT candidates, the top-4th SBERT candidate is added. GPT always rates four blind CH-ISCO occupation candidates per job ad, without access to rankings, similarity scores,

[6] We use the chat API of `https://openai.com`

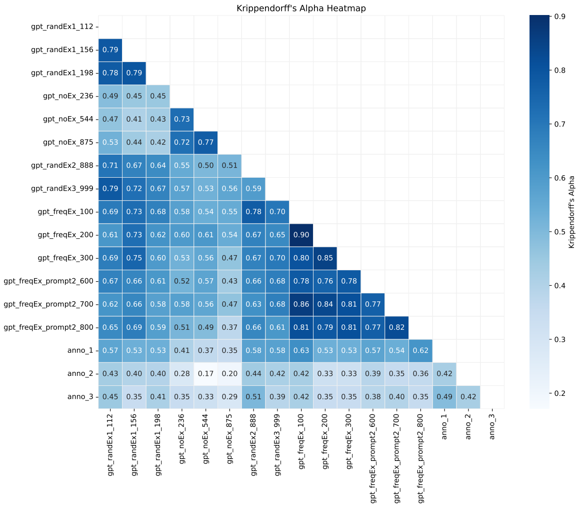

Figure 2: Heatmap of pairwise Krippendorff's α values between GPT runs and human annotators for rating ISCO classes in the challenge set (n=192).

or SJMM labels. Ratings follow a 3-value scheme: 1.0 for exactly matching, 0.5 for partial matches, and 0 for no match. This ensures a controlled setup for assessing GPT's rating quality, reliability, and alignment with human annotations. To evaluate the influence of input information and prompt structure on GPT's performance, we conducted multiple experimental configurations, repeating each setting several times. We performed three runs without examples (*noEx*), with using 5 randomly sampled examples (*randEx1*) and two additional runs with different random examples (*randEx2* and *randEx3*). For the frequency-based examples (*freqEx*), we used string matching to identify the most commonly mentioned occupations within each class in the SJMM dataset. Additionally, we tested the frequency-based setting with an updated prompt (*freqEx prompt2*), also for three runs. This updated prompt included a specific instruction to improve the classification of managerial roles, addressing observed misclassification related to the frequent mention of the word "manager' for smaller team leadership. The adapted prompt is shown in Figure

Each run was conducted using the *gpt-4o* model with a temperature setting of 0.5. To ensure reproducibility, a fixed random seed was used for each run, indicated by the numeric suffix in the run name (e.g., *gpt randEx1 112*).

Results in Figure 1 show high consistency in GPT ratings (Krippendorff's $\alpha > 0.85$) (Krippendorff, 2004) and confirm that in-context examples significantly improve rating consistency.

Table 3 nicely shows high LLM matches when SJMM and SBERT agree. Alignment of GPT with SJMM is better than with SBERT, suggesting a certain silver standard quality. SBERT semantic similarity ranking correlates with GPT average scores.

5.2 Human Annotation of Challenging Cases

How do LLM ratings compare to human annotations in cases with significant discrepancies between SJMM labels and SBERT predictions? In order to answer this question, 3 domain experts received the same randomly selected occupation examples as GPT (the "randEx1" setting) and focused on the "challenge set"—a subset of 48 particularly difficult cases where the SJMM occupation was not among the top three SBERT candidates. These annotations serve as a benchmark for evaluating GPT's rating quality and reliability in cases with significant classification discrepancies.

For this evaluation, we applied an adaptive binarization strategy: if an annotator did not assign a 1.0 rating to any suggestion, 0.5-rated suggestions were considered positive, and all others were treated as negative. When 1.0-rated suggestions were present, these were treated as positives, while the remaining ratings were considered negative. This approach helped balance differences in annotator stringency, accounting for both more lenient and more strict human raters, and allowed us to identify patterns in well-rated and poorly rated suggestions.

Low inter-annotator agreement (IAA) among human raters ($\alpha = 0.42$ on average) reflects the subset's difficulty and the lack of extensive training. GPT ratings show higher internal consistency ($\alpha = 0.64$), though their agreement with humans improves when examples are included ($\alpha = 0.45$).

Human annotators slightly favor SJMM occupations and SBERT's top candidates more than GPT (0.31–0.38 vs. 0.29–0.31; see Figure 2). While GPT evaluations are more stable, they may be slightly conservative, occasionally overlooking valid human-rated matches. Overall, GPT ratings demonstrate reliability and structured consistency, reinforcing their usefulness for refining training data.

	de	en	fr	it	all
Original					
SUFNotTop1	74k / 88%	4k / 5%	5k / 6%	1k / 1%	84k
SJMMNotInTop3	62k / 80%	6k / 8%	8k / 11%	1k / 1%	77k
SJMM&Top3	91k / 82%	7k / 6%	11k / 10%	2k / 2%	111k

Table 4: Distribution of training samples across different data selection scenarios, with raw counts and percentages per language.

6 CH-ISCO-Specific MNR Training Data Curation: Results and Discussion

The feasibility study in the preceding section confirms that GPT ratings provide a reliable means of evaluating CH-ISCO class candidates, demonstrating strong internal consistency and decent alignment with SJMM and SBERT predictions. However, observed disagreements—both among human annotators and between GPT and human ratings—highlight the need for careful curation when incorporating GPT-generated labels into training data.

To generate CH-ISCO-specific MNR training data, we prioritize cases where at least two sources—GPT, SBERT, or SJMM—agree on an occupation class. Job ad titles and descriptions serve as anchors, with validated CH-ISCO classes as positive targets. Hard negatives are selected based on GPT ratings, where CH-ISCO classes rated 0.0 are used directly, while additional negatives are drawn from mismatched SBERT predictions. This approach ensures that training samples emphasize high-confidence mappings, reducing noise and mitigating potential biases.

Using the Transformers library (Wolf et al., 2020) and following Gururangan et al. (2020), we apply **weighted sampling** to optimize updates: high-confidence matches (GPT=1.0 for SJMM or SBERT Top-1) receive a weight of 2 in MNR updates, while moderate matches (GPT=0.5) receive a weight of 1 to balance the training distribution. SJMM/SBERT agreements count as positive targets with weight 1.

6.1 Training Data Strategies for MNR Fine-Tuning

To enhance similarity-based classification accuracy, we fine-tune SBERT models using MNR with carefully curated training data. However, which training data should be used for MNR fine-tuning (and in which order)? What to do about language imbalance?

The training data is derived from the SJMM dataset (65k job ads), where discrepancies exist between the SJMM-assigned ISCO class and the top-ranked ISCO prediction from the best-performing SBERT model (50k). Each case is rated by GPT to determine the reliability of both SJMM and SBERT classifications. The dataset provides a rich multilingual resource, encompassing job postings in German, French, English, and Italian. However, due to language imbalances, adjustments are applied to maintain representativeness.

Model	Top-1	Top-2	Top-3
jobLMde-MNR	37.2	49.4	56.7
+ SJMMNotInTop3	49.8	64.0	70.5
+ SJMMNotTop1	51.8	65.1	71.5
+ SJMM&Top3	53.6	66.3	72.4
+ SJMM&Top3Large	58.3	70.6	76.3

Table 5: Top1–3 accuracy (%) in ISCO-5digit classification of model *jobLMde-MNR*, evaluated on SJMM data from 1990 onwards (n=65k), before and after initial updates with three different training data settings.

We explore three different training data configurations, each targeting varying degrees of agreement between SJMM and SBERT predictions:

SJMMNotInTop3 (43% of GPT-rated dataset) focuses on cases where the SJMM-assigned CH-ISCO class does not appear among the top three SBERT predictions. These cases highlight the strongest classification discrepancies. GPT assigns an average rating of 0.47 to SJMM occupations, aligning with the pretest rating (0.46). However, GPT assigns higher ratings to SBERT's top-ranked suggestions in this dataset (0.58 vs. 0.31 in pretests), indicating a greater concentration of frequent, well-defined occupations. This configuration produces approximately 77k training triplets.

SJMMNotTop1 (64% of GPT-rated dataset) covers all cases where the SJMM label is not the top-ranked SBERT prediction. GPT aligns more strongly with SJMM labels in this set, assigning them an average rating of 0.62, compared to 0.56 for SBERT's top candidate. This suggests that SJMM labels still provide useful occupation-specific information beyond what SBERT has learned from ontological pre-fine-tuning. This configuration generates approximately 84k training triplets.

SJMM&Top3 (100% of dataset) is the most comprehensive dataset, including all cases where GPT evaluated ISCO candidates. This configuration integrates cases where SJMM is not the top SBERT suggestion (*SJMMNotTop1*) alongside

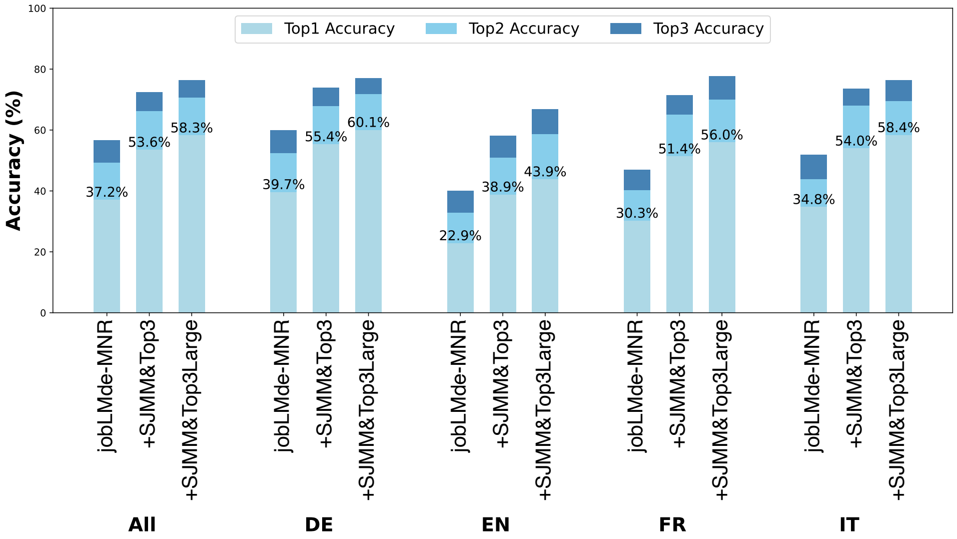

Figure 3: Stacked Top1–3 accuracy (%) in ISCO-5digit classification of *jobLMde-MNR*, before and after updates with two different training data settings, evaluated on SJMM data from 1990 onwards (n=65k). Top-1 accuracy is displayed on the bars.

cases where SJMM and SBERT's top-ranked prediction align (36% of the dataset). By capturing both agreement and discrepancy cases, this setting yields a total of 111k training triplets.

The results in Table 3 demonstrate that all MNR training configurations with curated job ad data improve the *jobLMde-MNR* model, which was initially fine-tuned on ontological data. Notably, larger datasets yield stronger improvements, and incorporating both positively and negatively rated SBERT predictions further enhances performance. The MNR fine-tuning process not only increases the semantic similarity scores for the top-ranked candidate but also refines overall ranking quality, ensuring better alignment between job ad descriptions and CH-ISCO classifications.

Language Balancing. As shown in Table 4, German dominates all datasets. Initial experiments with a balanced sampling strategy—downsampling German (0.25) and upsampling English and French (×2) and Italian (×4)—reduced German performance (see Table 5). To counteract this, we applied a more aggressive upsampling strategy, retaining all German samples while increasing English and French by a factor of 8 and Italian by 4 to prevent overfitting due to limited data.

The optimal training dataset, *SJMM&Top3Large*, consists of 232k samples, distributed as 38% German, 36% French, 23% English, and 3% Italian. Notably, Figure 3 shows that all languages benefit from positive cross-lingual transfer. While German, the dominant language in training, achieves the highest Top-1 accuracy (58.3%), French (56.0%) and Italian (58.4%) follow closely. Only English remains a challenge, with a substantially lower Top-1 accuracy of 43.9%.

6.2 SBERT Similarity and GPT Rating

How does GPT-based data curation affect cosine similarities? The impact on SBERT top candidates is also reflected in their similarity scores. Figure 4 presents violin plots illustrating the relationship between similarity scores and GPT ratings before and after the update.

Post-update, similarity scores for top candidates rated as 1.0 shift upward, ensuring that candidates with scores above 0.9 no longer receive a 0 rating. Conversely, candidates with scores below 0.6 are now more consistently identified as incorrect. This refinement not only improves classification accuracy but also establishes clearer similarity thresholds, enhancing the interpretability of predictions.

6.3 Ontology Retention.

Does the improvement in CH-ISCO job ad classification also enhance performance on the original pre-fine-tuning ontology-based CH-ISCO classification task?

Table 7 provides a clear answer. Top-1 accuracy on ontology data decreases by approximately 10 percentage points after fine-tuning, while Top-3 accuracy remains high at 95%. This decline reflects

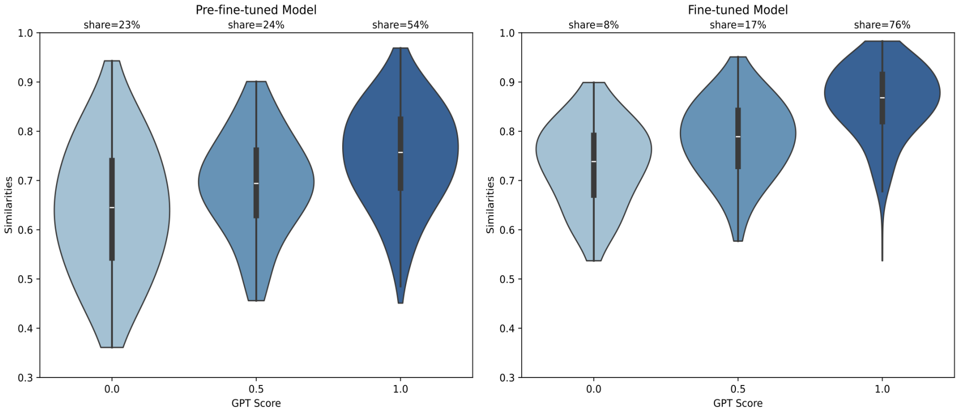

Figure 4: Violin plots showing the distribution of similarity scores for the SBERT top candidate, categorized by GPT rating, across test sets for both pre-fine-tuned and SJMM&Top3Large-updated *jobLMde-MNR* models (n=775). Each violin plot includes an embedded box plot, and the distribution of GPT ratings is indicated at the top of each plot.

a data shift between job ads and ontological data. Qualitative analysis suggests that many misclassifications are non-critical, with English occupation assignments frequently preferring 4-digit classes. This may stem from the initial scarcity of English data at CH-ISCO level 5.

6.4 Results: Evaluation on Held-Out Test Sets

To assess the model's generalization to unseen job ads, we evaluate *jobLMde-MNR* on a held-out test set, leveraging both SJMM data and GPT-based external validation. This analysis focuses on the best-performing *SJMM&Top3Large* fine-tuning.

GPT-based ratings on the human-validated random test set (100 job ads, 25 per language) confirm improved agreement for top-ranked SBERT suggestions, demonstrating the effectiveness of the fine-tuning approach. The *SJMM&Top3Large* update significantly enhances GPT alignment with top SBERT suggestions. As shown in Table 9, the proportion of top-ranked suggestions rated as acceptable or perfect matches (≥ 0.5) reaches 95% for German (from 84%), 92% for French (from 77%), 91% for English (from 76%), and 81% for Italian (from 73%). These results confirm substantial consistency gains through the update. At the same time, the ads with silver standard SJMM occupation classes on Top-1 SBERT rank raise from 43% to 60% for German, from 30% to 47% for French, from 37% to 47% for Italian, from 32% to 45% for English. Further manual validation is needed to assess actual issues in the silver data.

Overall, these results demonstrate the *SJMM&Top3Large* update's effectiveness in refining SBERT's top-ranked predictions, increasing GPT validation scores, and improving alignment with task-specific classification needs.

6.5 Multiple Rounds of LLM Refinement

In a last experiment, we tested whether a second round of GPT annotations on newly misaligned cases improves performance further (Clavié and Soulié, 2023). When performing incremental updates, performance degraded slightly. Small additional gains only occur when combining new data into a single large update. Sequential incremental updates can degrade generalization, suggesting it is more effective to do a comprehensive retraining with all curated data at once.

7 Conclusion and Future Work

We demonstrate that LLM-assisted data curation can effectively refine noisy SJMM occupation labels and significantly improve SBERT-based classification of multilingual job postings. While some ontology knowledge is attenuated by the task-specific updates, the overall classification accuracy on job ads improves substantially.

Future Directions. To further enhance multilingual occupation classification, future research should focus on (1) obtaining more reliable human annotations for critical occupations, (2) refining English labels to reduce granularity mismatches, and (3) exploring advanced fine-tuning strategies, such as iterative domain adaptation or synthetic job ad generation, to better retain ontology alignment while improving task performance.

Limitations

While our approach significantly improves ISCO classification of multilingual job postings, several limitations remain.

Dependence on Silver-Standard Data. The refinement process relies on silver-standard job labels from the Swiss Job Market Monitor (SJMM), which may contain systematic biases due to legacy classification schemes. Although LLM-based validation helps mitigate inconsistencies, it cannot fully resolve underlying errors in the original data.

Ontology Drift. Fine-tuning on job postings shifts the model away from ontology-based classification, reducing accuracy on structured occupational taxonomies. While Top-3 accuracy remains high (95%), the decline in Top-1 performance suggests that additional strategies are needed to balance task-specific adaptation with ontology retention.

Cross-Lingual Performance Gaps. Despite gains from multilingual fine-tuning, performance remains uneven across languages. English shows lower classification accuracy compared to German and French, likely due to limited training data and fewer fine-grained occupational labels at CH-ISCO level 5. Further adjustments to training data distribution may be necessary to close this gap.

LLM Annotation Stability. Although GPT-based rating is internally consistent, agreement with human annotators varies, particularly for ambiguous occupations. The system's reliance on in-context learning means that results may be sensitive to prompt variations, requiring careful tuning to ensure stable outputs across different rating tasks.

Computational Cost. Our approach requires multiple rounds of LLM annotation using a commercial API. The experiments could be extended to include strong open-source LLMs.

Lack of Independent Human Evaluation. While domain experts annotated a challenge set to benchmark GPT ratings, the broader evaluation lacks a fully independent human assessment of model outputs. A larger-scale human annotation effort would provide a more definitive validation of classification accuracy.

Future work should address these challenges by incorporating human-in-the-loop validation, exploring more efficient update strategies.

References

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) co-located with the 17th ACM Conference on Recommender Systems (RecSys 2023), Singapore, Singapore, 18th-22nd September 2023*, volume 3490 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2307.10778*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

European Commission. 2017. *ESCO handbook – European skills, competences, qualifications and occupations*. Publications Office.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient Natural Language Response Suggestion for Smart Reply. *arXiv preprint*. ArXiv:1705.00652 [cs].

International Labour Organization. 2012. *International Standard Classification of Occupations: ISCO-08: Structure, group definitions and correspondence tables*. International Labour Office, Geneva, Switzerland.

International Labour Organization. 2023. Classification of occupations — concepts and definitions. URL: `https://ilostat.ilo.`

org/methods/concepts-and-definitions/classification-occupation/. Accessed: 2024-12-17.

Klaus Krippendorff. 2004. Reliability in Content Analysis.: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433.

Antoine Magron, Anna Dai, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. 2024. JOBSKAPE: A Framework for Generating Synthetic Job Postings to Enhance Skill Matching.

Elena Pâslaru-Bontaş. 2007. *A contextual approach to ontology reuse: Methodology, methods, and tools for the semantic web*. Ph.D. thesis, Freie Universität Berlin, Berlin, Germany. URL: https://refubium.fu-berlin.de/handle/fub188/8347. Accessed: 2025-01-29.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Swiss Federal Statistical Office (FSO). 2017. Schweizer Berufsnomenklatur 2000 - SBN 2000. URL: https://www.bfs.admin.ch/bfs/de/home/statistiken/arbeit-erwerb/nomenclaturen/sbn2000.assetdetail.4082532.html. Accessed: 2024-12-17.

Swiss Federal Statistical Office (FSO). 2022. Schweizerische Berufsnomenklatur: CH-ISCO-19. URL: https://www.bfs.admin.ch/bfs/en/home/statistics/work-income/nomenclatures/ch-isco-19.html. Accessed: 2024-12-12.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

Lg.	Occupation	Description
de	Controllerin	Controller/Controllerinnen prüfen und analysieren Jahresabschlüsse, Budgets, Finanzberichte und Geschäftspläne auf fehlerbedingte oder betrügerische Unregelmäßigkeiten und beraten ihre Kunden zu Themen wie Finanzprognose und Risikoanalyse. Sie können Finanzdaten prüfen, Insolvenzfälle verwalten, Steuererklärungen erstellen und in Bezug auf die geltenden Rechtsvorschriften weitere Steuerberatung anbieten.
en	Software tester	Software testers perform software tests. They may also plan and design them. They may also debug and repair software although this mainly corresponds to designers and developers. They ensure that applications function properly before delivering them to internal and external clients.
fr	Médecin généraliste	Les médecins généralistes œuvrent à la promotion de la santé, à la prévention, décèlent des pathologies, posent des diagnostics et traitent des maladies; ils encouragent la guérison de maladies physiques et psychiques et de troubles de la santé de toutes sortes pour toutes les personnes, indépendamment de leur âge, de leur sexe ou du type de problème de santé qui les affecte.
it	Meccanica riparatore di motori diesel	I meccanici riparatori di motori diesel curano la riparazione e manutenzione di tutti i tipi di motori diesel. Utilizzano attrezzi manuali, strumenti di misurazione di precisione e macchine utensili per individuare guasti, problemi, smontare i motori, nonché esaminare e cambiare parti difettose ed eccessivamente usurate.

Table 6: Examples of occupations from ESCO in German (de), English (en), French (fr), and Italian (it) with their corresponding descriptions.

[BJT] Digital Channel Manager (m/f) [EJT] Responsibilities: - Manage the website from requirements to the realization. - Gather feedback from internal and external stakeholders. - Support and assure correct implementation within the ecosystem, including testing, monitoring and training. - Preparation, organization and implementation of testing periods. - Introduction and assurance of success monitoring for all digital touchpoints. ... Technical background and minimum 3-5 years of work experience with solid understanding of IT, specifically content management and tracking systems ... Experience in project management and profound perception of how IT applications and business processes are linked. Fluent in German and English. ... Flexible working time model.
[BJT] Lebensmittelverkäufer [EJT] möglichst mit Lehrabschlussprüfung. *[BJT] Food salesperson [EJT] Preferably with a vocational diploma.*

Table 7: Example of extracted and preprocessed occupation information from an English and a German job ad.

Input Text	Top-1 Suggestion	SJMM Classification
[BJT] Chief Architect / Technology Partner [EJT] Being part of the pan-European technology office of Banking and Financial Services industry practice, you will also have responsibility to collaborate with fellow architects and solution SMEs to ...	25111 Systemanalytiker, Architektur und Controlling (*System analyst, architecture and controlling*)	21610 Architekten (*Building Architects*)
[BJT] Stage di pratica in cure (studente medicina) [EJT] Lo stage di 4 settimane è richiesto da alcune facoltà di medicina della Svizzera ... (*Internship in care (medical student). The 4-week internship is required by certain medical faculties in Switzerland...*)	32560 Medizinische Assistenten (*Medical Assistants*)	22100 Ärzte, onA (*Medical Doctors, unspecified*)

Table 8: Shortened and translated examples of job ads with top-1 predictions from model *jobLMde-MNR* and corresponding SJMM classifications.

	Random sample							
	de		**en**		**fr**		**it**	
	orig	**updated**	**orig**	**updated**	**orig**	**updated**	**orig**	**updated**
% ads with GPT rating $\geq$ 0.5:								
in Rank 1	84%	95%	76%	91%	77%	92%	73%	81%
in Ranks 1-2	94%	98%	89%	98%	87%	96%	82%	93%
in Ranks 1-3	96%	99%	94%	98%	93%	99%	91%	97%
% ads with GPT rating = 1:								
in Rank 1	62%	83%	42%	68%	46%	67%	56%	67%
in Ranks 1-2	74%	90%	61%	79%	62%	79%	71%	83%
in Ranks 1-3	87%	95%	73%	92%	71%	92%	80%	90%
% ads with SJMM Occupation:								
in Rank 1	43%	60%	32%	45%	30%	47%	37%	47%
in Ranks 1-2	58%	71%	43%	57%	39%	64%	46%	58%
in Ranks 1-3	66%	82%	47%	66%	43%	74%	51%	65%
Mean Ratings:								
SJMM Occupation	0.75	0.78	0.70	0.66	0.77	0.73	0.70	0.68
SBERT Rank 1	0.73	0.89	0.59	0.80	0.62	0.79	0.64	0.74
SBERT Rank 2	0.38	0.48	0.42	0.45	0.34	0.43	0.36	0.42
SBERT Rank 3	0.36	0.34	0.32	0.40	0.26	0.33	0.24	0.28
Overall Mean Rating	0.44	0.48	0.44	0.48	0.42	0.44	0.40	0.43

Table 9: Performance of pre-fine-tuned and SJMM&Top3Large-updated *jobLMde-MNR* across random held-out test sets for German (de), English (en), French (fr), and Italian (it). The table shows percentages of job ads with 0.5- or 1-rated GPT suggestions in ranks 0–3, and SJMM occupations in ranks 0–3, along with mean rating scores for SJMM occupations and SBERT candidates. The random set includes 25 ads per language languages, with four suggestions per ad.

```
You are a specialist in assigning Swiss-
    specific CHISCO codes to job
    postings from Switzerland. CHISCO is
     an extension of the International
    Standard Classification of
    Occupations (ISCO), specific to
    Switzerland. Each class has a
    numerical identifier and a textual
    label (e.g., "12110 Führungskräfte
    im Bereich Finanzen").The first 4
    digits correspond to the ISCO code,
    and the 5th digit represents the
    Swiss-specific extension (0
    indicates a direct match to the ISCO
     code).

Task:
Evaluate the suggested CHISCO class
    candidates based on the job posting
    text provided.

Input:
A job ad (in German, French, English, or
     Italian) and a list of CHISCO class
     candidates, each with its code,
    label, and example occupational
    titles.

Evaluation:
Rate each CHISCO candidate based on
    relevance to the job ad:
- 1: The CHISCO candidate matches the
    job ad very well.
- 0.5: The CHISCO candidate somewhat
    matches the job ad.
- 0: The CHISCO candidate does not match
     the job ad at all.

Consider for your evaluation:
- More than one candidate can be
    partially or fully correct.

...
```

Figure 5: Excerpt of the system prompt for GPT's CH-ISCO-19 evaluation task in example-based settings, focusing on content instructions while omitting technical formatting details.

```
Consider for your evaluation:
- More than one candidate can be
    partially or fully correct.
- CHISCO candidates that start with '1'
    (e.g., 12212 Führungskräfte in
    Marketing) are reserved for
    executives and managers with
    significant decision-making
    authority and responsibility for an
    organizational unit (strategic,
    financial, or staffing), requiring a
     high skill level. Supervisors or
    positions involving team leadership
    do not qualify per se for these
    categories.
```

Figure 6: Expanded instruction for handling managerial positions in the system prompt for GPT ratings (*freqEx-promp2*).

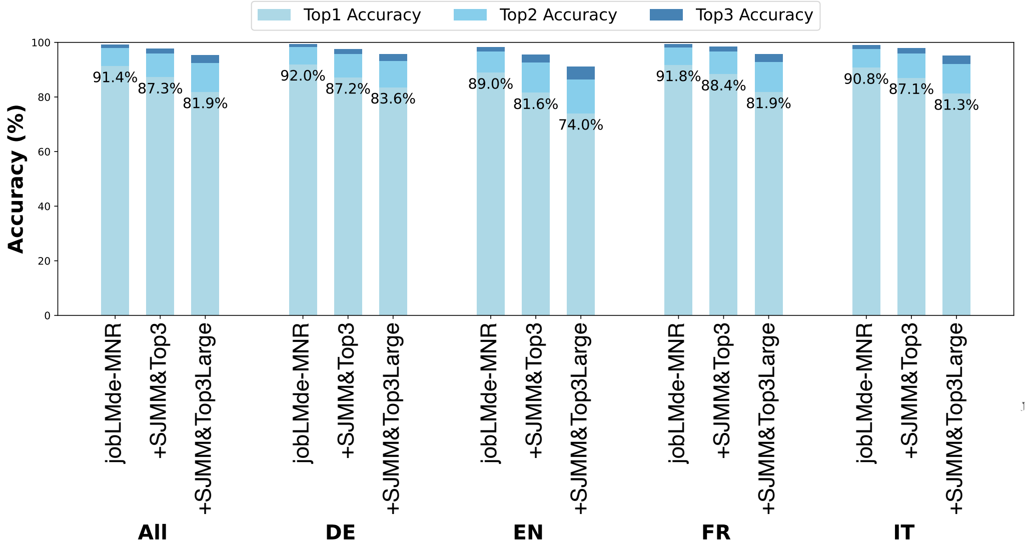

Figure 7: Top1–3 accuracy (%) on ontology evaluation data (n=108k) in ISCO-5digit classification of *jobLMde-MNR*, before and after updates with two different training data settings. Top1 accuracy percentages are displayed on the bars.