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

Ann-Sophie Gnehm
Department of Sociology
University of Zurich
gnehm@soziologie.uzh.ch

Simon Clematide
Department of Computational Linguistics
University of Zurich
simon.clematide@uzh.ch

Abstract

Classifying occupations in multilingual job postings is challenging due to noisy labels, language variation, and domain-specific terminology. We present a method that refines silver-standard ISCO labels by consolidating them with predictions from pre-fine-tuned models, using large language model (LLM) evaluations to resolve discrepancies. The refined labels are used in Multiple Negatives Ranking (MNR) training for SentenceBERT-based classification. This approach substantially improves performance, raising Top-1 accuracy on silver data from 37.2% to 58.3% and reaching up to 80% precision on held-out data—an over 30-point gain validated by both GPT and human raters. The model benefits from cross-lingual transfer, with particularly strong gains in French and Italian. These results demonstrate hat LLM-guided label refinement can substantially improve multilingual occupation classification in fine-grained taxonomies such as CH-ISCO with 670 classes.

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 education-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 dataset of 4.7 million Swiss job ads shows 80% are German, 11% French, 8% English, and under 1% Italian. Many postings intermix these languages, especially in job titles, and commercial machine translation services often left terms untranslated (e.g., "Compliance Officer"), or translated them to unusual German terms. 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 (hereafter CH-ISCO) 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 a natural target for classifying job ads in Switzerland.

To address the noise in existing labeled data, complex role descriptions, and multilingual content, our approach combines multilingual embedding adaptation with structured ontologies and uses LLMs to refine silver-standard training data. We first learn semantic similarities between in-domain texts and occupation labels via Masked Language Modeling (MLM) and MNR, optimized for German CH-ISCO classes but supporting multiple languages. LLM-based validation then consolidates high-confidence examples into a cleaner dataset.

Findings of the Association for Computational Linguistics: ACL 2025, pages 21834–21847

This pipeline preserves Swiss-specific detail and ISCO compatibility while addressing the multilingual complexity of job ads.

This paper makes the following key **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. (3) We provide an 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 *job titles*, offering limited robustness for noisy job ads. Traditional systems often emphasize structured taxonomies such as ISCO (International Labour Organization, 2012), yet typically overlook issues of label misalignment or noise (Deniz et al., 2024; Retyk et al., 2024). We build on CH-ISCO (Swiss Federal Statistical Office (FSO), 2022), augmenting its taxonomy with multilingual embeddings to better handle real-world classification challenges.

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 (Buchmann et al., 2024), specifically focusing on rating occupation candidates through in-context examples. This distinguishes our method from purely synthetic data approaches by leveraging existing—albeit noisy—labels 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 SentenceBERT (SBERT) that retains CH-ISCO alignment while addressing job-ad diversity and language imbalance. We then incorporate an LLM-based curation step, which filters and reweights training data, 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 classification target level. Examples in italics are our translations.

3 Ontologies and Datasets

CH-ISCO, 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 labels exist in German, French, Italian, and partly in English, coverage at the fifth level is very limited in English.

The FSO maps over **23k Swiss occupations** to CH-ISCO, 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. For newer IT or managerial roles (e.g., "DevOps Engineer"), English names are often adopted across languages. However, English remains underrepresented, prompting us to focus on German labels while continuing to support multilingual input in the classification pipeline.

ESCO (European Commission, 2017) extends coverage by providing around 15.5k German and French titles, 16k Italian titles, and 32k English titles, plus English descriptions for all 1-4-digit ISCO classes. Table 8 illustrates examples of the 350-character summaries for 3k occupations, describing typical tasks and responsibilities.

The **Swiss Job Market Monitor (SJMM)**[3] provides for the period 1990-2023 approximately 65k

[1] Official CH-ISCO documentation available at the FSO.
[2] See ISCO documentation at ILOstat.ilo.org.
[3] https://doi.org/10.48573/17e3-0t73

job ads with 5-digit CH-ISCO labels. However, these labels were derived from the older manual SBN2000 classifications (Swiss Federal Statistical Office (FSO), 2017) rather than being directly annotated according to CH-ISCO. Unlike the skill-based CH-ISCO, SBN2000 groups occupations by economic fields, e.g., classifying doctors and nurses together under healthcare, whereas CH-ISCO separates them by skill requirements. Since no direct crosswalk exists, mappings rely on occupational titles, often assigned with limited precision, focusing more on selecting a fitting SBN2000 class than ensuring an exact match. This conversion introduces misalignment that is difficult to quantify. Despite this limitation, the data remains a valuable "silver standard" for training and evaluation.

The CH-ISCO class distribution in the SUF data is highly diverse: 498 of 670 possible classes (74.3%) are represented, with a relative entropy of 81% (MacKay, 2003). This broad spread highlights the complexity of the classification task.

3.1 Data Preprocessing

We extract occupation-relevant information from **job ads** using a text zoning model. Trained in-house on a manually annotated dataset, the transformer-based model segments job ad text at the token level into zones such as skills, company information, and job descriptions (Gnehm et al., 2022). The primary job title, a key input for classification, is marked using special tags (`[BJT]`, `[EJT]`). We also extract additional job details—tasks, duties, education, and experience—by concatenating relevant fragments with ellipses. Extracted text length varies considerably across languages: English job ads average 1,800 characters, while German, French, and Italian range between 600 and 900. Table 9 illustrates these language-specific differences with examples.

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

Multilingual occupation classification requires models that handle job-specific terminology while supporting both German and multilingual inputs. German is key, as most ontological and job ad data are in this language, but multilingual capability is essential. To identify best practices, we compare two domain-adapted language models: one German-specific and one multilingual.

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* (Gnehm et al., 2022)[4], further pre-trained on 2 million spans from job ads, including task descriptions, work activities, and translated O*NET[5] 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 the Transformers library (Wolf et al., 2020) on 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).[6] The resulting domain-adapted multilingual model is referred to as *jobLMmulti*. We focus on these models as prior experiments showed that SBERT fine-tuning on domain-adapted language models consistently outperforms fine-tuning of general-purpose SBERT variants.

We fine-tune both models using the Sentence-BERT library (Reimers and Gurevych, 2019) with 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

[4]Available on Hugging Face: jobGBERT

[5]https://www.onetonline.org

[6]25 epochs, a batch size of 2048, and the Adam optimizer with tuned hyperparameters.

directly tunes the mapping of multilingual occupation data (titles and descriptions) to 670 German fifth level classes (*Occ2ISCOde* setting).

We also conducted MNR adaptation experiments using cross-lingual translation pairs from the available ontological data. While this approach substantially increased the amount of training data, it did not yield improvements on the target classification task.

4.1 Results of Ontological Adaptation

We evaluate MNR adaptation by predicting CH-ISCO classes for occupation titles drawn from the ontology. The best-performing *Occ2ISCOde* setting achieves a Top-1 accuracy of 91.4% for the German and 90.9% for the multilingual model. In 99% of cases, the correct class is ranked within the Top-3, indicating that MNR fine-tuning effectively aligns multilingual occupation titles with 670 German CH-ISCO labels.

However, when evaluated on SJMM job ads with CH-ISCO silver-standard labels, all models exhibit relatively modest Top-1 performance of around 37%, despite strong alignment with the ontology and a Top-3 accuracy of 57%. Manual inspection revealed cases where SBERT predictions were more specific and contextually appropriate than the silver-standard labels (see Table 10). To address this discrepancy and improve model quality at scale, we introduce LLM-assisted data validation and refinement.

5 LLM-Assisted Data Curation Approach

This approach targets the limitations of silver-standard SJMM labels, which—due to conversion issues—introduce noise that affects both training and evaluation. By leveraging LLM-based validation, we identify and resolve label discrepancies to enhance the reliability of training data.

5.1 Prompt Engineering for LLM Ratings

How accurate and consistent are LLM ratings across experimental settings?

We conducted a feasibility study to assess GPT-4o's[7] ability to rate the quality of CH-ISCO classification candidates. GPT receives SJMM labels and top SBERT suggestions, evaluating their relevance using structured prompts. Prior work has shown that LLMs can effectively re-rank classification candidates (Magron et al., 2024), especially

[7]Using the chat completion API of `https://openai.com`

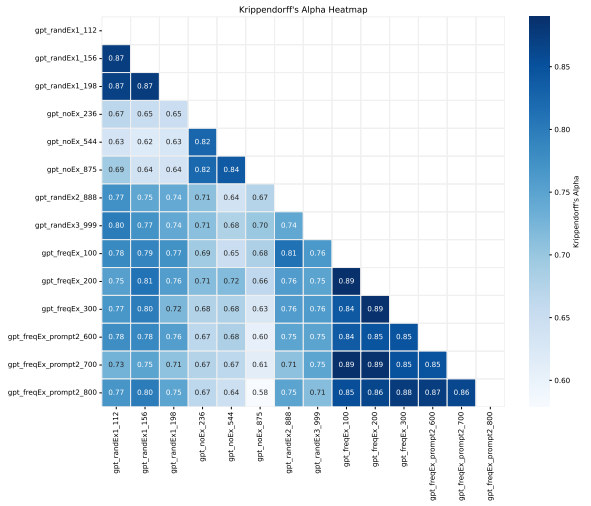

Figure 1: Heatmap of pairwise Krippendorff's α values between GPT runs for ISCO ratings on the small random test set (n=400 suggestions across 100 ads).

with structured input. We therefore test different prompt formats and in-context learning (ICL) examples, following suggestions by Dong et al. (2024). Figure 4 in the Appendix shows the ICL system prompt.

Setup: We randomly sample 100 SJMM-labeled job ads (25 per language) to create the *small random test set*, comparing SJMM occupations with the top 1–3 SBERT suggestions from *jobLMde-MNR*. If the SJMM label appears in the top three, the fourth SBERT candidate is added. GPT then rates these four blind CH-ISCO candidates per ad—without access to rankings, similarity scores, or SJMM labels—using a 3-point scale: 1.0 for exact matches, 0.5 for partial, and 0 for no match. This setup evaluates GPT's rating reliability and alignment with human ratings.

To evaluate the impact of input and prompt structure on GPT's performance, we ran several experimental configurations. These included three runs without examples (*noEx*), three runs with the same five random examples (*randEx1*), and two runs with different examples (*randEx2* and *randEx3*). For the frequency-based setting (*freqEx*), also tested in three runs, we used string matching to identify the most common occupations per class in the SJMM data. We also tested an updated version (*freqEx prompt2*) with a specific instruction to better classify managerial roles, addressing misclassifications linked to generic use of "manager." See Figure 5 in the Appendix for the prompt adaptation.

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 SBERT candidates across experimental settings on the small random test set (400 suggestions across 100 ads). R denotes the SBERT candidate's rank based on cosine similarity (e.g., R 1 = top-ranked suggestion, R 2 = second-ranked, etc.). GPT ratings follow a 3-point scale: 1 = exact match, 0.5 = partial match, 0 = no match. Scores reflect majority votes from three GPT runs. For RandExMix condition, the vote is based on RandEx2_888, RandEx3_999, and RandEx1_198. Highest scores per condition are shown in bold.

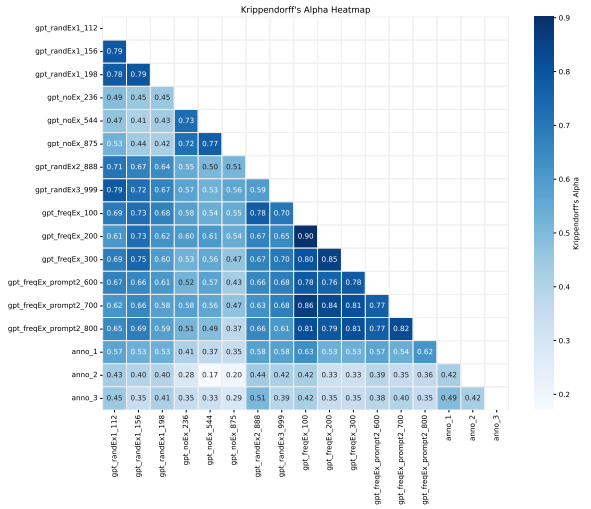

Figure 2: Heatmap of pairwise Krippendorff's α values between GPT runs and human raters for ISCO ratings on the challenge subset (192 suggestions across 48 ads).

Each run used the *gpt-4o* model with a temperature of 0.5. A fixed random seed ensured reproducibility and is 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 further shows that GPT scores are highest when SJMM and SBERT agree (up to 0.83). Notably, GPT aligns more closely with SJMM than with SBERT overall, suggesting that the SJMM silver-standard labels—despite their noise—retain meaningful occupational information. SBERT's similarity rankings correlate with GPT scores, indicating that its ranking signal reflects GPT's assessment of label relevance.

5.2 Human Annotation of Challenging Cases

How do LLM ratings compare to human annotations in cases with major discrepancies between SJMM labels and SBERT predictions? To assess this, three domain experts rated the same examples as GPT (using the "randEx1" setting), focusing on 48 particularly difficult cases—the *challenge set*—in which the SJMM-assigned label did not appear among SBERT's top three candidates. These human ratings serve as a benchmark for evaluating GPT's reliability in resolving substantial classification discrepancies.

To harmonize rating comparisons, we applied an adaptive binarization strategy. If an annotator did not assign a 1.0 rating to any candidate, those rated 0.5 were treated as positive; otherwise, only 1.0-rated suggestions were considered positive. This method accommodates variation in annotator strictness and helps distinguish between well- and poorly rated candidates.

Inter-annotator agreement (IAA) among human raters was low ($\alpha = 0.42$ on average), reflecting both the complexity of the cases and the absence of extensive annotator calibration. GPT ratings show higher internal consistency ($\alpha = 0.64$), and their agreement with humans improves when examples are included ($\alpha = 0.45$).

Human annotators slightly favor both SJMM occupations and SBERT's top candidates more than GPT, with scores of 0.52 vs. 0.33–0.48 and 0.42 vs. 0.28–0.31, respectively. GPT evaluations are more stable but somewhat conservative, occasionally missing valid human-rated matches. Overall, GPT ratings show reliable and consistent structure, supporting their use in refining training data.

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

The preceding feasibility study confirms that GPT ratings offer a reliable way to evaluate CH-ISCO class candidates, showing strong internal consistency and reasonable alignment with SJMM and SBERT predictions. However, disagreements—both among human annotators and between GPT and human ratings—underscore the need for careful curation when incorporating GPT-

Language	de	en	fr	it	all
SJMMNotTop1	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.

generated labels into training data.

To generate CH-ISCO-specific MNR training data, we select cases where at least two sources—GPT, SBERT, or SJMM—agree. Job ad titles and descriptions serve as anchors, with agreed classes as positives. Negative examples are drawn primarily from GPT-rated 0.0 candidates, with random sampling used as a fallback.

To ensure effectiveness, we apply weighted sampling to amplify the impact of high-confidence classifications. Classes rated 1.0 by GPT receive a weight of 2, reflecting strong confidence, while those rated 0.5 receive a weight of 1. If the SJMM silver label matches SBERT's top candidate, it is included as a positive example with weight 1—reinforcing agreement without overemphasis.

6.1 MNR Training Data Strategies

To improve similarity-based classification accuracy, we fine-tune SBERT models using MNR with carefully curated training data. This raises two key questions: which data should be selected for MNR fine-tuning, and how should language imbalance be addressed?

The training data is drawn from the SJMM dataset (1990-2023, 65k job ads), focusing on 50k cases where the SJMM-assigned CH-ISCO class differs from the top-ranked prediction of the best-performing SBERT model. Each case is rated by GPT to assess the reliability of both labels. The dataset is multilingual, covering German, French, English, and Italian job postings. To address language imbalance, sampling adjustments are applied to ensure representativeness.

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.

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 CH-ISCO-5digit classification of model *jobLMde-MNR*, evaluated on SJMM data from 1990 onwards (n=65k), before and after initial updates with different training data settings.

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 roughly 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 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 5 show that all MNR training configurations using curated job ad data enhance the performance of the *jobLMde-MNR* model, originally fine-tuned on ontological data, when applied to SJMM job ads. These gains, though measured against silver-standard labels, still reflect meaningful improvements in occupational alignment. Notably, larger datasets lead to stronger effects, and including both cases with large and small classification discrepancies boosts performance. The MNR fine-tuning process improves overall ranking quality, aligning job ad descriptions more closely with CH-ISCO classifications.

Language Balancing: As shown in Table 4, German dominates all datasets. Initial experiments with a balanced sampling strategy—downsampling German ($\times 0.25$) and upsampling English and

	Large 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 6: Performance of pre-fine-tuned and SJMM&Top3Large-updated *jobLMde-MNR* on the large held-out random test set for German (de), English (en), French (fr), and Italian (it). The table reports the percentage of job ads with SBERT suggestions or SJMM occupations rated 0.5 or 1 in ranks 1–3, along with mean GPT rating scores for both. The test set includes 100 ads per language, with four suggestions per ad.

French ($\times 2$) and Italian ($\times 4$)—resulted in smaller relative gains for German compared to other languages (see Figure 6 in the Appendix). 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.

This optimal training dataset *SJMM&Top3Large*, includes 232k samples, distributed as 38% German, 36% French, 23% English, and 3% Italian. Overall, we achieve a Top-1 accuracy of 58.3% and a Top-3 accuracy of 76.3% on the SJMM silver data (see Table 5). Notably, Figure 6 shows that all languages benefit from positive cross-lingual transfer. While German—the dominant training language—achieves the highest Top-1 accuracy (60.1%), Italian (58.4%) and French (56.0%) follow closely. English remains challenging, with a substantially lower Top-1 accuracy of 43.9%.

6.2 Ontology Retention

Does the improvement in CH-ISCO job ad classification also improve performance on the original ontology-based CH-ISCO classification task?

The results indicate that fine-tuning leads to a trade-off: Top-1 accuracy on ontology data decreases by roughly 10 points after fine-tuning, while Top-3 accuracy remains high at 95% (see Figure 7 in the Appendix). This decline reflects a data shift between job ads and ontological data. Qualitative analysis suggests that many misclassifications are non-critical (see Table 11 in the Appendix), with English occupation assignments frequently preferring more generic classes, which correspond to 4-digit codes. This may stem from the initial scarcity of English data at the more fine-grained 5-digit level.

6.3 Results: Evaluation on Held-Out Test Sets

To assess the model's generalization to unseen job ads, we evaluate it on a held-out sample of 400 randomly selected job ads—100 per language—referred to as the *large random test set*. This set includes the smaller 100-ad sample used in earlier evaluations and serves as the basis for assessing the best-performing *SJMM&Top3Large* fine-tuning. The expanded size allows for both overall validation and cross-language comparison.

GPT-based ratings on this large random test set confirm improved agreement for top-ranked SBERT suggestions, demonstrating the effectiveness of the fine-tuning approach. The *SJMM&Top3Large* update significantly enhances

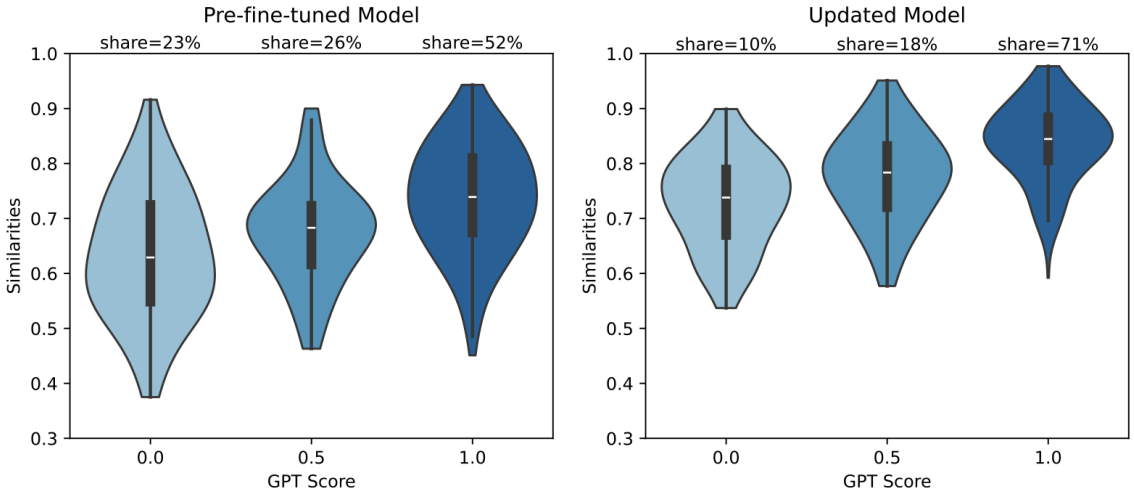

Figure 3: Violin plots showing the similarity score distributions for SBERT top candidates, categorized by GPT rating, across the large held-out random test set (n=400) for both pre-fine-tuned and SJMM&Top3Large-updated *jobLMde-MNR*. Each plot includes a box plot, and GPT rating distributions are indicated at the top of each plot.

GPT alignment with top SBERT suggestions. As shown in Table 6, the proportion of top-ranked suggestions rated as acceptable or perfect matches ($\geq$ 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 validation is needed to assess actual issues in the silver data.

These results demonstrate the *SJMM&-Top3Large* update's effectiveness in refining SBERT's top-ranked predictions, increasing GPT validation scores, and improving performance on realistic job ad classification scenarios.

6.4 SBERT Similarity and GPT Rating

How do GPT-curated updates affect SBERT's cosine similarity?

Figure 3 shows how similarity scores vary with GPT ratings before and after the update on the held-out set. The impact is evident in the similarity values of top-ranked candidates. 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 0.7 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.

6.5 Multiple Rounds of LLM Refinement and Final Human Evaluation

Finally, we tested whether a second round of GPT annotations on newly misaligned cases could further improve performance. This included a final human evaluation conducted alongside GPT ratings—limited to the small random sample due to annotation constraints. We assessed the top-1 suggestions from three model variants—the SJMM&Top3Large model, and those updated sequentially or in combination—alongside the silver-standard SJMM label. The goal was to evaluate how well the final SBERT predictions aligned with human judgment.

Model performance, based on majority votes from three human raters, reached a Top-1 precision of 0.79 for both the SJMM&Top3Large model and the combined update (see Table 7). The combined update also yielded at least one acceptable top-ranked suggestion (rated $\geq$ 0.5) in 86% of cases, indicating strong performance from a human perspective.

Perfect ratings (score = 1.0) were consistent across models, ranging from 72% to 73%, according to both humans and GPT. Differences appeared in at least 0.5-rated suggestions: humans saw a slight gain with the combined update (from 85% to 86%) but a drop with the sequential update (to 82%). GPT showed smaller differences, with the

	top3Large	comb.	seq.
% rated ≥ 0.5 in rank 1:			
by GPT	90%	89%	**91%**
by Humans	85%	**86%**	82%
% rated = 1 in rank 1:			
by GPT	73%	73%	73%
by Humans	**73%**	72%	**73%**
Mean rating of rank 1:			
by GPT	**0.82**	0.81	**0.82**
by Humans	**0.79**	**0.79**	0.78
Mean rating SJMM occ.:			
by GPT	0.71	0.71	0.71
by Humans	0.68	0.68	0.68
% ads with SJMM occ. in rank 1	53%	**55%**	52%

Table 7: Performance of the SJMM&Top3Large-updated model, evaluated before (top3Large) and after combined (comb.) and sequential (seq.) updates on the small held-out random sample (n=100 ads), based on human and GPT evaluations. The table shows the percentage of ads where rank-1 SBERT suggestions or SJMM occupations received a rating of 0.5 or 1.0, along with mean ratings for both. Best values per column are shown in bold.

sequential update slightly ahead. These results suggest that sequential updates may overfit to GPT-specific preferences, potentially reducing generalization as seen in human evaluations.

In summary, GPT-curated training data yields strong performance on this fine-grained classification task, achieving roughly 80% top-1 precision, and improving held-out performance by up to 30 points compared to initial evaluation (see Table 3). Additional refinement rounds offer minimal benefit, and update strategy comparisons remain inconclusive. Overall, a single round of curation proved both efficient and effective in our use case.

7 Conclusion and Future Work

We show that LLM-assisted data curation leads to marked improvements in SBERT-based multilingual occupation classification. The refinement process consolidates information from two sources—the original silver-standard labels in the SJMM data and predictions from a pre-fine-tuned SBERT model—using GPT ratings to resolve discrepancies and validate candidate assignments.

Fine-tuning SBERT with this GPT-refined data substantially improves alignment with the SJMM silver standard, raising Top-1 accuracy from 37.2% to 58.3%. More importantly, on held-out random test data, we achieve an average Top-1 precision of 80%—up from around 50% before fine-tuning—as confirmed by both GPT and human raters. This marks a significant gain in a fine-grained (670-class) and complex classification task.

While task-specific fine-tuning reduces Top-1 accuracy on the original ontology-based classification task by approximately 10 points, Top-3 accuracy remains high at 95%. The impact is limited and acceptable, given that ontology retention is not the primary objective.

In our case, GPT's consistency and structured scoring provide a reliable mechanism for curating high-quality training data. The results suggest that a single round of GPT-based refinement is both effective and efficient—delivering strong improvements in classification accuracy.

Future Directions: To further advance multilingual occupation classification, future work should aim to: (1) improve label quality in edge cases through targeted human annotation; (2) expand the coverage of English labels to reduce mismatch issues; and (3) develop ontology-aware fine-tuning strategies that better preserve structured knowledge during task adaptation.

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.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback and insightful comments. We are also grateful to Norma de Min, Jana Finkbeiner, Arkhip Lovin, and Jan Müller for their dedicated annotation efforts. This work was supported by the Swiss National Science Foundation under grant number 10FI 229649.

References

Marlis Buchmann, Helen Buchs, Eva Bühlmann, Ann-Sophie Gnehm, Debra Hevenstone, Yanik Kipfer, Urs Klarer, Jan Müller, Marianne Müller, Stefan Sacchi, Alexander Salvisberg, and Anna Von Ow. 2024. Swiss job market monitor 1950-2023 (9.0.0) [dataset].

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*.

Daniel Deniz, Federico Retyk, Laura García-Sardiña, Hermenegildo Fabregat, Luis Gasco, and Rabih Zbib. 2024. Combined unsupervised and contrastive learning for multilingual job recommendation. In *Proceedings of the 4th workshop on recommender systems for human resources (recsyshr 2024)*, pages 28–37.

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.

Ann-Sophie Gnehm, Eva Bühlmann, and Simon Clematide. 2022. Evaluation of transfer learning and domain adaptation for analyzing german-speaking job advertisements. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3892–3901, Marseille, France. European Language Resources Association.

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, Yunhsuan 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.

David JC MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press, Cambridge.

Antoine Magron, Anna Dai, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. 2024. JobSkape: A

framework for generating synthetic job postings to enhance skill matching. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 43–58, St. Julian's, Malta. Association for Computational Linguistics.

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.

Federico Retyk, Luis Gascó, Casimiro Pio Carrino, Daniel Deniz, and Rabih Zbib. 2024. MELO: An evaluation benchmark for multilingual entity linking of occupations. In *Proceedings of the 4th workshop on recommender systems for human resources (recsys in hr 2024)*, volume 3788 of *CEUR Workshop Proceedings*, pages 12–27. CEUR-WS.org.

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

This appendix contains supplementary figures and tables referenced throughout the main text. The materials are listed below in the order in which they appear:

- **Table 8**: Examples of occupations from ESCO in German, English, French, and Italian with corresponding descriptions.
- **Figure 4**: Excerpt of the system prompt used in GPT's CH-ISCO evaluation task, showing only content-relevant instructions.
- **Table 9**: Extracted and preprocessed occupation information from a German and an English job ad.
- **Figure 5**: Expanded instruction for GPT's CH-ISCO evaluation task, focusing on how to handle managerial positions.
- **Figure 6**: Top1–3 accuracy for *jobLMde-MNR* on silver-standard SJMM data before and after model updates using two different training setups.
- **Figure 7**: Top1–3 accuracy for *jobLMde-MNR* on ontology evaluation data before and after model updates using two different training setups.
- **Table 10**: Job ad examples with top-1 predictions from the pre-fine-tuned *jobLMde-MNR* model and the corresponding SJMM silver-standard labels.
- **Table 11**: Examples of non-critical misclassifications made by the SJMM&Top3Large-updated *jobLMde-MNR* model in ontology-based evaluation.

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

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

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

```
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 5: Expanded instruction for handling managerial positions in the system prompt for GPT ratings (*freqEx-prompt2*).

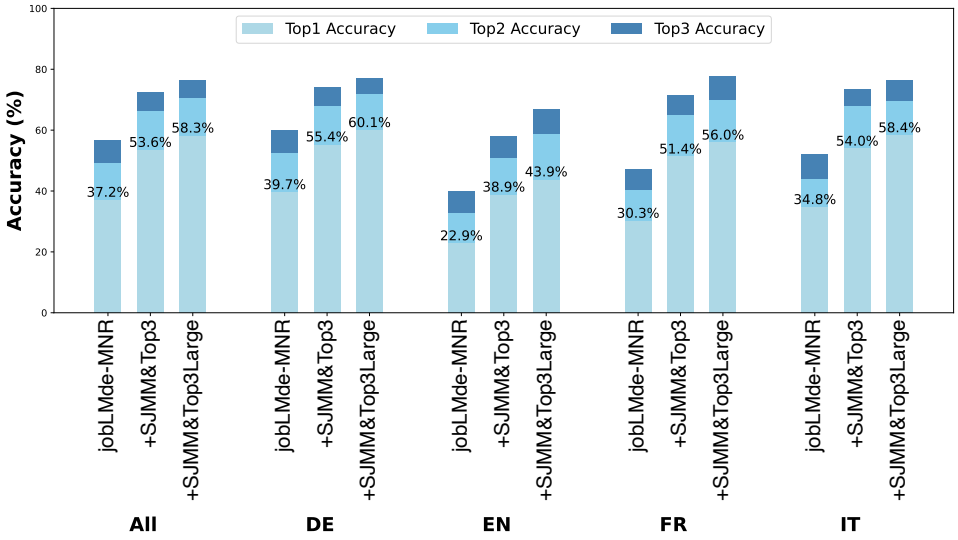

Figure 6: Stacked Top1–3 accuracy (%) in CH-ISCO-5-digit classification of *jobLMde-MNR*, before and after updates using two different training data settings: SJMM&Top3 with balanced language sampling, and SJMM&Top3Large with an aggressive language upsampling strategy. Evaluation is based on SJMM data from 1990 onwards (n=65k). Top-1 accuracy is shown on the bars.

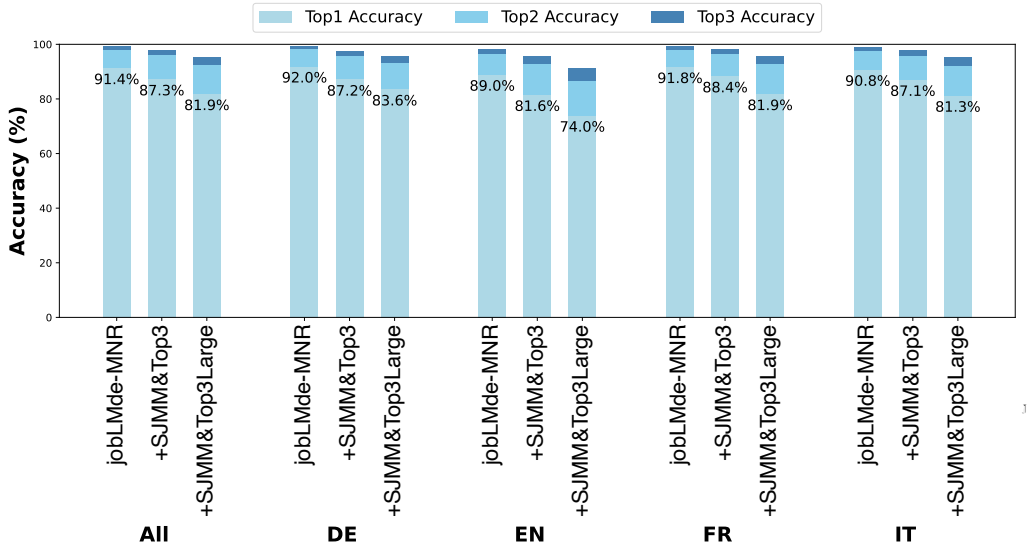

Figure 7: Top1–3 accuracy (%) on ontology evaluation data (n=108k) in ISCO-5digit classification of *jobLMde-MNR*, before and after updates with two different training data settings. Top-1 accuracy percentages are displayed on the bars.

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] Chef-fe de clinique [EJT] Votre mission: Garantir une prise en charge multidisciplinaire de qualité des patient-e-s. Superviser les médecins-assistant-e-s et étudiant-e-s en médecine. Participer à la formation des médecins-assistant-e-s et étudiant-e-s en médecine. ... (*Chief medical officer. Your task: To guarantee high-quality, multidisciplinary care for patients. Supervise assistant doctors and medical students. Participate in their training §...*)	13420 Führungskräfte in der Gesundheitsversorgung (*Health Service Managers*)	22120 Fachärzte (*Specialist Medical Practitioners*)

Table 10: Shortened and translated examples of job ads with top-1 predictions from model *jobLMde-MNR* and corresponding SJMM classifications.

Ontological Job Title	**Ontological Class (CH-ISCO Code)**	**SBERT Top Candidate (CH-ISCO Code)**	**SBERT Rank**
Life science engineer	21310 Biologist, botanist, zoologist	21300 Bioscientist, unspecified	3
Hotelier/Restaurateur	14110 Managers in hotels	14100 Managers in hotels and restaurants	2
Digital collaboration specialist	33410 Secretarial manager	35100 Technician for the operation of ICT and for user support	12
Pharmacy specialist	32130 Pharmaceutical technicians and assistants	22620 Pharmacist	2

Table 11: Examples of non-critical misclassifications by SJMM&Top3Large-updated *jobLMde-MNR* model in ontology-based evaluation.