Improving Parental Occupation Coding Procedures with AI

Research Report, April 2024

Daniel Duckworth (lead researcher)
International Association for the Evaluation of Educational Achievement (IEA)

Julian Fraillon (research supervisor)
International Association for the Evaluation of Educational Achievement (IEA)

CONTENTS

Abstract 3
1 Introduction 6
2 Literature review 13
2.1 Results from other studies 13
3 Research design 15
3.1 ISCO-08 classification model overview 15
3.2 Data and data processing 15
3.3 Structuring the data for model training and evaluation 19
3.4 Model training and evaluation process 21
4 Procedure for Evaluating the models 23
4.1 Evaluation measures 23
4.2 ISCO Hierarchical Accuracy Measure 25
5 Evaluation Results 29
6 Discussion and Conclusion 33
7 Appendices 37
7.1 Appendix A: Terms and abbreviations 37
7.2 Appendix B: ISCO code validation 39
7.3 Appendix C: Class report 40
7.4 Appendix D: Supplementary digital resources 50
8 References 51

Abstract

This research evaluates the efficacy of the pre-trained XLM-RoBERTa (Conneau, et al., 2019) model for use as an ISCO (International Standard Classification of Occupations) classifier of parental occupation. XLM-RoBERTa (XLM-R) is a large multilingual language model, trained on 2.5TB of filtered CommonCrawl text data comprising 100 different languages. The research also explores the potential value of hierarchical evaluation metrics that recognize partial correctness within the ISCO hierarchy.

The International Standard Classification of Occupations (ISCO) is a hierarchical taxonomy developed by the International Labour Organization (ILO) to categorize and compare job titles based on skill level and skill specialization. This classification system is structured into four distinct levels that reflect the complexity and responsibility of job tasks across various sectors of the economy. At the top of the hierarchy are the major groups, which delineate broad categories of occupations based on similar skill sets and tasks, such as managers or healthcare professionals. Under these are the sub-major groups, offering a more detailed classification within the broader categories, such as specialized managers or specific healthcare practitioners. The minor groups within the sub-major groups identify specific roles within each sub-major group, highlighting the diversity of occupations within a given field, such as nursing professionals. At the most detailed level, the unit groups provide a precise classification, distinguishing between job titles and roles based on specific job duties, skills, and qualifications, like general nurses versus specialized nurse practitioners. This hierarchical structure facilitates the standardized comparison of occupational data across different countries and regions.

Parental occupation is a valuable and frequently used measure of student socioeconomic status (SES) in international large-scale assessments (ILSAs). One way of collecting data on parental occupation is to ask students to name and describe the jobs of their parents. These responses can then be classified against the International Standard Classification of Occupations (ISCO). Quality control procedures of this coding typically involve double coding of student responses to measure inter-rater agreement and to identify and adjudicate discrepancies. While reliability coding within countries is relatively straightforward, it is more challenging to manage this across countries and languages. Recent advances in natural language processing (NLP) and multilingual large language models (LLMs) offer an opportunity to enhance the accuracy and efficiency of cross-national occupational data coding and ultimately improve the validity and reliability of the derived SES measures.

The text data used for model training and evaluation were sourced from the International Computer and Information Literacy Study (ICILS 2013, ICILS 2018) and the International Labour Organization (ILO). The ICILS data were processed to balance the distribution of samples across the ISCO groups using undersampling and oversampling techniques that prioritized double coded samples in agreement. The student descriptions and occupation titles were used as the text feature and the human assigned 4-digit ISCO codes (at the unit group level) as the reference labels. The ILO data comprised a coded index of occupation titles in

English which were machine translated with OpenAI's GPT-4 model into eleven of the ICILS languages.

Three models were developed by fine-tuning the XLM-R model with the two datasets and a third merged dataset, using a data division of training (80%), validation (10%), and testing (10%). This methodology aims to ascertain if the features learned from the training and validation data can be effectively applied to unseen data in the testing set. Twelve monolingual models were designed to evaluate their performance on language-specific subsets of the ICILS data, enabling a direct comparison with the multilingual model using the same test splits.

Model accuracy was evaluated by comparing the predicted labels against the reference labels from the test split, with accuracy defined as the ratio of correct predictions to total predictions. Furthermore, for a valid comparison of accuracies, the models fine-tuned with the ILO dataset and the combined dataset (ICILS and ILO) were also evaluated using only the test split of the ICILS dataset.

The evaluations further included measures of precision and recall for each ISCO-08 unit group and introduced a novel implementation of hierarchical precision and recall that recognizes partial correctness within the ISCO hierarchy. For example, if a specific occupation is misclassified at the unit group level, within a correctly classified minor group the classification is still more correct than a misclassification of the occupation at the minor group level within a correctly classified sub-major group. The traditionally used measures of accuracy would treat both examples as incorrect classifications, without being sensitive to the relative accuracy of each misclassification.

Multilingual model evaluation results are summarized as follows:

- Model 1 (ICILS data) accuracy was 63%. It predicted the exact same occupation group as human coders across 13 countries for 63% of the test split records.
- Model 2 (ILO data) accuracy was 92%. It predicted the exact same occupation group as coded by the ILO for 92% of the test split records.
- Model 3 (ICILS+ILO data) accuracy was 80%. It predicted the exact same occupation group as ICILS and ILO coders for 80% of the combined test split records.
- When evaluated with only the test split of the ICILS data, the accuracy of Model 2 and Model 3 was 36% and 62% respectively.
- The hierarchical F-measure (hFβ) scores for each model evaluated with only the test split of the ICILS data were 0.89, 0.94, and 0.95 respectively.

The improvement in hFβ scores, particularly for Model 3, suggests that combining datasets can enhance the model's ability to make partially correct predictions within the ISCO hierarchy. This indicates the potential value of hierarchical evaluation metrics in contexts where partial correctness is relevant. The fact that the inclusion of ILO data did not improve the accuracy of Model 3 over Model 1 when evaluated on ICILS data but did improve the hFβ score implies

that for applications where hierarchical correctness is valuable, merging diverse datasets could be beneficial, even if it does not increase overall accuracy.

To help make this research more accessible and transparent, we have included several supplementary resources:

- ICILS Multilingual ISCO-08 Parental Occupation Corpus Dataset is a private[1] git repository with the dataset configurations and splits used in the research and can be explored online and analyzed using the Hugging Face Datasets Python package.
- The ISCO-08 Hierarchical Accuracy Measure metric card is a publicly accessible implementation of the evaluation measure with the Hugging Face Evaluate Python package and demo space for testing the implementation.
- ICILS XLM-R-ICILS-ILO model card and Hosted Inference API is a gated access[2] git repository with the fine-tuned model weights, configuration files, and parameters used for training.
- ICILS ISCO R&D Model Trainings Reports is an automatically generated report of the model's accuracy over time captured during the training procedure with machine learning experiment monitoring software.
- ICILS ISCO AI Git Repository is a publicly accessible code repository with the Python code for reproducing the model training and evaluations.

The outcomes of this research demonstrate the potential for multilingual LLM to contribute to occupation coding in ILSAs. The model's ability to interpret and classify occupations from textual descriptions in twelve languages addresses the challenge of accurately coding parental occupation data across countries and hence contributes to the consistency and reliability of comparative SES measures. The development of hierarchical accuracy measures for ISCO-08 opens avenues for more sophisticated quality monitoring of not only model performance, but also of coder training feedback mechanisms that support the differentiation of socioeconomic nuances within occupational groups, laying the groundwork for future innovations in occupation coding procedures.

Overall, this research contributes to the broader discourse on leveraging technology to address long-standing challenges in educational and occupational research, promoting greater inclusivity and precision in international studies, and contributes to global educational equity by ensuring that SES measures accurately reflect diverse occupational contexts around the world.

[1] The dataset is securely hosted and is accessible only to IEA organization members.
[2] Access to the model repository requires a user to share their contact information (email and username), to agree to IEA's terms of use, and for the request to be manually approved by the repository owners.

1 INTRODUCTION

In this section we provide an overview of the challenges associated with coding parental occupations in international educational research, specifically within the context of the International Standard Classification of Occupations (see section ISCO Hierarchical Accuracy Measure). We identify issues such as the difficulty of achieving high inter-rater agreement, the potential for systematic biases in occupational classification, and the implications of these biases on the reliability of socioeconomic status (SES) measures and, consequently, for the validity of trend and cross-national comparisons.

Cross-national parental occupation coding is a procedure in comparative social research that empowers researchers to compare occupational structures across countries. Such comparisons can enable insights into the architecture of labor markets, social stratification, and economic development across diverse national contexts. In International Large-Scale Assessments (ILSAs) the procedure entails the classification of parental occupations, as described by students, into internationally standardized categories. By employing a common occupation classification scheme such as the International Standard Classification of Occupations (International Labour Organization, 2012), differences in national schemes are reconciled by mapping national schemes to the international standard. The information obtained from analyses of data that rely on coded parental occupations is an essential component for understanding cross-national variations in social and economic inequality, alongside analyzing labor market dynamics and policy implications. In a typical ILSA, parental occupation data is collected from participating countries, and the coding of these data, coordinated by national study centers, contribute to measures of social status (Wolf C. , 1997) from Socio-Economic Index scale (Ganzeboom, De Graaf, & Treiman, 1992).

The computation of an International Socio-Economic Index (ISEI) (Ganzeboom H. , 2010) scale of occupational status from ISCO coded data is a prevalent practice in ILSAs. The International Computer and Information Literacy Study (ICILS) 2018 cycle employed a composite index of students' socio-economic background created with an ISEI combined with an indices of the highest level of parental education and the number of books in the home (Fraillon, Ainley, Schulz, Friedman, & Duckworth, 2018, p. 170). Composite indices of this nature support the formulation of responses to key research questions and fulfill a significant reporting function that aids in explaining variations in students' educational achievement. It is well recognized in the research literature that background attributes such as socioeconomic status (SES) consistently predict student achievement. A meta-analysis encompassing 71 research papers concerning ILSA studies (Holzberger et al., 2020) analyzed common student and school background characteristic variables. The analysis identified that SES composition showed the most robust relationship with achievement (r = .30), followed by out-of-school activities (r = .18). While it is contended that background attributes like SES cannot be influenced by educators, it warrants acknowledgment that the continual reporting of consistent findings is important for contributing to an evidence based public discourse on

education (a cultural resource) and income (an economic resource), which are the principal forms of power in contemporary societies (Treiman, 2013).

Accurately coding parental occupations with over 400 occupation groups from ISCO-08 is a challenging task for human coders and significant operational cost for national study centers. Assessing the accuracy of occupation coding is also methodologically challenging because it is an inherently subjective evaluation with no objective reference point. Instead, proxy measures of accuracy such as inter-rater reliability are used to assess the degree of agreement among independent coders within countries. An OECD report on the theory and adoption of ISCO-88 reported the outcomes of seven recoding studies, noting that employing a 3-digit coding frame (350 categories) made obtaining inter-rater agreement exceeding 75 percent a challenge (Elias, 1997).

However, even a degree of agreement among coders within countries does not necessarily mean that there is also a high degree of agreement across countries. If systematic biases are involved in the classification process and those biases manifest differently across countries, nationally reported reliability measures may not detect systematic differences across countries.

Consider for instance the classification of a description of a farmer with ISCO-08. The description could be coded as Managers in Agriculture (1310), Market-oriented Skilled Agricultural Workers (6100), or Subsistence Farmers (6200). The differences in ISEI scores (Ganzeboom H. , 2010) among these groups illustrate a range of economic statuses within the agricultural sector. Managers in agriculture (ISEI score of 49) have a notably higher socio-economic status compared to market-oriented skilled agricultural workers (ISEI score of 18) and subsistence farmers (ISEI score of 10). Hence systematic differences in how descriptions of farmers are classified between coders within and across countries could affect the derived SES measures. Hence the validity of trend and cross-national comparisons relying on derived measures of SES requires careful examination.

We can summarize this problem as the differential in SES from the misclassification of occupations in semantically similar occupation groups. Comparing inter-rater agreement between coders and across countries can help address the problem; however, the descriptions parental occupations described by students are in the language of the study instruments, requiring either multilingual coders or cross-nationally translated occupation descriptions. Hence addressing the problem across national contexts is challenging.

Therefore, the problem remaining for ISLAs like ICILS is in measuring cross-national reliability of occupation coding and the development of practices for ensuring the consistent application of ISCO-08 across national contexts.

Three approaches to dealing with the issue of cross-national consistency that we will briefly discuss are:

- Cross-national coder training: The use of cross-nationally translated samples in national coder training and monitoring practices.
- AI occupation coding assistant: The adoption of a multilingual LLM as occupation coding assistant with humans in the loop. This strategy has two opposing use cases. Given a student description of a job, the assistant recommends the 3-5 most relevant occupation groups as a function integrated into:
    - the occupation coding software used by professional human coders; or
    - the student questionnaire, presenting the recommendations as options for students to choose.
- AI automation: The adoption of a multilingual LLM as a fully automated occupation classification service.

Multilingual LLMs could be used to support translation for the first approach and are at the core of the remaining two approaches. Using multilingual LLMs does, however, present some challenges in this context. The first challenge is in sourcing and developing a comprehensive dataset comprising job descriptions and their corresponding ISCO-08 codes. The dataset would need to sufficiently represent every ISCO-08 occupation group for every language. While there is no one-size-fits-all for the number of samples required for a dataset to be sufficient, let us assume the minimum is 100. In a multilingual context, the number of samples per class must be multiplied by the number of target languages. Hence a minimally sufficient dataset for training a multilingual ISCO classifier across 10 languages is 1,000 examples per class, which for the 433 ISCO-08 occupation groups is 433,000 total examples.

Other challenges include safeguarding against biases inherent in the training data, ensuring the model's adaptability to nuanced and evolving language use in job descriptions, and maintaining transparency and explainability in the model's decision-making processes. To address these challenges, ongoing model evaluation with human oversight in the form of quality checks with robust measures are essential components of adopting LLMs into occupation coding procedures.

Cross-national coder training

The implementation of cross-national coder training involves a variation of 'Gold Standard Coding' and 'Seed Coding.' Gold standard coding refers to the use of examples that have been pre-coded by a consensus among experts. Typically, the samples, without the occupation code, are shown to coders in training and the coders are then expected to assign the same code as the experts. Variations in the assignment of occupation codes are addressed by the trainer with a view to achieving consensus amongst the coders in training. Seed coding refers to a practice where a set of gold standard samples are included (or seeded) amongst the data to be coded and are presented to coders at random as if they were regular samples. In this scenario, differences can be acted on with interventional training or simply included in reliability reporting as the proportion of correctly coded samples.

These practices are typically coordinated within countries and hence do not address the problem of measuring cross-national inter-rater agreement because we do not know whether the accuracy of coders from different countries varies for the same set of gold standard samples. This can be addressed, however, by employing machine translation tools to translate a set of gold standard samples from each country participating in the study into the languages of every other participating country. The use of cross-nationally translated samples in coder training and monitoring can enable the measurement and reporting of cross-national inter-rater agreement. This approach can provide a direct measure of cross-national inter-rater agreement (for the selected samples) and may indirectly contribute to some increased consistency of occupation coding across countries.

AI occupation coding assistant

Assisting occupation coders

A research report published by the European Commission on "Machine Learning Assisted Mapping of Multilingual Occupational Data to ESCO" (ESCO Publications, 2022) offers a thorough examination of the methodologies employed to map national occupational classifications to the occupations pillar of the multilingual classification of European Skills, Competences, Qualifications, and Occupations classification (European Commission and Directorate-General for Employment, Social Affairs and Inclusion, 2022). The report stresses the role of expert validation in confirming the types of matches, spanning from exact to broader, narrower, and close matches, between the national source concept of an occupation and the ESCO concept. The researchers used representation learning techniques to handle free text across 28 languages, “thereby aligning the embedding space for the different languages” (ESCO Publications, 2022). The top 5 accuracy ranged between 75% and 83% for Member State classifications and reached up to 94% for the O*NET classification.

In an effort to enhance the comparison of cross-national occupation taxonomies, the European Commission harnessed the BERT language model to create a "crosswalk between ESCO, the European multilingual classification of Skills, Qualifications, and Occupations, developed by DG Employment, Social Affairs and Inclusion of the European Commission, and O*NET, the Occupational Information Network, developed by the U.S. Department of Labor/Employment and Training Administration (US DOL)" (European Commission, 2022). The best model’s performance for an exact match was 84% and increased to 100% when the correct prediction was one of the top 7 predictions ranked by probabilities.

The use of pre-trained multilingual LLMs as classifiers, such as the ones developed by the European Commission, can also be used to classify student descriptions of parental occupations against ISCO. Using existing occupation data collected in a study, the student descriptions are the text feature and the 4-digit ISCO-08 codes assigned by human coders are the reference labels. The model can then be trained using a ‘fine-tuning’ technique which preserves the original weights of the language model and adds classification layer with new weights learned from the data. A trained model can then be deployed as a web API service,

taking text as an input, and returning predictions that can be consumed by other software applications such as those used by human coders in occupation coding procedures and those used by questionnaire developers to create study instruments.

Occupation coding software is used by trained coders to map job descriptions or titles into standardized occupation codes and typically features several key components: randomized allocation of samples, presentation of job description texts, user interface (UI) for code entry, and mechanisms for ensuring and reporting reliability. The randomized allocation of samples to coders is a fundamental methodology that ensures the objectivity and reliability of the occupation coding process. This is typically achieved with randomization algorithms that distribute job descriptions or titles among a team of coders to minimize biases. The goal is to ensure that no single coder is overrepresented in coding specific types of jobs, which could skew the data. It can also include load balancing where the software dynamically adjusts the distribution of samples to ensure an even workload among coders, considering their availability and coding speed.

The way job descriptions are presented to coders is important for accurate interpretation and efficient coding. Some coding software tools apply text normalization techniques to standardize text, such as removing irrelevant symbols, correcting typos, and sometimes translating languages to fit the coder's proficiency. Important terms or phrases within the job description that are likely to influence the occupation code selection can be highlighted to draw the coder's attention.

The user interface (UI) for entering codes is designed to be intuitive and efficient, enabling coders to quickly find and select the appropriate occupation codes. Some tools include search and autocomplete functions that enable coders to search for occupation codes using keywords, and the information retrieval system suggests matches as they type, speeding up the selection process. The UI typically also displays the hierarchy of occupation codes and provides descriptions for each, helping coders understand the relationships between different codes and ensuring they choose the most relevant one.

Reliability of coding is fundamental to coding tools, and the software incorporates several features for this purpose:

- Inter-coder reliability measures: Randomly selected samples are coded by multiple coders independently to assess consistency in their selections. Discrepancies can be reviewed to identify misunderstandings or ambiguities in code definitions and can be escalated to a more experienced coder for adjudication.
- Coder training and feedback loops: The system provides feedback to coders based on reliability checks, highlighting areas for improvement (e.g., indicating when a coder's decisions are overly representing an occupation group relative to other coders), and contributes to ongoing training.

- Reliability reporting: The software can generate reports detailing coding accuracy, inter-coder reliability scores, and other metrics important for assessing the quality of the coding process.

Integrating a multilingual LLM as an assistant into occupation coding software represents a significant advancement in the standardization and consistency of occupation coding across national contexts. This integration leverages the capabilities of LLMs to understand and classify texts in multiple languages, ensuring that the classification is consistent, regardless of the original language of the job description which aligns with the purpose of the ISCO-08 framework. Hence, we propose the 'Initial LLM Classification' method which presents the LLM's suggested ISCO-08 codes as preliminary classifications for each job description, alongside the description in the coder's workspace. Under this method, human coders review the LLM's suggestions, and make the final decision based on their expertise and understanding of nuances that the model might miss. Corrections and adjustments made by human coders can be fed back into a LLM training and evaluation pipeline (see Appendix D: Supplementary digital resources). When there is a measurable improvement in the new model's accuracy compared to a baseline benchmark, it can be used to replace the old model. We also suggest the implementation of a semantic search function that enhances the search and autocomplete functions with the LLM's suggestions, making it easier for coders to find and confirm the most accurate codes.

The 'Initial LLM Classification' method could significantly improve the consistency, accuracy, and efficiency of occupation coding across various national contexts. The method theoretically leverages the strengths of both AI and human expertise, offering a potential solution to the challenges of cross-national occupation classification. However, the method needs to be empirically tested and carefully evaluated in comparison with the existing methods to determine whether it is truly effective.

Assisting respondents

Self-identification (Tijdens, 2015), or self-coding in student questionnaires, proposes that instead of using open question formats, respondents choose an occupation from a series of nested lists based on the hierarchy of ISCO. The idea of self-coding certainly has merit. First it would eliminate the need for open responses to be coded. Second, given a student is a better judge of their intention than an external human coder is of the student's writing, and assuming the student fully comprehends the meaning of the options presented to them, self-identification may provide a more accurate representation of parental occupation than external coding. However, a list derived from ISCO-08 minor groups would include 130 items and represent an excessively onerous response burden (Martin, Rust, & Adams, 1999, p. 26). Conversely, a more reasonably manageable list derived from the 10 major groups or even 43 sub-major groups of ISCO-08 is insufficient for SES measures because ISEI scores must be averaged for the occupations within those groups. These occupations can have significantly

different educational requirements and earnings, which are nuanced details that get lost when averaged together. Yet the principle of self-coding should not be overlooked.

Hence for the second application of the AI coding assistant strategy, we propose a variation of self-identification that retains the open format question for use with an AI classification assistant in-place of nested lists. Under this self-coding model, a student's description of a parent's occupation is submitted to an ISCO classification model deployed as a web API service and, in a follow-up question, the student is presented with the top 3-5 predictions returned by the model and asked to choose the most relevant occupation.

To address the issue of students struggling with the technical language used to define ISCO occupation groups, we propose a solution. By utilizing state-of-the-art, multilingual generative AI models, such as GPT-4 (OpenAI, 2023) and Gemini (Gemini Team, et al., 2023), we aim to simplify complex occupational content. These models can rephrase the content into simpler, age-appropriate language while retaining the original meaning with a level of quality considered to be on par with human annotators (Feng, Qiang, Li, Yuan, & Zhu, 2023). Importantly, this adaptation occurs in the language of the test, ensuring accessibility and comprehension for all students.

AI automation

AI automation represents the culmination of leveraging advanced machine learning technologies to streamline and enhance the accuracy of occupation classification, eliminating the need for human involvement in the coding process. This strategy employs LLMs to fully automate the task of mapping job descriptions or titles to standardized occupation codes. The approach is predicated on the model's ability to interpret and classify text inputs across multiple languages with sufficient accuracy. Accuracy must therefore be evaluated with a comprehensive dataset. As a service, the LLM could be integrated into various data collection and analysis workflows for coding occupations in ILSAs. The advantages of this approach include significant reductions in time and resource expenditures associated with manual coding, as well as the potential for increased accuracy and consistency in occupation classification across different languages and national contexts.

2 LITERATURE REVIEW

The classification of occupations using large language models such as BERT and XLM-Roberta has gained significant attention in recent years. Open-source pre-trained large language models offer practical and efficient means to creating high quality task-oriented models. In a comparative study of BERT and GPT3 for automated occupation coding the researchers noted that "training deep language models on our data is time-consuming and computationally expensive. Pre-trained language models are, therefore, attractive because they reduce the burden on practitioners to provide the appropriate resources (time, hardware, and data) to train the models" (Safikhani, Avetisyan, Föste-Eggers, & Broneske, Automated occupation coding with hierarchical features: a data-centric approach to classification with pre-trained language models, 2023). The four main approaches to AI-based classification of occupations are text-based classification, multi-modal classification, ontology-based knowledge representation, and transfer learning with deep learning language models (i.e., large language models). Many studies have focused on using natural language processing (NLP) techniques to extract relevant features from job descriptions and classify occupations. These approaches often involve methods like text embedding, topic modelling, and supervised machine learning algorithms. Ontology-based approaches involve the use of occupation taxonomies (e.g., KldB, ISCO-08 and ESCO) to facilitate classification. Taxonomies provide structured knowledge representations of occupations, including hierarchical relationships, skill requirements, and job tasks. By leveraging ontological knowledge, researchers have developed ontology-based classification frameworks that achieve high accuracy in occupation classifications (Safikhani, Avetisyan, Föste-Eggers, & Broneske, Automated occupation coding with hierarchical features: a data-centric approach to classification with pre-trained language models, 2023; Zhang, Goot, & Plank, 2023). Transfer learning is a technique where a model's representation of knowledge learned from one task is transferred to another related down-stream task. Efforts have been made to establish benchmark datasets and evaluation metrics for multilingual natural language understanding such as the "MASSIVE and Massively Multilingual NLU 2022" (Yadav, et al., 2019). These benchmarks allow researchers to compare the performance of different approaches and track progress in the field helping to ensure the reproducibility and reliability of research findings.

2.1 Results from other studies

Several pertinent studies have delved into the application of large language models for the task of occupation coding. In a comparative analysis of BERT and GPT-3 (Safikhani, Avetisyan, Föste-Eggers, & Broneske, 2023), the researchers employed the German Classification of Occupations 2010 (KldB 2010) with the goal of outputting predictions for subsets of the KldB digits, given that some studies only utilize segments of the full classification. The KldB 2010 is a hierarchical 5-digit classification schema that encapsulates both occupational group and skill level. The study draws upon occupation and task descriptions from the DZHW Graduate Survey and School Leaver Survey collected from 2005 to 2015. The data was coded manually

to ensure a reliable reference for training and evaluation. Both BERT and GPT-3 were fine-tuned on the full 5-digit KldB labels, framing it as a multi-class text classification task. Additional experiments were run, fine-tuning models to predict each KldB digit independently as a sequence of single-label classification tasks, thus facilitating partial predictions. Hierarchical features from the KldB descriptions were also integrated into the fine-tuning by appending literal descriptions of preceding digit combinations to the model input and constructing the validation and test sets incrementally predicated on previous predictions, to better model the relationships between digits. BERT markedly surpassed GPT-3 and the previous state-of-the-art on full 5-digit prediction. The researchers reported model performances of 64.22% versus 29.93% and 48.5% respectively[3].

ESCOXLM-R (Zhang, Goot, & Plank, 2023) is based on XLM-Roberta-Large and employs domain-adaptive pre-training on the ESCO taxonomy. Encompassing the 27 languages of ESCO, ESCOXLM-R incorporates dynamic masked language modelling alongside a novel additional objective for inducing multilingual taxonomical ESCO relations. The authors assessed the performance of ESCOXLM-R on 6 sequence labelling and 3 classification tasks across 4 languages. The researchers asserted that ESCOXLM-R surpassed XLM-Roberta-Large "on job-related downstream tasks in 6 out of 9 datasets, particularly when the task was relevant to the ESCO taxonomy and context was important" (Zhang, Goot, & Plank, 2023).

[3] The researchers referred to the Cohen's kappa performance measures as percentages, but it is not clear why given Cohen's kappa is a coefficient ranging from -1 and 1.

3 RESEARCH DESIGN

3.1 ISCO-08 classification model overview

For the development of a multilingual ISCO classification model, we used several open-source machine learning software libraries (Molino, Dudin, & Miryala, 2019; Wolf, et al., 2020) alongside Facebook AI's multilingual sentence encoder XLM-RoBERTa (XLM-R), initially proposed in 2018 (Conneau, et al., 2018) and publicized in 2021. In short, XLM-R is a transformer-based encoder-decoder language model pre-trained on 2.5TB of text data from the Common Crawl corpus encompassing 100 languages. The Facebook AI researchers illustrated that the model surpassed existing models on a variety of cross-lingual understanding benchmarks, such as XNLI (Conneau, et al., 2018), where the objective is to ascertain textual entailment between a premise and hypothesis and MLQA (Lewis, Oğuz, Rinott, Riedel, & Schwenk, 2019), and where the goal is to extract a span of text as an answer to a question from a contextual passage.

We sourced data collected by the ICILS 2013[4] and ICILS 2018 student questionnaire instruments which contained student descriptions of the jobs held by their parents in response to questions about parental occupation. As a standard process in the occupation coding procedure employed by ICILS, these descriptions were coded with 4-digit ISCO-08 codes by trained occupation coders from the countries participating in ICILS. The descriptions of parental occupations served as the text feature (text to be classified) and the 4-digit ISCO-08 codes were the reference labels. Hence, we can refer to the task as a single label multiclass classification problem.

3.2 Data and data processing

Data

Classified text data in this research were sourced from ICILS 2013, ICILS 2018[5], and the International Labor Organization (ILO).

The ILO data comprised a list of 7,019 job titles in English with corresponding 4-digit ISCO-08 unit group codes. We used OpenAI's GPT-4 model to translate the job titles from English into 11 of the ICILS languages, which resulted in a total of 72,003 records.

The source data from ICILS 2013 comprised 6,325 student records. The ICILS 2018 source data comprised 39,156 student records from 13 countries, representing 12 different languages. In ICILS, students were asked to write a job title and a short summary of the tasks in that job for two parents[6] (or one parent where applicable). Students had to give the title

[4] Data from ICILS 2013 included only samples from Australia.
[5] ICILS 2013 data were included from Australia only; ICILS 2018 data were included from all participating countries (excluding benchmarking participants).
[6] Countries were able to adapt these terms to 'parent' and/or 'guardian' or 'mother' or 'father'.

and summary of the present job for each parent or for the last job of a parent who was not currently working in a paid job.

As part of ICILS, each pair of responses (title and description) were coded according to the international standard by professional human coders from the respective countries with 10% percent of the samples coded by a second person for measuring inter-rater reliability. Hence, all text response pairs had corresponding human-assigned ISCO codes.

Data processing method

The data processing begins with the creation of 'engineered features' that are used in the LLM dataset configuration and LLM training. The features were:

- the 2-digit language identification code of language of testing for the student;
- the 3-digit country identification code for the student;
- the job title and description for each parent provided by each ICILS student; and
- the human generated ISCO code for that record and the ILO occupation group name for that ISCO code.

These engineered features form the core of the *ICILS Multilingual ISCO-08 Parental Occupation Corpus Dataset*: a private[7] git repository with the dataset structure used in the research that can be explored online and analyzed using the Hugging Face Datasets Python package (see Appendix D: Supplementary digital resources).

Below is a summary of the key processes used to create the engineered features for the ICILS Multilingual ISCO-08 Parental Occupation Corpus:

Data validation, standardization, and preparation

- Verification that ISCO codes and metadata labels used in ICILS 2013 and ICILS 2018 corresponded with those from ILO (see Appendix E: Supplementary digital resources for Python notebooks).
- Assignment of the 3-digit ISO country code 2-digit ISO language code to the ICILS 2013 records (to be consistent with the ICILS 2018 records).
- Preparation of the ICILS reliability scores (the second coder ISCO allocations and corresponding text responses) for inclusion in the main dataset.

Feature engineering and data cleaning

- Feature engineering[8]: Job titles and descriptions of current and previous jobs, along with the corresponding ISCO codes, were merged into a single data table with unified text features (JOB and DUTIES) and categorical features (ISCO, ISCO_REL). The JOB

[7] The dataset is securely hosted and is accessible only to IEA organization members.
[8] This process involves creating new features (variables) from existing ones to better capture the information relevant for the analysis or predictive modeling.

and DUTIES features were concatenated to form a single text feature named JOB_DUTIES[9].

- Occupation titles from an ILO ISCO datafile were joined using the ISCO code as a new feature named ISCO_TITLE. A description corresponding to the five ICILS auxiliary codes was also created and joined using the ISCO feature. ISCO and ISCO_TITLE were concatenated to form an additional categorical feature named ISCO_CODE_TITLE and was used as the reference label in model training.
- Exclusions: Records were dropped if ISCO was fewer than 4 characters[10] or had a value denoting missing ('9999') or JOB_DUTIES was an empty string.

Table 3.1 shows the source variables and the features engineered for machine learning tasks derived from those variables.

[9] While it is possible to use multiple text features with tied weights, this increases the memory overhead needed for model training and had no positive effect on model performance in our experiments.

[10] Some of these records were double coded and in disagreement, dropping these records improved the quality of the data for model training. However, by dropping these records, the inter-rater reliability for the ICILS data used in this research may differ slightly from those reported in ICILS.

Table 3.1: ICILS database variables and engineered features for machine learning

Variable name	Variable description	Engineered feature name	Engineered feature type	Engineered feature description
CNTRY[ab]	3-digit ISO country code	COUNTRY	Categorical	3-digit ISO country code
LANGUAGE[c]	2-digit ISO language code	LANGUAGE	Categorical	2-digit ISO language code
IS1G07A[a] IS2G07A[b]	Parent 1 current job title (if currently working).	JOB	Text	Parent 1 or parent 2 current job title or previous job title from ICILS 2013 or ICILS 2018.
IS1G07C[a] IS2G07B[b]	Parent 1 previous job title (if not currently working).			
IS1G10A[a] IS2G11A[b]	Parent 2 current job title (if currently working).			
IS1G10C[a] IS2G08B[b]	Parent 2 previous job title (if not currently working).			
IS1G07B[a] IS2G08A[b]	Parent 1 current job description (if currently working).	DUTIES	Text	Parent 1 or parent 2 current job description or previous job description from ICILS 2013 or ICILS 2018.
IS1G07D[a] IS2G08B[b]	Parent 1 previous job description (if not currently working).			
IS1G10B[a] IS2G12A[b]	Parent 2 current job description (if currently working).			
IS1G10D[a] IS2G12B[b]	Parent 2 previous job description (if not currently working).			
ISCO_M[a] ISCO1[b]	Parent 1 ISCO code assigned by coder 1.	ISCO	Categorical	Parent 1 or parent 2 ISCO code assigned by coder 1 from ICILS 2013 or ICILS 2018.
ISCO_F[a] ISCO2[b]	Parent 2 ISCO code assigned by coder 1.			
ISCO_M_REL[a] ISCO1_REL[b]	Parent 1 ISCO code assigned by coder 2 (if double coded).	ISCO_REL	Categorical	Parent 1 or parent 2 ISCO code assigned by coder 2 (if double coded) from ICILS 2013 or ICILS 2018.
ISCO_F_REL[a] ISCO2_REL[b]	Parent 2 ISCO code Assigned by coder 2 (if double coded).			
ISCO Title EN[d]	ILO ISCO occupation group name	ISCO_TITLE	Categorical	ILO ISCO occupation group name
English Title[d]	ILO ISCO job titles examples	JOB	Text	ILO ISCO job titles examples
ISCO-08	ISCO-08	ISCO	Categorical	ISCO code assigned to English Title
JOB_DUTIES[c]	Combined JOB and DUTIES features	JOB_DUTIES	Text	Combined JOB and DUTIES features
ISCO_CODE_TITLE[c]	Combined ISCO and ISCO_TITLE features	ISCO_CODE_TITLE	Categorical	Combined ISCO and ISCO_TITLE features

[a] Source = ICILS 2013
[b] Source =ICILS 2018
[c] Source = Created for feature engineering
[d] Source = ILO

Below is a descriptive summary of the data after processing:

- Total records: 58,211 (containing student response and corresponding ISCO code)
- Unique ISCO/ICILS codes: 428 ISCO codes and 5 ICILS auxiliary codes
- Unique languages: 12

- Most frequently occurring occupation codes: 9705 (vague response), 9701 (home duties), 5223 (shop sales assistants), 2221 (nursing professionals), and 8332 (heavy vehicle and lorry drivers) (3.99% to 1.87% of the dataset)
- Most frequently used languages: English (29.42%), Spanish (14.14%), Portuguese (7.45%), French (7.29%), and Russian (6.29%).
- Double coded responses in agreement (ISCO == ISCO_REL): 12,906 (22.17%) records, indicating direct match between reported and validated ISCO codes.
- Single coded responses (ISCO_REL is blank, "9998", or "9999"): 43,740 (75.14%) entries, where the reliability of the ISCO code is uncertain.
- Double coded responses with discrepancies: (ISCO_REL ≠ "9998", "9999", and ≠ ISCO): 1,565 (2.68%) entries, indicating a discrepancy between coder 1 and coder 2.

3.3 Structuring the data for model training and evaluation

Model training

The process of training a model involves a series of trials utilizing different divisions of data from a larger dataset. The divided data are organized into distinct groups known as *splits*.

The *training split* includes data that the model initially learns from. These data help the model understand the associations between text responses and various occupation groups.

The *validation split* comprises data that are used to periodically evaluate the model's performance so that the parameters[11] that govern the model's learning strategy can be optimized.

The *test split* consists of data that are not used during the training phase. Instead, it is used after the training is complete to assess how well the model performs, providing a measure of its effectiveness in real-world scenarios.

Balancing the ICILS samples for inclusion in the splits

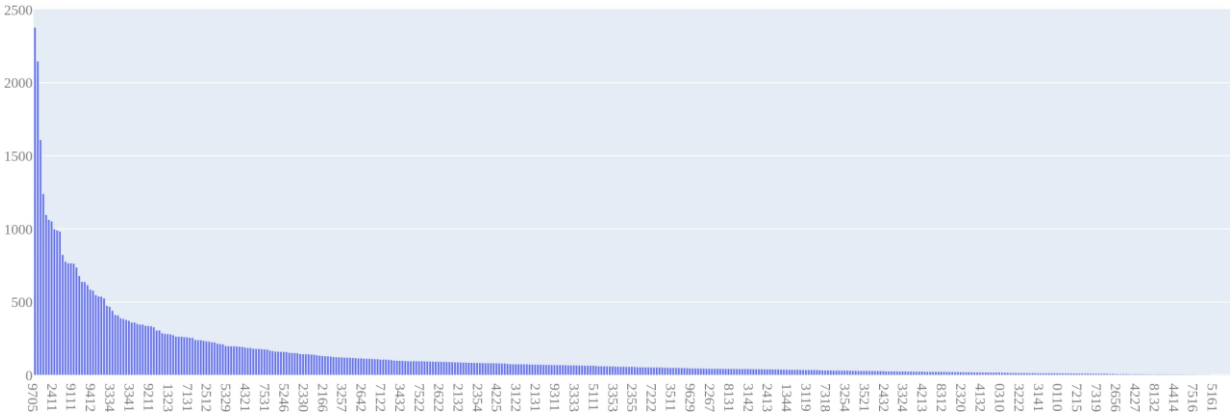
If data are randomly allocated to the training or validation splits, there is a risk that all records pertaining to a specific occupation group are allocated to the test split and not represented in the training split. In such cases, the model's evaluation would (unfairly) include records from occupation groups unknown by the model, resulting in incorrect classifications and poorer reported performance. Furthermore, overly represented classes can cause the model to overfit on a small number of classes which does not generalize well to new data.

The distribution of 58,211 records across the ISCO occupation unit groups and languages showed a significant imbalance. The maximum imbalance ratio (Tarekegn, Giacobini, & Michalak, 2021) is the ratio of the most common class against the rarest one. For the ICILS

[11] Often called hyperparameters. Examples include learning rate, batch size, maximum sequence length, and the number of epochs (complete passes over all records in the training split).

data, the most common occupation code was the ICILS auxiliary code 9705 (Vague responses) and the rarest was the ISCO code 6224 (Hunters and Trappers). The maximum imbalance ratio in the dataset was 2327:1.

Figure 3.1: Distribution of samples across ISCO groups before balancing



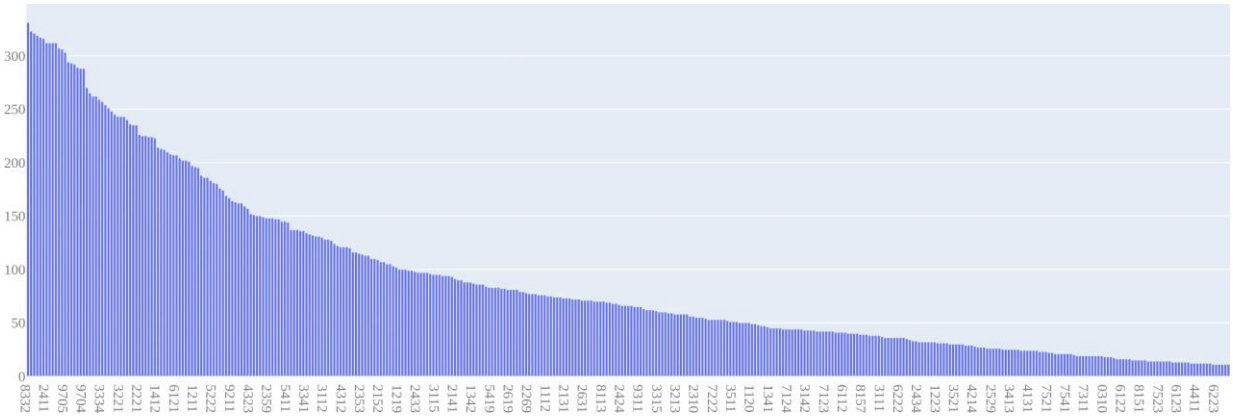
To address these problems, we developed a generalizable sampling strategy for occupation coded data that aimed to create a balanced and diverse dataset that maximizes the quality of data for training a language model. The dataset to be sampled for training and evaluating the model was first refined with a view to address any detectable issues of imbalance. This process included the following:

- If an occupation group was represented by only one record, this record was excluded (effectively excluding the class from the training process).
- Records where ISCO == ISCO_REL were prioritized for selection as they are likely to be the most accurate.
- For records where ISCO_REL is unknown ("9998", "9999", or blank), we used stratified sampling to ensure that the allocation of records to occupation groups reflected their proportional representation within the dataset.
- Records where there was a known discrepancy between ISCO and ISCO_REL, were deprioritized for selection. Stratified sampling was applied to include a small proportion of these samples, ensuring they did not dominate any specific ISCO group.
- We allowed the oversampling of underrepresented groups and undersampling of overrepresented groups to balance the proportions of occupation groups represented across the dataset.
- We used equal proportions of occupation groups within each of the three splits (training, validation and test).

The balanced dataset comprised 35,178 records. 28,142 records were allocated to the training split and 3,518 were allocated to each of the validation and test splits. Records containing double-coded responses represented 10% of each split. The average inter-rater agreement of these records was 99% within each split. The distribution of records across

occupation groups is illustrated in Figure 3.2. This final dataset serves as the cornerstone for all subsequent model trainings and evaluations conducted in this research.

Figure 3.2: Balanced distribution of ISCO groups in combined training, validation and test splits



Summary of final datasets used in model training and validation

Each record used in model trainings included the engineered features (see Table 3.1), JOB_DUTIES (the combined text describing the job title and the description of the duties), and ISCO_CODE_TITLE (the ILO ISCO code and group name).

Three different multilingual models were generated. These models were generated using data from all 12 languages available in the ICILS data. The model training differed according to the source data (ICILS, ILO, or both) used to generate the engineered features (Table 3.2).

Table 3.2: Summary of data sources corresponding to the three multilingual trained models

Model name	Training dataset (training & validation splits)	Evaluation dataset (test split)
Model 1	ICILS	ICILS
Model 2	ILO	ILO
Model 2	ILO	ICILS
Model 3	ICILS+ILO	ICILS+ILO
Model 3	ICILS+ILO	ICILS

In addition, 12 monolingual models were generated using the language-specific subsets of the ICILS data. These 12 monolingual models did not use ILO data.

3.4 Model training and evaluation process

Data used in all model training used a data division of training split (80%), validation split (10%), and testing split (10%).

The model training and evaluation process did not vary across the models. The only difference across the different model training and evaluation exercises was the source data used[12]. This meant that differences in the evaluation results across models can be interpreted as resulting from differences in the data used to generate each model rather than differences in the configuration and steps used to generate each model.

The model training and evaluation process comprised the following broadly described steps:

1. The XLM-R language model was loaded.
2. The parameters of the training were configured (see Appendix D: Supplementary digital resources, *ISCO R&D model training report* for the complete list of the configured training parameters).
3. The training is initiated using the training split.
4. Once the training has been completed the model with the highest reported accuracy evaluated with the validation split from the training trials is then loaded.
5. The loaded model is then evaluated using a test split. The evaluation includes all model predictions for all examples given so that the evaluation metrics can be calculated.

[12] The only exception to this was Model 1 which was trained for more epochs than other models to determine when the performance of the model (on the period evaluations) stopped improving. This was used to configure the number of epochs for the other model trainings.

4 PROCEDURE FOR EVALUATING THE MODELS

4.1 Evaluation measures

The following evaluation measures were used to assess the performance of each model.

Accuracy

Accuracy answers the question, '*How often is the model right?*'.

Accuracy measures the proportion of all predictions that the model got right.

Accuracy is computed with:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Precision

Precision helps answer the question, '*What proportion of predictions pertaining to an occupation group were correct?*'.

Precision is the ratio of the number of correct predictions (pertaining to a specific occupation group) to the total number of predictions (pertaining to a specific occupation group).

While accuracy measures the overall correctness of the model, precision measures the correctness of the model for a given class.

Precision is computed with:

$$Precision = \frac{TP}{TP + FP}$$

This is the same as:

$$Precision = \frac{Number\ of\ correct\ predictions\ (pertaining\ to\ an\ occupation\ group)}{Total\ number\ of\ predictions\ (pertaining\ to\ an\ occupation\ group)}$$

Recall

Recall helps answer the question, '*What proportion of records that belong to a particular occupation group were correctly identified?*'.

Recall is the ratio of records that belong to an occupation group that were correctly predicted to belong to that occupation group.

While precision measures the correctness of the model for a specific class, recall measures how well the model can identify all records that belong to a particular occupation group.

Recall is computed with:

$$Recall = \frac{TP}{(TP + FN)}$$

F1 Score

The F1 Score is a measure that combines precision and recall into a single metric by taking their harmonic mean[13]. It provides a balance between precision and recall, offering insight into the overall efficiency of the model when both false positives and false negatives are crucial. The F1 Score is particularly useful when the distribution of class labels is uneven.

F1 Score is computed with:

$$F1 = \frac{2 * TP}{2 * (TP + FP + FN)}$$

This is the same as:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Micro, Macro, and Weighted Averages

These averages provide methods to aggregate precision, recall, and the F1 Score across multiple occupation groups, offering a more comprehensive evaluation of model performance.

Micro Average aggregates all predictions across all occupation groups to compute the average metric. In micro-averaging, all True Positives, False Positives, and False Negatives for each occupation group are summed up and then the average is calculated. This approach is beneficial when class imbalance is present, as it gives an equal weight to each prediction.

Macro Average computes the metric independently for each class and then takes the average (hence treating all classes equally). This method gives equal weight to each occupation group, not each prediction, which can be useful when you want to know how the model performs overall across the set of occupation groups. The Macro Average does not take the imbalance of records across occupation groups into account.

Weighted Average calculates the metric for each occupation group independently, but when taking the average, it weights each occupation group's metric by the number of correct predictions for each occupation group. This approach accounts for an imbalanced representation of occupation groups by giving more weight to groups with more records. It is particularly useful when you care more about the performance on the more dominant occupation groups.

[13] The harmonic mean is a type of average, specifically designed to find the mean of rates or ratios.

4.2 ISCO Hierarchical Accuracy Measure

The ISCO Hierarchical Accuracy Measure is a version of hierarchical precision (HP), recall (HR), and F-measure (HF) (Kiritchenko & Famili, 2005) that is tailored to the ISCO-08 hierarchical classification system (see Appendix D: Supplementary digital resources for the implementation of the measure used in the evaluations). The measure rewards more precise classifications that correctly identify an occupation's placement down to the specific unit group and applies penalties for misclassifications based on the hierarchical distance between the reference and assigned categories.

Hierarchical precision

Hierarchical precision evaluates the proportion of correctly predicted labels (including their ancestors) out of all *predicted labels* and their ancestors. It penalizes predictions that are hierarchically distant from the reference label more than those that are closer.

Hierarchical recall

Hierarchical recall assesses the proportion of correctly predicted labels (including their ancestors) out of all *reference labels* and their ancestors. It rewards predictions that cover the reference label and its hierarchical context, even if the label is not matched exactly.

Hierarchical F-measure (hF-measure)

Hierarchical F-measure combines hP and hR into a single score that balances precision and recall, providing a unified metric to evaluate the correctness of predictions within the hierarchical structure.

The hierarchical structure of the ISCO-08 classification scheme is organized into four levels, delineated by the specificity of their codes (Table 4.1). "Looking at the hierarchical structure of ISCO-08 from the top down, each of the ten major groups is made up of one or more sub-major groups, which in turn are made up of one or more minor groups. Each of the 130 minor groups is made up of one or more-unit groups. In general, each unit group is made up of several 'occupations' that have a high degree of similarity in terms of skill level and skill specialization. Each group in the classification is designated by a title and code number and is associated with a description that specifies the scope of the group." (International Labour Organization, 2024).

Table 4.1: Number of groups at each level of ISCO-08

Major Groups (1-digit)	Sub Major Groups (2-digits)	Minor Groups (3-digits)	Unit Groups (4-digits)
1 Managers	4	11	31
2 Professionals	6	27	92
3 Technicians and Associate Professionals	5	20	84
4 Clerical Support Workers	4	8	29

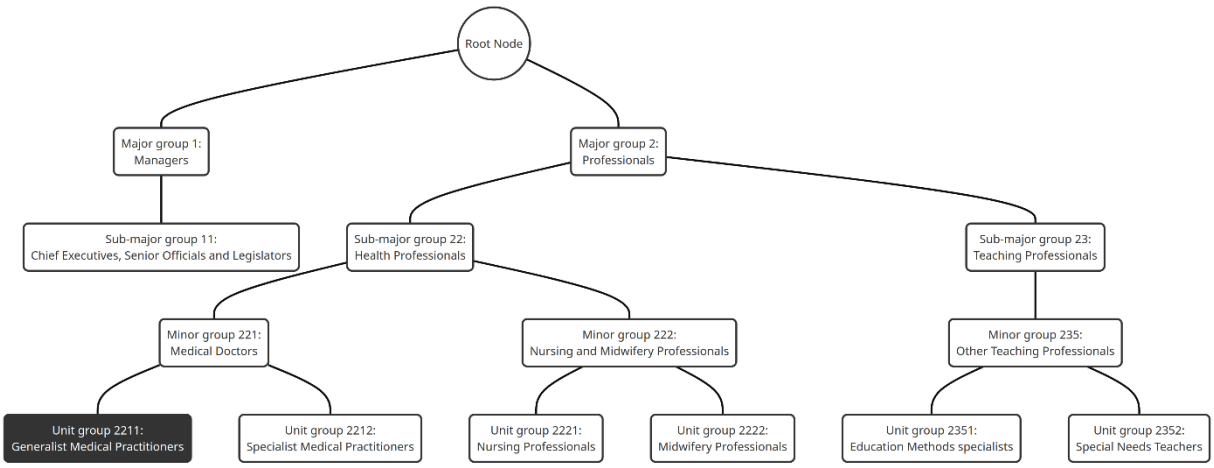
5 Service and Sales Workers	4	13	40
6 Skilled Agricultural, Forestry and Fishery Workers	3	9	18
7 Craft and Related Trades Workers	5	14	66
8 Plant and Machine Operators, and Assemblers	3	14	40
9 Elementary Occupations	6	11	33
0 Armed Forces Occupations	3	3	3
Total number of groups	**43**	**130**	**436**

In this context, the Hierarchical Accuracy Measure is specifically designed to evaluate classifications within this structured framework. It emphasizes the importance of precision in classifying occupations at the correct level of specificity:

- Major groups (1-digit codes): The broadest category that encompasses a wide range of occupations grouped based on their fundamental characteristics and role in the job market. Misclassification among different major groups implies a fundamental misunderstanding of the occupational role.
- Sub-major groups (2-digit codes): Provide a more detailed classification within each major group, delineating categories that share specific occupational characteristics. Errors within a major group but across different sub-major groups are less severe than those across major groups but are still significant due to the deviation from the correct occupational category.
- Minor groups (3-digit codes): Offer further granularity within sub-major groups, classifying occupations that require similar skill sets. Misclassifications within a sub-major group but across different minor groups are penalized, yet to a lesser extent than errors across sub-major or major groups, as they still fall within a broader correct occupational category.
- Unit groups (4-digit codes): The most specific level, identifying occupations that are most similar in terms of skill level and skill specialization. Misclassification within a minor group, between different unit groups, represents the least severe error, as it occurs within the context of closely related occupations.

The hierarchical accuracy measure rewards more precise classifications that correctly identify an occupation's placement down to the specific unit group level and applies penalties for misclassifications based on the hierarchical distance between the correct and assigned categories. This approach allows for a more nuanced evaluation of classification systems, acknowledging partially correct classifications while penalizing inaccuracies more severely as they deviate further from the correct level in the hierarchical structure.

Figure 4.1: Tree graph representation of the ISCO-08 hierarchical structure



For example, a misclassification into *Unit group 2211: Generalist medical practitioners* (Figure 4.1) when the correct category is *Unit group 2212: Specialist medical practitioners* should incur a lesser penalty than misclassification into *Unit group 2352: Special needs teachers*, as unit groups 2211 and 2212 are within the same minor group (*221: Medical doctors*) and share a closer hierarchical relationship compared to *Minor group 235: Other teaching professionals* which is in a different sub-major group (*23: Teaching professionals*).

The penalty is higher for errors that occur between more distant categories within the hierarchy:

1. Correct classifications at a more specific level (e.g., minor group 2211) are evaluated more favorably than classifications at a more general level within the same hierarchy (e.g., sub-major group 22), since the former indicates a closer match to the correct category.
2. Conversely, incorrect classifications at a more specific level are penalized more heavily than those at a more general level. For instance, classifying into *Minor group 222: Nursing and midwifery professionals* is considered worse than classifying into its parent category, sub-major group 22, if the correct classification is minor group 221, because the incorrect specific classification indicates a greater deviation from the correct category.

Misclassification among sibling categories (e.g., between different unit groups within the same minor group) is less severe than misclassification at a higher hierarchical level (e.g., between different sub-major groups).

The measure extends the concepts of precision and recall into a hierarchical context, introducing hierarchical precision (hP) and hierarchical recall (hR). In this framework, each sample belongs not only to its designated class but also to all ancestor categories in the hierarchy. This adjustment allows the measure to account for the hierarchical structure of the classification scheme, rewarding more accurate location of a sample within the hierarchy and penalizing errors based on their hierarchical significance.

The following is a passage from the original research paper with the equations for reference.

> *To calculate the hierarchical measure, we extend the set of real classes* $C_i = \{G\}$ *with all ancestors of class* G: $\overrightarrow{C_i} = \{B, C, E, G\}$.
>
> *We also extend the set of predicted classes* $C_i' = \{F\}$ *with all ancestors of class* F: $\overrightarrow{C_i'} = \{C, F\}$.
>
> *So, class C is the only correctly assigned label from the extended set:* $\left|\overrightarrow{C_i} \cap \overrightarrow{C_i'}\right| = 1$ *I.*
>
> *There are* $\left|\overrightarrow{C_i'}\right| = 2$ *assigned labels and* $\left|\overrightarrow{C_i}\right| = 4$ *real classes.*
>
> *Therefore, we get:*
>
> $$hP = \frac{\left|\overrightarrow{C_i} \cap \overrightarrow{C_i'}\right|}{\left|\overrightarrow{C_i'}\right|} = \frac{1}{2}$$
>
> $$hR = \frac{\left|\overrightarrow{C_i} \cap \overrightarrow{C_i'}\right|}{\left|\overrightarrow{C_i}\right|} = \frac{1}{2}$$
>
> *We can also combine the two values* hP *and* hR *into one hF-measure:*
>
> $$hF_\beta = \frac{(\beta^2 + 1) \cdot hP \cdot hR}{(\beta^2 \cdot hP + hR)}, \beta \in [0, +\infty)$$
>
> *(Kiritchenko & Famili, 2005, p. 5)*

5 EVALUATION RESULTS

Model 1 (ICILS data) predicted the exact same ISCO unit group as human coders across 12 languages for 63% of the test split samples (Table 5.1). By language (Table 5.3), accuracy ranged from 42% (sv) to 73% (ru).

Multilingual model results overall

Table 5.1: Summary of multilingual model overall accuracies and hF-measure scores

Model name	Training dataset (training & validation splits)	Evaluation dataset (test split)	Accuracy	hFβ
Model 1	ICILS	ICILS	63%	0.89
Model 2	ILO	ILO	92%	0.99
Model 2	ILO	ICILS	36%	0.94
Model 3	ICILS+ILO	ICILS+ILO	80%	0.93
Model 3	ICILS+ILO	ICILS	62%	0.95

Table 5.2: Multilingual models' accuracy, precision, and recall with ICILS test split

Model	support	accuracy	micro avg precision	micro avg recall	micro avg f1	macro avg precision	macro avg recall	macro avg f1	weighted avg precision	weighted avg recall	weighted avg f1	hP	hR	hFβ
Model 1 (icils)	3518	0.63	0.63	0.63	0.63	0.44	0.45	0.43	0.61	0.63	0.61	1.00	0.81	0.89
Model 2 (ilo)	3418	0.36	0.36	0.36	0.36	0.31	0.30	0.28	0.42	0.36	0.35	0.96	0.91	0.94
Model 3 (combined)	3518	0.62	0.62	0.62	0.62	0.48	0.47	0.46	0.62	0.62	0.60	0.98	0.92	0.95

Multilingual model results grouped by language

Table 5.3: Model 1 (ICILS) accuracy, hierarchical precision, recall and hF-measure score with ICILS test split grouped by language

Language	Accuracy	hP	hR	hFβ
da	0.56	0.84	0.78	0.81
de	0.61	0.68	0.74	0.71
en	0.62	0.95	0.82	0.88
es	0.67	0.89	0.84	0.87
fi	0.63	0.84	0.87	0.85
fr	0.57	0.81	0.77	0.79
it	0.60	0.79	0.78	0.78
kk	0.69	0.83	0.83	0.83
ko	0.69	0.84	0.79	0.81
pt	0.62	0.83	0.80	0.81
ru	0.73	0.80	0.80	0.80
sv	0.42	0.64	0.60	0.62
Average	**0.62**	**0.81**	**0.79**	**0.80**

The discrepancy between the hFβ for Model 1 in Table 1 and the average hFβ computed for predictions by language, as shown in Table 2, can be attributed to the nature of hFβ as a metric. The hFβ score is a harmonic mean of precision (hP) and recall (hR), with the β parameter determining the weight of recall in the calculation. It is designed to balance precision and recall, providing a single metric that can be used to evaluate the model's overall performance on a dataset. The hFβ score for Model 1 in Table 1 represents an aggregated measure of performance across the entire evaluation dataset. This aggregation considers the total counts of true positives, false positives, and false negatives across all instances, leading to a single hFβ score. In contrast, the average hFβ across languages involves calculating an hFβ for each language independently and then averaging these scores. The averaging process can mask disparities in performance across languages, especially if some languages have disproportionately high or low scores.

Table 5.4: Model 2 (ILO) accuracy, hierarchical precision, recall and hF-measure score with ILO test split grouped by language

Language	Accuracy	hP	hR	hFβ
da	0.95	0.98	0.98	0.98
de	0.93	0.99	0.99	0.99
en	0.97	0.99	0.98	0.99
es	0.97	0.99	0.98	0.99
fi	0.91	0.97	0.97	0.97
fr	0.95	0.98	0.98	0.98
it	0.95	0.99	0.99	0.99
kk	0.75	0.93	0.93	0.93
ko	0.89	0.96	0.97	0.96
pt	0.95	0.99	0.98	0.99
ro	0.95	0.97	0.99	0.98
Average	**0.92**	**0.98**	**0.98**	**0.98**

Table 5.5: Model 2 (ILO) accuracy, hierarchical precision, recall and hF-measure score with ICILS test split grouped by language

Language	Accuracy	hP	hR	hFβ
da	0.32	0.69	0.77	0.72
de	0.30	0.56	0.65	0.60
en	0.38	0.89	0.83	0.86
es	0.32	0.72	0.78	0.75
fi	0.34	0.68	0.73	0.71
fr	0.35	0.67	0.71	0.69
it	0.34	0.67	0.70	0.69
kk	0.26	0.60	0.70	0.65
ko	0.29	0.64	0.78	0.70
pt	0.39	0.73	0.76	0.74
ru	0.34	0.59	0.69	0.63
sv	0.42	0.45	0.50	0.48
Average	**0.34**	**0.66**	**0.72**	**0.68**

Table 5.6: Model 3 (ICILS+ILO) accuracy, hierarchical precision, recall and hF-measure score with ICILS+ILO test split grouped by language

Language	Accuracy	hP	hR	hFβ
da	0.83	0.97	0.95	0.96
de	0.87	0.98	0.98	0.98
en	0.75	0.99	0.95	0.97
es	0.83	0.98	0.97	0.97
fi	0.80	0.94	0.96	0.95
fr	0.83	0.95	0.96	0.96
it	0.83	0.96	0.96	0.96
kk	0.70	0.97	0.93	0.95
ko	0.79	0.94	0.95	0.94
pt	0.80	0.96	0.94	0.95
ro	0.93	0.98	0.98	0.98
ru	0.74	0.82	0.80	0.81
sv	0.42	0.59	0.60	0.60
Average	**0.78**	**0.93**	**0.92**	**0.92**

Table 5.7: Model 3 (ICILS+ILO) accuracy, hierarchical precision, recall and hF-measure score with ICILS test split grouped by language

Language	Accuracy	hP	hR	hFβ
da	0.54	0.79	0.82	0.81
de	0.61	0.68	0.77	0.72
en	0.62	0.93	0.90	0.91
es	0.65	0.87	0.85	0.86
fi	0.58	0.80	0.83	0.82
fr	0.56	0.78	0.78	0.78
it	0.57	0.76	0.80	0.78
kk	0.68	0.84	0.85	0.84
ko	0.64	0.83	0.79	0.81
pt	0.59	0.79	0.82	0.80
ru	0.74	0.82	0.80	0.81
sv	0.42	0.59	0.60	0.60
Average	**0.60**	**0.79**	**0.80**	**0.79**

Monolingual model results overall

Table 5.8: Summary of monolingual model overall accuracies and hF-measure scores with language specific test splits

Language	Accuracy	hFβ
da	0.48	0.77
de	0.46	0.66
en	0.60	0.86
es	0.64	0.82
fi	0.38	0.70
fr	0.47	0.73
it	0.46	0.70
kk	0.57	0.73
ko	0.59	0.72
pt	0.51	0.73
ru	0.65	0.74
sv	0.05	0.26
Average	**0.49**	**0.70**

6 DISCUSSION AND CONCLUSION

The research leverages the latest advancements in natural language processing and machine learning, specifically the capabilities of multilingual large language models like XLM-RoBERTa. This approach addresses the challenge of accurately coding parental occupation data across multiple countries and languages, thereby contributing to the consistency and reliability of derived SES measures in ILSAs.

Three multilingual models were developed and evaluated for ISCO classification: Model 1 utilizing ICILS data, Model 2 employing ILO data, and Model 3 combining ICILS and ILO data. Model 1 achieved an accuracy of 63% on the ICILS test split, indicating its capability to match human coder accuracy across 12 languages for over half of the test samples. Model 2, trained exclusively on ILO data, displayed a remarkable accuracy of 92% on the ILO test split, demonstrating its high precision in a controlled dataset. However, its performance dropped to 35% when evaluated against the ICILS test split, highlighting challenges in cross-dataset application and likely attributable to the very specific and narrow range of descriptions in the ILO test split in comparison to the wide variety of descriptions in the ICILS dataset. Model 3, which combined both datasets, showed improved versatility with an accuracy of 80% on the combined test split and 62% on the ICILS test split alone.

In addition, 12 monolingual models were designed to evaluate their performance on language-specific subsets of the ICILS data, enabling a direct comparison with the multilingual model using the same test splits. The multilingual model surpassed the performance of each monolingual model in every language group, with an average accuracy of 62% compared to the 49% average accuracy achieved by the monolingual models. This outcome underscores the multilingual model's enhanced ability in cross-lingual understanding. An important implication of this result is that the general guideline for the number of samples per class, adjusted for language, may be more adaptable. This means that descriptions of parental occupation belonging to an ISCO occupation group in one language can improve the model’s accuracy in classifying similar descriptions in another language.

The research introduced hierarchical precision and recall metrics specifically designed to use the ISCO-08 taxonomy for extending classes with their ancestor classes, acknowledging the hierarchical nature of ISCO classification.

The hierarchical nature of ISCO means that some occupation groups are closer to each other than others, depending on their position within the hierarchy. The 'distance' between two occupation groups can be thought of in terms of how many levels up in the hierarchy one must go before finding a common parent. For example, two occupations within the same minor group are closer (less 'distance') than two occupations in different major groups.

The hF-measure computes a score based on hierarchical precision (hP) and hierarchical recall (hR), incorporating the concept of 'distance' in its evaluation. Hierarchical precision (hP) reflects the proportion of correctly predicted occupation groups (and their ancestors in the

hierarchy) out of all predicted occupations groups (and their ancestors). It penalizes predictions that are hierarchically distant from the reference more than those that are closer. Hierarchical recall (hR) indicates the proportion of correctly predicted occupation groups (and their ancestors) out of all reference occupation groups (and their ancestors). It rewards predictions that cover the correct occupation group and its hierarchical context, even if the occupation group is not an exact match. Hierarchical F-measure (hF-measure) combines hP and hR into a single score that balances precision and recall, providing a unified metric to evaluate the extent of partial correctness within the hierarchical structure of ISCO.

When evaluated with the ICILS test split, the hFβ scores for Model 1 (0.89) and Model 3 (0.95) suggest that combining datasets can measurably enhance the model's ability to make partially correct predictions within the ISCO hierarchy.

Discrepancies can be observed between accuracy and hFβ scores across languages, where a model may show higher accuracy for one language, but lower hFβ scores compared to another. These discrepancies point to the complexity of the classification task, especially given the hierarchical structure of ISCO and the linguistic nuances inherent in occupation descriptions across different languages. An example of this discrepancy for Model 1 (ICILS data) is German (de) vs. Danish (da): where accuracy for German was higher (61%) compared to Danish (56%), suggesting that the model was more adept at exactly matching the ISCO codes for occupations described in German. However, the hF-measure score for German (0.71) was notably lower than that of Danish (0.81), which indicates a discrepancy in the model's ability to correctly classify occupations within the broader hierarchical context. This discrepancy could suggest that, while the model struggles with exact matches in Danish, it simultaneously faces challenges in recognizing broader occupational categories for German.

A similar example for Model 2 (ILO data) is English (en) vs. Kazakh (kk): Model 2, trained on ILO data, demonstrated a high accuracy in English (97%) compared to a significantly lower accuracy in Kazakh (75%) which could imply a problem with the quality of the machine translated samples. However, the hF-measure score for Kazakh (0.93) was robust, indicating that despite the lower accuracy, the model was quite effective in making partially correct classifications within the occupational hierarchy. This scenario underscores the importance of hierarchical precision and recall metrics in assessing model performance, especially in contexts where an exact match might be challenging to achieve.

For Model 3 (ICILS+ILO Data) the accuracy in English was higher compared to Swedish. Specifically, while English showed a respectable accuracy (75%), Swedish accuracy was significantly lower (42%). However, the hF-measure score for Swedish (0.60) was not as high as one might expect given the lower accuracy, which implies that the model's performance in Swedish was not only poor in terms of exact matches but also in making hierarchical classifications. This is likely due the small number of samples available in Swedish within the ICILS data.

These examples of discrepancies highlight several important aspects of model performance in multilingual contexts:

- Linguistic nuances and contextual understanding: The model's performance varies across languages, possibly due to differences in how occupational descriptions are phrased or understood in different linguistic and cultural contexts. Languages with more complex syntax or less training data might see lower accuracy and hF-measure scores. Languages less represented in the original pre-training of the XLM-R model might also be a factor.
- Hierarchical classification challenges: The discrepancies between accuracy and hF-measure scores indicate that correctly predicting an occupation's exact ISCO code is challenging. However, the model may still recognize the broader occupational category, emphasizing the value of hierarchical metrics for understanding partial correctness.
- Implications for model training and evaluation: These findings suggest that model evaluation should consider both accuracy and hierarchical metrics to gain a comprehensive understanding of model performance. Additionally, they underscore the importance of incorporating diverse linguistic and cultural contexts into the training data to improve model generalization across languages.

The observed discrepancies between accuracy and hF-measure scores across languages reveal the nuanced challenges of multilingual occupation classification. They stress the need for a balanced approach to model evaluation that considers both exact and hierarchical correctness to truly gauge a model's effectiveness in real-world applications.

While the hierarchical measures were primarily used to evaluate model performance, we can extend their use as a measure of coder reliability that complements inter-rater agreement by conceptualizing the hF-measure as a quantification of disagreement (or partial agreement) between coders' classifications within the hierarchical structure of ISCO. A 70% agreement implies a relatively high level of consensus among coders on the exact match of the ISCO codes, without accounting for the nature of partial correctness. In contrast, an hF-measure of coder agreement is expected to exceed the direct match agreement rate because it recognizes partial correctness. For example, when two coders classify an occupation into different occupation groups, the hF-measure can quantify the extent of their disagreement by considering the hierarchical 'distance' between the chosen groups.

Hence, a minor disagreement is one when one coder classifies a job as a 'General Nurse' and another as a 'Specialized Nurse,' because these are within the same minor group but different unit groups. The hierarchical 'distance' is small, so the disagreement impacts the hF-measure less severely. However, this would not contribute to an exact match inter-rater agreement but can still result in a relatively high hF-measure score because the classifications are hierarchically close. Conversely, if the second coder classifies the job as 'Software Developer',

this is a major disagreement because the hierarchical 'distance' is greater, impacting the hF-measure more severely.

Therefore, the hF-measure can offer a nuanced interpretation of coder agreement and disagreement, explicitly accounting for the hierarchical nature of occupation classifications. It can enable researchers to quantify disagreements not just as binary (agree/disagree) but in terms of the relative closeness of the chosen categories, providing insights into the coders' understanding of the classification scheme and the clarity of occupation descriptions. This approach could be particularly valuable in complex hierarchical schemes like ISCO, where recognizing degrees of agreement and the specifics of disagreement can inform improvements in coding procedures, training, and the development of classification tools.

In summary, the exploration of multilingual LLMs for classifying parental occupations within the ISCO framework marks a significant advance in integrating artificial intelligence into education research. This research has demonstrated these models' potential to improve occupation coding accuracy and efficiency across multiple languages and countries while also highlighting the complexity and challenges of this task. By employing a careful and structured methodology—developing hierarchical evaluation metrics for both model performance and coder agreement—and a commitment to transparency and reproducibility, this study offers valuable insights into AI models' nuanced performance in recognizing and categorizing occupational data. However, our findings also emphasize the need for a critical perspective on current technologies' limitations, the importance of human oversight, and further research to address the variability and cultural specificity of occupation descriptions. As we move forward, it is evident that leveraging technology to improve occupation coding procedures is both promising and challenging. Embracing this dual reality will be crucial for driving innovation, ensuring the relevance and reliability of socioeconomic status measures in international large-scale assessments (ILSAs), and ultimately contributing to a more inclusive and precise understanding of global educational equity.

7 APPENDICES

7.1 Appendix A: Terms and abbreviations

Training split: This is the data on which the model is trained. The model learns to make predictions based on the input data and the corresponding labels provided in this subset.

Validation split: The validation set is used to evaluate the model during the training process, but it is not used for training the model directly. It provides a measure of the model's performance on data it has not seen before, which helps in tuning the model's hyperparameters and making decisions about which model versions to save or further refine. The validation set plays a crucial role in preventing the model from overfitting to the training data.

Test split: After a model has been trained and validated, the test set is used to evaluate its performance. Like the validation set, the test set consists of unseen data that was not used during the training process. However, the key difference is that the test set is used only after the model has been fully trained and validated. It provides a final measure of how well the model is expected to perform in real-world scenarios or on data it has never seen before.

Class: A category within a classification system, here referring to an occupation group according to the ISCO-08 standard.

Label: A unique identifier for a class, in this context, a 4-digit code representing an occupation group under the ISCO-08 classification.

Text: A student's description of a parent's job.

Reference Label: The 4-digit ISCO-08 code assigned by a human expert to classify the described job in the text.

Sample: An instance from the dataset that includes both the text (job description) and its corresponding reference label.

Predicted label: The outcome of the model's classification, assigning a 4-digit ISCO-08 code to the given text.

True positive (TP): Occurs when the model accurately predicts a sample's class, with the predicted label matching the reference label.

True negative (TN): For a given class, a true negative is when the model correctly predicts that a sample does not belong to that class. However, in multi-class settings like ISCO-08, true negatives per class are less directly applicable. Hence for multi-class settings the concept of TN is broadened to include all correct classifications outside a particular class.

False positive (FP): Occurs when the model incorrectly predicts that a sample belongs to a certain class. For instance, the model assigns a sample a predicted label that does not match the reference label, and the *incorrect label* is the class of interest.

False Negative (FN): Happens when the model fails to identify that a sample belongs to a specific class. In other words, the model's predicted label does not match the reference label, and the *reference label* is the class of interest, indicating that the model missed identifying the sample's true class.

7.2 Appendix B: ISCO code validation

The data validation processed entailed a straightforward verification to ascertain that the ISCO codes within the ICILS data table align with the ISCO codes in the ILO ISCO data table. The outcome of this validation procedure is a compilation of ICILS ISCO codes that do not find a match in the ILO ISCO data table, along with the frequency of their occurrence. The initial five rows of the non-matching codes are shown in Figure 7.1.

Figure 7.1: ICILS ISCO codes not found in ILO ISCO

	ISCO object 0000 0.7% 1000 0.7% 143 others 98.6%	Description object Protective services workers 1.4% Assemblers 1.4% 140 others 97.2%	Count float64 2.0 - 3108.0
0	0000	Armed forces occupations	144
1	1000	Managers	198
2	1100	Chief executives, senior officials and legislators	120
3	1110	Legislators and senior officials	2
9	1200	Administrative and commercial managers	20

The auxiliary codes (9701 to 9705) or codes denoting missing values (9999 and 9998) are anticipated. Nonetheless, there are 140 other codes where the ICILS ISCO code does not correspond with the ILO ISCO code. Within the ICILS data, ISCO codes at level 1, 2, and 3 are documented as a 4-digit code by appending trailing zeros, up to the fourth digit. For instance, the ICILS ISCO code 1200 (Administrative and commercial managers) aligns with the ILO ISCO code 12 (Administrative and commercial managers). The correspondence between the ICILS and ISCO codes is one-to-one; however, the reported level of detail may vary.

The conventions for ICILS 2013 and ICILS 2018 diverge. In ICILS 2013, the protocol is to employ the ILO codes without alteration. Hence, the ICILS 2013 ISCO code 12 (Administrative and commercial managers) corresponds to the ILO ISCO code 12 (Administrative and commercial managers), and there are no instances of 1200. Consequently, the schemas between the ICILS 2013 and ICILS 2018 datasets are not directly comparable. To reconcile the disparity and enable merging the two datasets with consistent coding schemes, we mapped the ICILS outlier codes to equivalent ISCO codes by applying a string replacement method with a regular expression to replace digits from the 4th, 3rd, and 2nd positions if they were zero.

Table 7.1: Recoding ISCO codes process

0000 becomes 0	1000 becomes 1
0100 becomes 01	1100 becomes 11
0200 becomes 02	...
0300 becomes 03	9620 becomes 962

7.3 Appendix C: Class report

Table 2: Precision, recall, and f1 scores for Model 1 (ICILS data) and Model 3 (ICILS+ILO data) evaluated with the ICILS data test split

ISCO-08 unit group	M1 precision	M1 recall	M1 f1-score	M3 precision	M3 recall	M3 f1-score	support
0110 Commissioned Armed Forces Officers	0.50	1.00	0.67	0.00	0.00	0.00	1
0310 Armed Forces Occupations, Other Ranks	0.00	0.00	0.00	0.00	0.00	0.00	2
1111 Legislators	0.50	0.50	0.50	0.40	1.00	0.57	2
1112 Senior Government Officials	0.25	0.43	0.32	0.60	0.43	0.50	7
1114 Senior Officials of Special-interest Organizations	0.00	0.00	0.00	0.00	0.00	0.00	2
1120 Managing Directors and Chief Executives	0.25	0.20	0.22	0.33	0.20	0.25	5
1211 Finance Managers	0.74	0.67	0.70	0.72	0.62	0.67	21
1212 Human Resource Managers	0.67	0.75	0.71	0.67	0.75	0.71	16
1213 Policy and Planning Managers	0.00	0.00	0.00	0.00	0.00	0.00	4
1219 Business Services and Administration Managers Not Elsewhere Classified	0.30	0.27	0.29	0.50	0.36	0.42	11
1221 Sales and Marketing Managers	0.64	0.67	0.65	0.73	0.52	0.61	21
1222 Advertising and Public Relations Managers	0.00	0.00	0.00	0.00	0.00	0.00	2
1223 Research and Development Managers	0.00	0.00	0.00	0.25	0.33	0.29	3
1311 Agricultural and Forestry Production Managers	0.63	0.71	0.67	0.71	0.71	0.71	7
1312 Aquaculture and Fisheries Production Managers	0.00	0.00	0.00	0.00	0.00	0.00	1
1321 Manufacturing Managers	0.50	0.47	0.49	0.50	0.53	0.51	19
1322 Mining Managers	0.57	0.80	0.67	0.44	0.80	0.57	5
1323 Construction Managers	0.50	0.67	0.57	0.45	0.62	0.52	21
1324 Supply, Distribution and Related Managers	0.42	0.48	0.44	0.30	0.38	0.33	21
1330 Information and Communications Technology Service Managers	0.25	0.67	0.36	0.25	0.67	0.36	3
1341 Child Care Service Managers	0.50	0.25	0.33	0.60	0.75	0.67	4
1342 Health Service Managers	0.38	0.63	0.48	0.47	0.88	0.61	8
1343 Aged Care Service Managers	0.33	0.50	0.40	0.40	0.50	0.44	4
1344 Social Welfare Managers	0.00	0.00	0.00	0.17	0.25	0.20	4
1345 Education Managers	0.63	0.71	0.67	0.68	0.71	0.70	21
1346 Financial and Insurance Services Branch managers	0.82	0.69	0.75	0.77	0.77	0.77	13
1349 Professional Services Managers Not Elsewhere Classified	0.50	0.14	0.22	0.00	0.00	0.00	7
1411 Hotel Managers	1.00	0.75	0.86	0.83	0.63	0.71	8
1412 Restaurant Managers	0.67	0.73	0.70	0.63	0.68	0.65	22
1420 Retail and Wholesale Trade Managers	0.45	0.45	0.45	0.35	0.64	0.45	11
1431 Sports, Recreation and Cultural Centre Managers	0.43	0.75	0.55	0.45	0.63	0.53	8
1439 Services Managers Not Elsewhere Classified	0.30	0.27	0.29	0.33	0.09	0.14	11
2111 Physicists and Astronomers	0.00	0.00	0.00	1.00	1.00	1.00	1
2112 Meteorologists	0.00	0.00	0.00	0.00	0.00	0.00	1
2113 Chemists	0.50	0.50	0.50	0.50	0.50	0.50	4
2114 Geologists and geophysicists	0.67	1.00	0.80	0.67	1.00	0.80	4
2131 Biologists, Botanists, Zoologists and Related Professionals	0.63	0.71	0.67	0.56	0.71	0.63	7
2132 Farming, Forestry and Fisheries Advisers	0.86	0.75	0.80	1.00	0.75	0.86	8
2133 Environmental Protection Professionals	0.50	0.60	0.55	0.50	0.80	0.62	5
2141 Industrial and Production Engineers	0.57	0.44	0.50	0.36	0.44	0.40	9

ISCO-08 unit group	M1 precision	M1 recall	M1 f1-score	M3 precision	M3 recall	M3 f1-score	support
2142 Civil Engineers	0.75	0.86	0.80	0.83	0.71	0.77	21
2143 Environmental Engineers	0.00	0.00	0.00	0.00	0.00	0.00	2
2144 Mechanical Engineers	0.67	0.67	0.67	0.67	0.76	0.71	21
2145 Chemical Engineers	0.67	0.50	0.57	0.44	1.00	0.62	4
2146 Mining Engineers, Metallurgists and Related Professionals	0.60	0.50	0.55	0.50	0.33	0.40	6
2149 Engineering Professionals Not Elsewhere Classified	0.44	0.71	0.55	0.25	0.41	0.31	17
2151 Electrical Engineers	0.60	0.82	0.69	0.50	0.73	0.59	11
2152 Electronics Engineers	0.53	0.73	0.62	0.75	0.55	0.63	11
2153 Telecommunications Engineers	0.00	0.00	0.00	0.33	0.33	0.33	3
2161 Building Architects	0.76	0.73	0.74	0.75	0.68	0.71	22
2162 Landscape Architects	1.00	0.60	0.75	0.80	0.80	0.80	5
2163 Product and Garment Designers	0.80	0.75	0.77	0.80	0.75	0.77	16
2164 Town and Traffic Planners	0.00	0.00	0.00	0.00	0.00	0.00	3
2165 Cartographers and Surveyors	0.80	1.00	0.89	0.80	1.00	0.89	4
2166 Graphic and Multimedia Designers	0.80	0.92	0.86	0.85	0.85	0.85	13
2211 Generalist Medical Practitioners	0.86	0.86	0.86	1.00	0.77	0.87	22
2212 Specialist Medical Practitioners	0.77	0.77	0.77	0.72	0.82	0.77	22
2221 Nursing Professionals	0.94	0.81	0.87	0.86	0.57	0.69	21
2222 Midwifery Professionals	0.71	0.83	0.77	0.83	0.83	0.83	6
2230 Traditional and Complementary Medicine Professionals	0.00	0.00	0.00	0.00	0.00	0.00	1
2240 Paramedical Practitioners	0.00	0.00	0.00	0.33	1.00	0.50	1
2261 Dentists	0.94	0.94	0.94	1.00	0.94	0.97	16
2262 Pharmacists	0.80	0.67	0.73	0.81	0.72	0.76	18
2263 Environmental and Occupational Health and Hygiene Professionals	0.00	0.00	0.00	0.50	0.33	0.40	3
2264 Physiotherapists	0.81	1.00	0.90	0.65	0.85	0.73	13
2265 Dieticians and Nutritionists	0.83	0.83	0.83	0.83	0.83	0.83	6
2266 Audiologists and Speech Therapists	1.00	1.00	1.00	0.80	1.00	0.89	4
2267 Optometrists and Ophthalmic Opticians	0.67	0.80	0.73	0.57	0.80	0.67	5
2269 Health Professionals Not Elsewhere Classified	0.40	0.25	0.31	0.25	0.13	0.17	8
2310 University and Higher Education Teachers	0.80	0.80	0.80	0.50	0.40	0.44	5
2320 Vocational Education Teachers	0.00	0.00	0.00	0.50	1.00	0.67	1
2330 Secondary Education Teachers	0.92	0.86	0.89	0.92	0.79	0.85	14
2341 Primary School Teachers	0.84	0.95	0.89	0.81	0.95	0.88	22
2342 Early Childhood Educators	0.70	0.76	0.73	0.71	0.71	0.71	21
2351 Education Methods Specialists	0.50	0.20	0.29	1.00	0.20	0.33	5
2352 Special Needs Teachers	0.86	0.90	0.88	0.81	0.81	0.81	21
2353 Other Language Teachers	1.00	0.78	0.88	1.00	0.61	0.76	18
2354 Other Music Teachers	0.89	1.00	0.94	0.89	1.00	0.94	8
2355 Other Arts Teachers	0.83	0.83	0.83	0.83	0.83	0.83	6
2356 Information Technology Trainers	1.00	0.33	0.50	0.50	0.33	0.40	3
2359 Teaching Professionals Not Elsewhere Classified	0.62	0.72	0.67	0.53	0.50	0.51	18
2411 Accountants	0.80	0.76	0.78	0.65	0.62	0.63	21
2412 Financial and Investment Advisers	0.71	0.63	0.67	0.76	0.68	0.72	19
2413 Financial Analysts	0.00	0.00	0.00	1.00	0.50	0.67	4
2421 Management and Organization Analysts	0.38	0.83	0.53	0.36	0.67	0.47	6

ISCO-08 unit group	M1 precision	M1 recall	M1 f1-score	M3 precision	M3 recall	M3 f1-score	support
2422 Policy Administration Professionals	0.00	0.00	0.00	0.00	0.00	0.00	4
2423 Personnel and Careers Professionals	0.60	0.33	0.43	0.67	0.44	0.53	9
2424 Training and Staff Development Professionals	0.67	0.33	0.44	0.57	0.67	0.62	6
2431 Advertising and Marketing Professionals	0.53	0.71	0.61	0.47	0.64	0.55	14
2432 Public Relations Professionals	0.33	0.50	0.40	0.00	0.00	0.00	2
2433 Technical and Medical Sales Professionals (excluding ICT)	0.63	0.50	0.56	0.60	0.60	0.60	10
2434 Information and Communications Technology Sales Professionals	0.00	0.00	0.00	0.00	0.00	0.00	3
2511 Systems Analysts	0.28	0.45	0.34	0.38	0.55	0.44	11
2512 Software Developers	0.65	0.71	0.68	0.50	0.67	0.57	21
2513 Web and Multimedia Developers	1.00	0.50	0.67	0.00	0.00	0.00	2
2514 Applications Programmers	0.58	0.64	0.61	0.75	0.55	0.63	11
2519 Software and Applications Developers and Analysts Not Elsewhere Classified	0.00	0.00	0.00	0.00	0.00	0.00	2
2521 Database Designers and Administrators	0.00	0.00	0.00	0.33	0.50	0.40	2
2522 Systems Aministrators	0.00	0.00	0.00	1.00	0.50	0.67	2
2523 Computer Network Professionals	0.00	0.00	0.00	0.00	0.00	0.00	1
2529 Database and Network Professionals Not Elsewhere Classified	0.50	0.33	0.40	0.50	0.33	0.40	3
2611 Lawyers	1.00	0.81	0.89	1.00	0.67	0.80	21
2612 Judges	0.60	0.75	0.67	0.60	0.75	0.67	4
2619 Legal Professionals Not Elsewhere Classified	0.33	0.50	0.40	0.44	0.50	0.47	8
2621 Archivists and Curators	0.00	0.00	0.00	1.00	1.00	1.00	1
2622 Librarians and Related Information Professionals	0.80	0.89	0.84	0.67	0.89	0.76	9
2631 Economists	0.82	0.75	0.78	0.83	0.83	0.83	12
2632 Sociologists, Anthropologists and Related Professionals	0.67	1.00	0.80	0.50	1.00	0.67	2
2633 Philosophers, Historians and Political Scientists	0.00	0.00	0.00	0.00	0.00	0.00	1
2634 Psychologists	0.86	0.86	0.86	1.00	0.86	0.93	22
2635 Social Work and Counselling Professionals	0.52	0.76	0.62	0.70	0.67	0.68	21
2636 Religious Professionals	0.70	1.00	0.82	0.50	1.00	0.67	7
2641 Authors and Related Writers	0.67	1.00	0.80	0.60	0.75	0.67	4
2642 Journalists	0.71	0.83	0.77	0.83	0.83	0.83	12
2643 Translators, Interpreters and Other Linguists	1.00	1.00	1.00	1.00	1.00	1.00	7
2651 Visual Artists	0.58	1.00	0.74	0.58	1.00	0.74	7
2652 Musicians, Singers and Composers	0.82	1.00	0.90	0.64	0.78	0.70	9
2653 Dancers and Choreographers	0.00	0.00	0.00	1.00	1.00	1.00	1
2654 Film, Stage and Related Directors and Producers	0.50	0.80	0.62	0.63	1.00	0.77	5
2655 Actors	1.00	0.50	0.67	1.00	0.50	0.67	2
2656 Announcers on Radio, Television and Other Media	0.00	0.00	0.00	0.00	0.00	0.00	1
2659 Creative and Performing Artists Not Elsewhere Classified	0.00	0.00	0.00	0.33	0.50	0.40	2
3111 Chemical and Physical Science Technicians	0.40	0.50	0.44	0.33	0.25	0.29	4
3112 Civil Engineering Technicians	0.31	0.31	0.31	0.22	0.15	0.18	13
3113 Electrical Engineering Technicians	0.60	0.25	0.35	0.47	0.58	0.52	12
3114 Electronics Engineering Technicians	0.00	0.00	0.00	1.00	0.33	0.50	3
3115 Mechanical Engineering Technicians	0.00	0.00	0.00	0.38	0.33	0.35	9
3116 Chemical Engineering Technicians	0.00	0.00	0.00	0.00	0.00	0.00	1
3117 Mining and metallurgical technicians	0.00	0.00	0.00	0.00	0.00	0.00	1

ISCO-08 unit group	M1 precision	M1 recall	M1 f1-score	M3 precision	M3 recall	M3 f1-score	support
3118 Draughtspersons	0.57	0.57	0.57	0.60	0.43	0.50	7
3119 Physical and Engineering Science Technicians Not Elsewhere Classified	0.00	0.00	0.00	0.00	0.00	0.00	3
3121 Mining Supervisors	0.50	0.50	0.50	0.40	0.50	0.44	4
3122 Manufacturing Supervisors	0.25	0.14	0.18	0.20	0.14	0.17	7
3123 Construction Supervisors	0.60	0.53	0.56	0.56	0.59	0.57	17
3131 Power Production Plant Operators	0.00	0.00	0.00	0.00	0.00	0.00	3
3132 Incinerator and Water Treatment Plant Operators	0.80	1.00	0.89	0.67	1.00	0.80	4
3134 Petroleum and Natural Gas Refining Plant Operators	0.00	0.00	0.00	0.00	0.00	0.00	1
3135 Metal Production Process Controllers	0.00	0.00	0.00	0.00	0.00	0.00	1
3139 Process Control Technicians Not Elsewhere Classified	0.00	0.00	0.00	0.00	0.00	0.00	3
3141 Life Science Technicians (excluding Medical)	0.00	0.00	0.00	0.00	0.00	0.00	1
3142 Agricultural Technicians	0.00	0.00	0.00	0.00	0.00	0.00	4
3143 Forestry Technicians	0.00	0.00	0.00	0.00	0.00	0.00	1
3151 Ships' Engineers	0.00	0.00	0.00	0.50	0.50	0.50	2
3152 Ships' Deck Officers and Pilots	0.50	1.00	0.67	0.50	1.00	0.67	2
3153 Aircraft Pilots and Related Associate Professionals	1.00	0.83	0.91	1.00	1.00	1.00	6
3154 Air Traffic Controllers	1.00	1.00	1.00	1.00	1.00	1.00	1
3211 Medical Imaging and Therapeutic Equipment Technicians	0.50	0.67	0.57	0.50	0.50	0.50	6
3212 Medical and Pathology Laboratory Technicians	0.70	0.88	0.78	0.86	0.75	0.80	8
3213 Pharmaceutical Technicians and Assistants	0.43	1.00	0.60	0.38	0.83	0.53	6
3214 Medical and Dental Prosthetic Technicians	1.00	0.80	0.89	0.83	1.00	0.91	5
3221 Nursing Associate Professionals	0.79	0.90	0.84	0.69	0.86	0.77	21
3222 Midwifery Associate Professionals	0.00	0.00	0.00	0.50	1.00	0.67	1
3240 Veterinary Technicians and Assistants	1.00	0.50	0.67	1.00	0.50	0.67	2
3251 Dental Assistants and Therapists	0.85	1.00	0.92	0.83	0.91	0.87	11
3252 Medical Records and Health Information Technicians	0.00	0.00	0.00	0.33	0.50	0.40	2
3253 Community Health Workers	0.00	0.00	0.00	0.00	0.00	0.00	1
3254 Dispensing Opticians	0.67	0.67	0.67	0.50	0.33	0.40	3
3255 Physiotherapy Technicians and Assistants	0.46	0.86	0.60	0.50	0.29	0.36	7
3256 Medical Assistants	0.80	0.80	0.80	0.75	0.60	0.67	10
3257 Environmental and Occupational Health Inspectors and Associates	0.48	0.83	0.61	0.55	0.50	0.52	12
3258 Ambulance Workers	0.67	1.00	0.80	0.75	0.75	0.75	4
3259 Health Associate Professionals Not Elsewhere Classified	0.00	0.00	0.00	0.00	0.00	0.00	3
3311 Securities and Finance Dealers and Brokers	0.50	0.83	0.63	0.50	0.50	0.50	6
3312 Credit and Loans Officers	0.45	0.50	0.48	0.60	0.60	0.60	10
3313 Accounting Associate Professionals	0.61	0.67	0.64	0.58	0.52	0.55	21
3314 Statistical, Mathematical and Related Associate Professionals	0.00	0.00	0.00	0.00	0.00	0.00	1
3315 Valuers and Loss Assessors	1.00	0.33	0.50	0.80	0.67	0.73	6
3321 Insurance Representatives	0.58	0.90	0.70	0.63	0.90	0.75	21
3322 Commercial Sales Representatives	0.43	0.62	0.51	0.57	0.38	0.46	21
3323 Buyers	0.67	0.67	0.67	0.88	0.58	0.70	12
3324 Trade Brokers	0.00	0.00	0.00	0.00	0.00	0.00	3
3331 Clearing and Forwarding Aents	1.00	0.25	0.40	1.00	0.50	0.67	4
3332 Conference and Event Planners	0.63	0.50	0.56	0.67	0.40	0.50	10

ISCO-08 unit group	M1 precision	M1 recall	M1 f1-score	M3 precision	M3 recall	M3 f1-score	support
3333 Employment agents and contractors	0.67	0.67	0.67	0.80	0.67	0.73	6
3334 Real Estate Agents and Property Managers	0.73	0.90	0.81	0.69	0.86	0.77	21
3339 Business Services Agents Not Elsewhere Classified	0.00	0.00	0.00	0.00	0.00	0.00	2
3341 Office Supervisors	0.47	0.43	0.45	0.47	0.43	0.45	21
3342 Legal Secretaries	0.67	0.50	0.57	0.67	0.50	0.57	4
3343 Administrative and Executive Secretaries	0.41	0.52	0.46	0.61	0.81	0.69	21
3344 Medical Secretaries	0.65	0.73	0.69	0.73	0.73	0.73	15
3351 Customs and Border Inspectors	0.71	1.00	0.83	0.80	0.80	0.80	5
3352 Government Tax and Excise Officials	0.67	0.89	0.76	0.80	0.89	0.84	9
3353 Government Social Benefits Officials	0.33	0.17	0.22	0.33	0.33	0.33	6
3354 Government Licensing Officials	0.00	0.00	0.00	0.00	0.00	0.00	4
3355 Police Inspectors and Detectives	0.63	0.71	0.67	0.83	0.71	0.77	7
3359 Government Regulatory Associate Professionals Not Elsewhere Classified	0.63	0.79	0.70	0.65	0.79	0.71	19
3411 Legal and Related Associate Professionals	0.50	0.57	0.53	0.50	0.71	0.59	7
3412 Social Work Associate Professionals	0.36	0.36	0.36	0.32	0.50	0.39	14
3413 Religious Associate Professionals	1.00	0.33	0.50	1.00	0.33	0.50	3
3421 Athletes and Sports Players	1.00	0.67	0.80	0.50	0.67	0.57	3
3422 Sports Coaches, Instructors and Officials	0.71	0.56	0.63	1.00	0.33	0.50	9
3423 Fitness and Recreation Instructors and Programme Leaders	0.91	0.71	0.80	0.68	0.93	0.79	14
3431 Photographers	0.80	0.89	0.84	0.89	0.89	0.89	9
3432 Interior Designers and Decorators	0.73	0.80	0.76	0.73	0.80	0.76	10
3433 Gallery, Museum and Library Technicians	0.00	0.00	0.00	0.00	0.00	0.00	1
3434 Chefs	0.90	0.90	0.90	0.95	0.90	0.93	21
3435 Other Artistic and Cultural Associate Professionals	0.50	0.25	0.33	0.67	0.50	0.57	4
3511 Information and Communications Technology Operations Technicians	0.00	0.00	0.00	0.33	0.20	0.25	5
3512 Information and Communications Technology User Support Technicians	0.31	0.56	0.40	0.33	0.33	0.33	9
3513 Computer Network and Systems Technicians	0.33	0.33	0.33	0.29	0.22	0.25	9
3514 Web Technicians	0.00	0.00	0.00	0.00	0.00	0.00	2
3521 Broadcasting and Audio-visual Technicians	0.00	0.00	0.00	0.50	0.33	0.40	3
3522 Telecommunications Engineering Technicians	0.00	0.00	0.00	0.00	0.00	0.00	1
4110 General Office Clerks	0.25	0.40	0.31	0.50	0.40	0.44	5
4120 Secretaries (general)	0.50	0.80	0.62	0.80	0.80	0.80	5
4131 Typists and Word Processing Operators	0.00	0.00	0.00	0.00	0.00	0.00	2
4132 Data Entry Clerks	0.00	0.00	0.00	0.50	0.50	0.50	2
4211 Bank Tellers and Related Clerks	0.80	0.76	0.78	0.76	0.76	0.76	21
4212 Bookmakers, Croupiers and Related Gaming Workers	0.75	1.00	0.86	0.67	0.67	0.67	3
4213 Pawnbrokers and Money-lenders	0.50	0.50	0.50	0.67	1.00	0.80	2
4214 Debt Collectors and Related Workers	1.00	0.67	0.80	1.00	0.33	0.50	3
4221 Travel Consultants and Clerks	0.78	0.78	0.78	0.80	0.89	0.84	9
4222 Contact Centre Information Clerks	0.25	0.25	0.25	0.20	0.25	0.22	8
4223 Telephone Switchboard Operators	0.00	0.00	0.00	0.00	0.00	0.00	3
4224 Hotel Receptionists	0.75	0.75	0.75	0.67	0.50	0.57	4
4225 Inquiry Clerks	0.44	0.50	0.47	0.50	0.50	0.50	8
4226 Receptionists (general)	0.65	0.50	0.56	0.86	0.55	0.67	22
4227 Survey and Market Research Interviewers	0.00	0.00	0.00	0.00	0.00	0.00	1

ISCO-08 unit group	M1 precision	M1 recall	M1 f1-score	M3 precision	M3 recall	M3 f1-score	support
4229 Client Information Workers Not Elsewhere Classified	0.00	0.00	0.00	0.00	0.00	0.00	3
4311 Accounting and Bookkeeping Clerks	0.54	0.62	0.58	0.64	0.43	0.51	21
4312 Statistical, Finance and Insurance Clerks	0.20	0.15	0.17	0.20	0.15	0.17	13
4313 Payroll Clerks	0.57	0.44	0.50	0.50	0.44	0.47	9
4321 Stock Clerks	0.53	0.44	0.48	0.67	0.56	0.61	18
4322 Production Clerks	0.00	0.00	0.00	0.00	0.00	0.00	3
4323 Transport Clerks	0.50	0.60	0.55	0.56	0.67	0.61	15
4411 Library Clerks	0.00	0.00	0.00	0.00	0.00	0.00	1
4412 Mail Carriers and Sorting Clerks	0.85	0.85	0.85	0.83	0.75	0.79	20
4415 Filing and Copying Clerks	0.00	0.00	0.00	0.00	0.00	0.00	1
4416 Personnel Clerks	0.40	0.29	0.33	0.50	0.43	0.46	7
4419 Clerical Support Workers Not Elsewhere Classified	0.00	0.00	0.00	0.00	0.00	0.00	5
5111 Travel Attendants and Travel Stewards	0.80	0.67	0.73	0.83	0.83	0.83	6
5112 Transport Conductors	0.67	0.50	0.57	1.00	0.75	0.86	4
5113 Travel Guides	0.00	0.00	0.00	0.00	0.00	0.00	2
5120 Cooks	0.63	0.83	0.71	0.67	0.67	0.67	6
5131 Waiters	0.73	0.76	0.74	0.73	0.76	0.74	21
5132 Bartenders	0.89	0.89	0.89	0.80	0.89	0.84	9
5141 Hairdressers	1.00	0.90	0.95	1.00	0.90	0.95	21
5142 Beauticians and Related Workers	0.90	0.86	0.88	0.76	0.90	0.83	21
5151 Cleaning and Housekeeping Supervisors in Offices, Hotels and Other Establishments	0.60	0.33	0.43	0.67	0.44	0.53	9
5152 Domestic Housekeepers	0.00	0.00	0.00	0.00	0.00	0.00	5
5153 Building Caretakers	0.57	0.76	0.65	0.52	0.71	0.60	21
5162 Companions and Valets	0.00	0.00	0.00	0.00	0.00	0.00	1
5163 Undertakers and Embalmers	1.00	0.50	0.67	0.50	1.00	0.67	2
5164 Pet Groomers and Animal Care Workers	0.67	0.86	0.75	0.75	0.86	0.80	7
5165 Driving Instructors	1.00	1.00	1.00	1.00	1.00	1.00	3
5169 Personal Services Workers Not Elsewhere Classified	1.00	0.67	0.80	1.00	0.67	0.80	3
5211 Stall and Market Salespersons	0.64	0.73	0.68	0.61	0.64	0.62	22
5212 Street Food Salespersons	1.00	0.33	0.50	1.00	0.67	0.80	3
5221 Shopkeepers	0.65	0.62	0.63	0.67	0.57	0.62	21
5222 Shop Supervisors	0.57	0.44	0.50	0.41	0.39	0.40	18
5223 Shop Sales Assistants	0.63	0.55	0.59	0.50	0.45	0.48	22
5230 Cashiers and Ticket Clerks	0.33	0.50	0.40	0.50	0.75	0.60	4
5242 Sales Demonstrators	0.36	0.36	0.36	0.44	0.50	0.47	14
5243 Door-to-door Salespersons	0.00	0.00	0.00	0.50	0.20	0.29	5
5244 Contact Centre Salespersons	0.33	0.38	0.35	0.56	0.63	0.59	8
5245 Service Station Attendants	0.71	0.71	0.71	1.00	0.71	0.83	7
5246 Food Service Counter Attendants	0.38	0.40	0.39	0.45	0.33	0.38	15
5249 Sales Workers Not Elsewhere Classified	0.29	0.36	0.32	0.67	0.18	0.29	11
5311 Child Care Workers	0.83	0.71	0.77	0.80	0.76	0.78	21
5312 Teachers' Aides	0.67	0.67	0.67	0.65	0.62	0.63	21
5321 Health Care Assistants	0.64	0.76	0.70	0.61	0.67	0.64	21
5322 Home-based Personal Care Workers	0.63	0.55	0.59	0.65	0.59	0.62	22
5329 Personal Care Workers in Health Services Not Elsewhere Classified	0.50	0.37	0.42	0.50	0.37	0.42	19

ISCO-08 unit group	M1 precision	M1 recall	M1 f1-score	M3 precision	M3 recall	M3 f1-score	support
5411 Fire Fighters	0.87	0.87	0.87	0.81	0.87	0.84	15
5412 Police Officers	0.82	0.86	0.84	0.77	0.95	0.85	21
5413 Prison Guards	0.75	1.00	0.86	0.75	1.00	0.86	9
5414 Security Guards	0.68	0.86	0.76	0.65	0.91	0.75	22
5419 Protective Services Workers Not Elsewhere Classified	0.33	0.13	0.18	0.33	0.25	0.29	8
6111 Field Crop and Vegetable Growers	0.62	0.67	0.64	0.75	0.50	0.60	12
6112 Tree and Shrub Crop Growers	1.00	0.25	0.40	1.00	0.25	0.40	4
6113 Gardeners, Horticultural and Nursery Growers	0.52	0.67	0.59	0.50	0.50	0.50	18
6114 Mixed Crop Growers	0.00	0.00	0.00	0.00	0.00	0.00	1
6121 Livestock and Dairy Producers	0.58	0.68	0.62	0.64	0.64	0.64	22
6122 Poultry Producers	0.00	0.00	0.00	0.00	0.00	0.00	2
6123 Apiarists and Sericulturists	0.00	0.00	0.00	0.50	1.00	0.67	1
6129 Animal Producers Not Elsewhere Classified	0.00	0.00	0.00	0.00	0.00	0.00	1
6130 Mixed Crop and Animal Producers	0.00	0.00	0.00	0.50	1.00	0.67	2
6221 Aquaculture Workers	0.00	0.00	0.00	0.00	0.00	0.00	1
6222 Inland and Coastal Waters Fishery Workers	0.67	1.00	0.80	0.75	0.75	0.75	4
6223 Deep-sea Fishery Workers	0.00	0.00	0.00	0.00	0.00	0.00	1
7111 House Builders	0.71	0.71	0.71	0.79	0.52	0.63	21
7112 Bricklayers and Related Workers	0.76	0.90	0.83	0.85	0.81	0.83	21
7113 Stonemasons, Stone Cutters, Splitters and Carvers	1.00	0.25	0.40	1.00	0.50	0.67	4
7114 Concrete Placers, Concrete Finishers and Related Workers	0.80	0.80	0.80	0.86	0.60	0.71	10
7115 Carpenters and Joiners	0.74	0.81	0.77	0.76	0.76	0.76	21
7119 Building Frame and Related Trades Workers Not Elsewhere Classified	0.29	0.25	0.27	0.00	0.00	0.00	8
7121 Roofers	1.00	0.43	0.60	0.80	0.57	0.67	7
7122 Floor Layers and Tile Setters	0.85	1.00	0.92	0.85	1.00	0.92	11
7123 Plasterers	0.40	0.50	0.44	0.50	0.50	0.50	4
7124 Insulation Workers	0.80	1.00	0.89	0.80	1.00	0.89	4
7125 Glaziers	0.40	0.40	0.40	1.00	0.60	0.75	5
7126 Plumbers and Pipe Fitters	0.80	0.95	0.87	0.73	0.90	0.81	21
7127 Air Conditioning and Refrigeration Mechanics	0.67	0.75	0.71	0.78	0.88	0.82	8
7131 Painters and Related Workers	0.90	0.90	0.90	0.79	0.90	0.84	21
7132 Spray Painters and Varnishers	0.50	0.80	0.62	0.75	0.60	0.67	5
7133 Building Structure Cleaners	0.00	0.00	0.00	0.00	0.00	0.00	1
7211 Metal Moulders and Coremakers	0.00	0.00	0.00	0.00	0.00	0.00	2
7212 Welders and Flame Cutters	0.76	0.90	0.83	0.79	0.90	0.84	21
7213 Sheet Metal Workers	0.29	0.25	0.27	0.22	0.25	0.24	8
7214 Structural Metal Preparers and Erectors	0.33	0.29	0.31	0.33	0.29	0.31	7
7215 Riggers and Cable Splicers	0.00	0.00	0.00	0.00	0.00	0.00	1
7221 Blacksmiths, Hammersmiths and Forging Press Workers	0.56	0.63	0.59	0.42	0.63	0.50	8
7222 Toolmakers and Related Workers	0.33	0.20	0.25	0.00	0.00	0.00	5
7223 Metal Working Machine Tool Setters and Operators	0.50	0.50	0.50	0.67	0.50	0.57	8
7224 Metal Polishers, Wheel Grinders and Tool Sharpeners	0.00	0.00	0.00	0.00	0.00	0.00	1
7231 Motor Vehicle Mechanics and Repairers	0.77	0.81	0.79	0.95	0.90	0.93	21
7232 Aircraft Engine Mechanics and Repairers	0.80	0.80	0.80	0.50	0.80	0.62	5

ISCO-08 unit group	M1 precision	M1 recall	M1 f1-score	M3 precision	M3 recall	M3 f1-score	support
7233 Agricultural and Industrial Machinery Mechanics and Repairers	0.39	0.65	0.49	0.43	0.76	0.55	17
7234 Bicycle and Related Repairers	0.00	0.00	0.00	1.00	1.00	1.00	1
7311 Precision-instrument Makers and Repairers	0.00	0.00	0.00	1.00	0.50	0.67	2
7313 Jewellery and Precious Metal Workers	0.67	1.00	0.80	0.80	1.00	0.89	4
7314 Potters and Related Workers	0.00	0.00	0.00	1.00	1.00	1.00	1
7316 Sign Writers, Decorative Painters, Engravers and Etchers	0.00	0.00	0.00	1.00	1.00	1.00	1
7317 Handicraft Workers in Wood, Basketry and Related Materials	0.22	0.50	0.31	0.50	0.50	0.50	4
7318 Handicraft Workers in Textile, Leather and Related Materials	0.50	0.33	0.40	0.00	0.00	0.00	3
7319 Handicraft Workers Not Elsewhere Classified	0.00	0.00	0.00	0.00	0.00	0.00	1
7321 Pre-press Technicians	0.00	0.00	0.00	0.00	0.00	0.00	1
7322 Printers	0.42	0.71	0.53	0.56	0.71	0.63	7
7323 Print Finishing and Binding Workers	0.00	0.00	0.00	0.00	0.00	0.00	1
7411 Building and Related Electricians	0.68	0.90	0.78	0.67	0.86	0.75	21
7412 Electrical Mechanics and Fitters	0.13	0.13	0.13	0.25	0.13	0.17	15
7413 Electrical Line Installers and Repairers	0.67	0.29	0.40	0.75	0.43	0.55	7
7421 Electronics Mechanics and Servicers	0.33	0.13	0.18	0.50	0.50	0.50	8
7422 Information and Communications Technology Installers and Servicers	0.42	0.56	0.48	0.50	0.44	0.47	9
7511 Butchers, Fishmongers and Related Food Preparers	0.74	0.91	0.82	0.76	0.86	0.81	22
7512 Bakers, Pastry-cooks and Confectionery Makers	0.89	0.81	0.85	0.85	0.81	0.83	21
7513 Dairy Products Makers	1.00	0.25	0.40	1.00	0.25	0.40	4
7514 Fruit, Vegetable and Related Preservers	0.00	0.00	0.00	0.00	0.00	0.00	1
7515 Food and Beverage Tasters and Graders	0.00	0.00	0.00	0.00	0.00	0.00	2
7521 Wood Treaters	0.00	0.00	0.00	0.00	0.00	0.00	2
7522 Cabinet-makers and Related Workers	0.42	0.56	0.48	0.55	0.67	0.60	9
7523 Woodworking Machine Tool Setters and Operators	0.00	0.00	0.00	0.00	0.00	0.00	1
7531 Tailors, Dressmakers, Furriers and Hatters	0.79	0.88	0.83	0.71	0.88	0.79	17
7532 Garment and Related Patternmakers and Cutters	0.83	0.83	0.83	0.80	0.67	0.73	6
7533 Sewing, Embroidery and Related Workers	0.75	0.67	0.71	0.56	0.56	0.56	9
7534 Upholsterers and Related Workers	0.50	0.33	0.40	0.00	0.00	0.00	3
7535 Pelt Dressers, Tanners and Fellmongers	0.00	0.00	0.00	0.00	0.00	0.00	1
7536 Shoemakers and Related Workers	0.60	0.60	0.60	0.60	0.60	0.60	5
7541 Underwater Divers	1.00	0.50	0.67	1.00	1.00	1.00	2
7542 Shotfirers and Blasters	0.00	0.00	0.00	0.00	0.00	0.00	1
7543 Product Graders and Testers (except Foods and Beverages)	0.40	0.40	0.40	0.00	0.00	0.00	5
7544 Fumigators and Other Pest and Weed Controllers	1.00	0.67	0.80	0.75	1.00	0.86	3
7549 Craft and Related Workers not Elsewhere Classified	0.00	0.00	0.00	0.00	0.00	0.00	1
8111 Miners and Quarriers	0.33	0.20	0.25	0.50	0.30	0.37	10
8112 Mineral and Stone Processing Plant Operators	0.00	0.00	0.00	0.00	0.00	0.00	1
8113 Well Drillers and Borers and Related Workers	0.71	0.71	0.71	0.46	0.86	0.60	7
8114 Cement, Stone and Other Mineral Products Machine Operators	0.00	0.00	0.00	0.00	0.00	0.00	1
8121 Metal Processing Plant Operators	0.56	0.83	0.67	1.00	0.33	0.50	6
8122 Metal Finishing, Plating and Coating Machine Operators	0.00	0.00	0.00	0.00	0.00	0.00	2
8131 Chemical Products Plant and Machine Operators	0.14	0.25	0.18	0.20	0.25	0.22	4

ISCO-08 unit group	M1 precision	M1 recall	M1 f1-score	M3 precision	M3 recall	M3 f1-score	support
8132 Photographic Products Machine Operators	0.00	0.00	0.00	0.00	0.00	0.00	1
8141 Rubber Products Machine Operators	0.00	0.00	0.00	1.00	0.50	0.67	2
8142 Plastic Products Machine Operators	0.00	0.00	0.00	1.00	0.33	0.50	3
8143 Paper Products Machine Operators	0.00	0.00	0.00	0.00	0.00	0.00	2
8151 Fibre Preparing, Spinning and Winding Machine Operators	0.00	0.00	0.00	0.00	0.00	0.00	1
8152 Weaving and Knitting Machine Operators	0.00	0.00	0.00	0.00	0.00	0.00	1
8153 Sewing Machine Operators	0.00	0.00	0.00	0.00	0.00	0.00	3
8154 Bleaching, Dyeing and Fabric Cleaning Machine Operators	0.00	0.00	0.00	0.00	0.00	0.00	1
8156 Shoemaking and Related Machine Operators	0.00	0.00	0.00	0.00	0.00	0.00	2
8157 Laundry Machine Operators	0.38	0.75	0.50	0.00	0.00	0.00	4
8159 Textile, Fur and Leather Products Machine Operators Not Elsewhere Classified	0.00	0.00	0.00	1.00	0.50	0.67	2
8171 Pulp and Papermaking Plant Operators	1.00	0.67	0.80	0.50	0.33	0.40	3
8172 Wood Processing Plant Operators	0.40	0.50	0.44	0.50	0.50	0.50	4
8181 Glass and Ceramics Plant Operators	0.00	0.00	0.00	1.00	1.00	1.00	2
8182 Steam Engine and Boiler Operators	0.50	0.33	0.40	0.50	0.33	0.40	3
8183 Packing, Bottling and Labelling Machine Operators	0.00	0.00	0.00	0.25	0.14	0.18	7
8189 Stationary Plant and Machine Operators Not Elsewhere Classified	0.44	0.67	0.53	0.36	0.67	0.47	6
8211 Mechanical Machinery Assemblers	0.48	0.74	0.58	0.56	0.53	0.54	19
8212 Electrical and Electronic Equipment Assemblers	0.00	0.00	0.00	0.00	0.00	0.00	3
8219 Assemblers Not Elsewhere Classified	0.00	0.00	0.00	0.14	0.17	0.15	6
8311 Locomotive Engine Drivers	0.67	0.67	0.67	0.75	0.50	0.60	6
8312 Railway Brake, Signal and Switch Operators	0.00	0.00	0.00	0.50	0.50	0.50	2
8322 Car, Taxi and Van Drivers	0.86	0.90	0.88	0.74	0.67	0.70	21
8331 Bus and Tram Drivers	0.87	0.95	0.91	0.91	0.95	0.93	21
8332 Heavy Truck and Lorry Drivers	0.73	0.76	0.74	0.59	0.62	0.60	21
8341 Mobile Farm and Forestry Plant Operators	1.00	0.60	0.75	0.70	0.70	0.70	10
8342 Earthmoving and Related Plant Operators	0.56	0.69	0.62	0.53	0.62	0.57	13
8343 Crane, hoist and related plant operators	0.69	0.75	0.72	0.67	0.50	0.57	12
8344 Lifting Truck Operators	0.80	0.67	0.73	0.82	0.75	0.78	12
8350 Ships' Deck Crews and Related Workers	0.00	0.00	0.00	0.00	0.00	0.00	1
9111 Domestic Cleaners and Helpers	0.77	0.81	0.79	0.65	0.81	0.72	21
9112 Cleaners and Helpers in Offices, Hotels and Other Establishments	0.60	0.86	0.71	0.79	0.71	0.75	21
9121 Hand Launderers and Pressers	0.00	0.00	0.00	0.50	0.50	0.50	4
9122 Vehicle Cleaners	0.80	1.00	0.89	1.00	0.50	0.67	4
9123 Window Cleaners	1.00	1.00	1.00	1.00	1.00	1.00	1
9129 Other Cleaning Workers	0.50	0.25	0.33	0.60	0.75	0.67	4
9211 Crop Farm Labourers	0.60	0.71	0.65	0.65	0.62	0.63	21
9212 Livestock Farm Labourers	0.22	0.40	0.29	0.29	0.40	0.33	5
9213 Mixed Crop and Livestock Farm Labourers	0.00	0.00	0.00	0.17	0.33	0.22	3
9214 Garden and Horticultural Labourers	0.59	0.63	0.61	0.56	0.56	0.56	16
9215 Forestry Labourers	0.00	0.00	0.00	0.00	0.00	0.00	3
9216 Fishery and Aquaculture Labourers	0.00	0.00	0.00	0.00	0.00	0.00	2
9311 Mining and Quarrying Labourers	0.71	0.83	0.77	0.56	0.83	0.67	6
9312 Civil Engineering Labourers	0.33	0.43	0.38	0.19	0.43	0.26	7
9313 Building Construction Labourers	0.43	0.57	0.49	0.40	0.67	0.50	21

ISCO-08 unit group	M1 precision	M1 recall	M1 f1-score	M3 precision	M3 recall	M3 f1-score	support
9321 Hand Packers	0.50	0.71	0.59	0.50	0.71	0.59	14
9329 Manufacturing Labourers Not Elsewhere Classified	0.67	0.55	0.60	0.38	0.55	0.44	11
9333 Freight Handlers	0.38	0.43	0.40	0.45	0.43	0.44	21
9334 Shelf Fillers	0.43	0.63	0.51	0.50	0.63	0.56	16
9411 Fast Food Preparers	0.80	0.36	0.50	0.83	0.45	0.59	11
9412 Kitchen Helpers	0.71	0.77	0.74	0.69	0.91	0.78	22
9611 Garbage and Recycling Collectors	0.67	1.00	0.80	0.50	0.25	0.33	4
9612 Refuse Sorters	0.00	0.00	0.00	0.00	0.00	0.00	1
9613 Sweepers and Related Labourers	0.63	0.83	0.71	0.71	0.83	0.77	6
9621 Messengers, Package Deliverers and Luggage Porters	0.68	0.62	0.65	0.57	0.57	0.57	21
9622 Odd Job Persons	0.43	0.33	0.38	0.50	0.22	0.31	9
9623 Meter Readers and Vending-machine Collectors	0.00	0.00	0.00	0.67	1.00	0.80	2
9624 Water and Firewood Collectors	0.00	0.00	0.00	0.00	0.00	0.00	1
9629 Elementary Workers Not Elsewhere Classified	1.00	0.20	0.33	0.33	0.40	0.36	5
9701 Home duties	0.78	0.64	0.70	0.74	0.64	0.68	22
9702 Student	0.78	0.58	0.67	0.70	0.58	0.64	12
9703 Social beneficiary	0.81	0.81	0.81	0.74	0.81	0.77	21
9704 'I don't know' and similar responses	0.89	0.81	0.85	0.89	0.81	0.85	21
9705 Vague responses	0.44	0.19	0.27	0.19	0.14	0.16	21

7.4 Appendix D: Supplementary digital resources

ICILS Multilingual ISCO-08 Parental Occupation Corpus Dataset is a private[14] git repository with the dataset configurations and splits used in the research and can be explored online and analyzed using the Hugging Face Datasets Python package.

The ISCO-08 Hierarchical Accuracy Measure metric card is a publicly accessible implementation of the evaluation measure with the Hugging Face Evaluate Python package and demo space for testing the implementation.

ICILS XLM-R-ICILS-ILO model card and Hosted Inference API is a gated access[15] git repository with the fine-tuned model weights, configuration files, and parameters used for training.

ICILS ISCO R&D Model Trainings Reports is an automatically generated report of the model's accuracy over time captured during the training procedure with machine learning experiment monitoring software.

ICILS ISCO AI Git Repository is a publicly accessible code repository with the Python code for reproducing the model training and evaluations.

[14] The dataset is securely hosted and is accessible only to IEA organization members.
[15] Access to the model repository requires a user to share their contact information (email and username), to agree to IEA's terms of use, and for the request to be manually approved by the repository owners.

8 REFERENCES

Bethmann, Arne & Schierholz, Malte & Wenzig, Knut & Nester, & Markus. (2014). *Automatic Coding of Occupations. Using Machine Learning Algorithms for Occupation Coding in Several German Panel Surveys.*

Conneau, A., Lample, G., Rinott, R., Williams, A., S. R., B., Schwenk, H., & Stoyanov, V. (2018). *XNLI: Evaluating cross-lingual sentence representations.* arXiv preprint.

Conneau, Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., . . . Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, *. Retrieved from https://arxiv.org/pdf/1810.04805.pdf

Elias, P. (1997). Occupational Classification (ISCO-88): Concepts, Methods, Reliability, Validity and Cross-National Comparability. *OECD Labour Market and Social Policy Occasional Papers*(20).

ESCO Publications. (2022). *Machine Learning Assisted Mapping of Multilingual Occupational Data to ESCO.* Eurpean Commission. Retrieved from https://esco.ec.europa.eu/en/about-esco/publications/publication/machine-learning-assisted-mapping-multilingual-occupational

European Commission. (2022). The crosswalk between ESCO and O*NET: Technical Report. 5.

European Commission and Directorate-General for Employment, Social Affairs and Inclusion. (2022). *European skills, competences, qualifications and occupations – Annual report 2021.* Publications Office of the European Union. doi:doi/10.2767/575560

Feng, Y., Qiang, J., Li, Y., Yuan, Y., & Zhu, Y. (2023). Sentence Simplification via Large Language Models. *arXiv*. doi:10.48550/arXiv.2302.11957

Fraillon, Ainley, Schulz, Friedman, & Duckworth. (2018). *International Computer and Information Literacy Study 2018 Technical Report.* Amsterdam: IEA.

Ganzeboom, H. (2010). A new International Socio-Economic Index (ISEI) of occupational status for the International Standard Classification of Occupation 2008 (ISCO-08) constructed with data from the ISSP 2002–2007.

Ganzeboom, H. B., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, 1-56.

Gemini Team, G. T., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., . . . Hauth, A. (2023). Gemini: A Family of Highly Capable Multimodal Models. *arXiv*.

Holzberger et al. (2020). A meta-analysis on the relationship between school characteristics and student outcomes in science and maths – evidence from large-scale studies. *Studies in Science Education, 56*, 1-34.

International Labour Office. (2007). International Standard Classification of Occupations: ISCO-08. Geneva.

International Labour Organization. (2012). *International Standard Classification of Occupations 2008 (ISCO-08) Volume 1: Structure, Group Definitions and Correspondence Tables.* Geneva: International Labour Office.

International Labour Organization. (2024, March 31). *International Standard Classification of Occupations (ISCO)*. Retrieved from ILOSTAT: https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/

Kiritchenko, S., & Famili, F. (2005). Functional Annotation of Genes Using Hierarchical Text Categorization. *Proceedings of BioLink SIG, ISMB*.

Lewis, P., Oğuz, B., Rinott, R., Riedel, S., & Schwenk, H. (2019). *MLQA: Evaluating cross-lingual extractive question answering.* arXiv preprint.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv*, *.

Liang, D., Gonen , H., Mao , Y., Hou, R., Goyal, N., Ghazvininejad , M., . . . Khabsa, M. (2023). XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models. *arXiv*, *.

Martin, M., Rust, K., & Adams, R. (1999). *Technical standards for IEA studies.* Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Molino, P., Dudin, Y., & Miryala, S. S. (2019). Ludwig: a type-based declarative deep learning toolbox. arXiv. doi:arXiv:1909.07930

Okkalioglu, M. (2023). TF-IGM revisited: Imbalance text classification with relative imbalance ratio. *Expert Systems with Applications, 217*, 119578. doi:10.1016/j.eswa.2023.119578

OpenAI. (2023). GPT-4 Technical Report. *arXiv*, *.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2023). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv*, *.

Safikhani, P., Avetisyan, H., Föste-Eggers, D., & Broneske , D. (2023). Automated occupation coding with hierarchical features: a data-centric approach to classification with pre-trained language models. *Discover Artificial Intelligence 3, 6*. doi:https://doi.org/10.1007/s44163-023-00050-y

Safikhani, P., Avetisyan, H., Föste-Eggers, D., & Broneske, D. (2023, February 13). Automated occupation coding with hierarchical features: a data-centric approach to classification with pre-trained language models. *Discover Artificial Intelligence*, *.

Schierholz, & Schonlau. (2021, November). Machine Learning for Occupation Coding—A Comparison Study. *Journal of Survey Statistics and Methodology, 9*(5), 1013-1034.

Schierholz, M., & Matthias, S. (2021). Machine Learning for Occupation Coding—A Comparison Study. *Journal of Survey Statistics and Methodology, Volume 9, Issue 5*, 1013–1034. doi:https://doi.org/10.1093/jssam/smaa023

Tarekegn, A. N., Giacobini, M., & Michalak, K. (2021). A review of methods for imbalanced multi-label classification. *Pattern Recognition, 118*, 107965. doi:10.1016/j.patcog.2021.107965

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., . . . G, A. (2023). Gemini: A Family of Highly Capable Multimodal Models. *arXiv*.

Tijdens, K. (2015). Self-identification of occupation in web surveys: requirements. *Survey Insights: Methods from the Field*. doi:10.13094/SMIF-2015-00008

Treiman. (2013). *Occupational Prestige in Comparative Perspective.* United States: Elsevier Science.

Wolf, C. (1997). The ISCO-88 International Standard Classification of Occupations in Cross-National Survey Research. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 54*(1), 23–40.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38--45). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-demos.6

Yadav, D., Jain, R., Agrawal, H., Chattopadhyay, P., Singh, T., Jain, A., . . . Batra, D. (2019). EvalAI: Towards Better Evaluation Systems for AI Agents. *arXiv*, *.

Yuan, Y., Wei, J., Huang, H., Jiao, W., Wang, J., & Chen, H. (2023). Review of resampling techniques for the treatment of imbalanced industrial data classification in equipment condition monitoring. *Engineering Applications of Artificial Intelligence, 126*, 106911. doi:doi.org/10.1016/j.engappai.2023.106911

Zhang, M., Goot, R. v., & Plank, B. (2023). ESCOXLM-R: Multilingual Taxonomy-driven Pre-training for the Job Market Domain. *arXiv*, 2305.12092.