Can Large Language Models Advance Crosswalks? The Case of Danish Occupation Codes

Bolei Ma[♣,♡] **Cynthia A. Huang**[♠] **Anna-Carolina Haensch**[♣,◇]

[♣]LMU Munich [♡]Munich Center for Machine Learning
[♠]Monash University [◇]University of Maryland, College Park

{bolei.ma, c.haensch}@lmu.de, cynthia.huang@monash.edu

Abstract

Crosswalks, which map one classification system to another, are critical tools for harmonizing data across time, countries, or frameworks. However, constructing crosswalks is labor-intensive and often requires domain expertise. This paper investigates the potential of Large Language Models (LLMs) to assist in creating crosswalks, focusing on two Danish occupational classification systems from different time periods as a case study. We propose a two-stage, prompt-based framework for this task, where LLMs perform similarity assessments between classification codes and identify final mappings through a guided decision process. Using four instruction-tuned LLMs and comparing them against an embedding-based baseline, we evaluate the performance of different models in crosswalks. Our results highlight the strengths of LLMs in crosswalk creation compared to the embedding-based baseline, showing the effectiveness of the interactive prompt-based framework for conducting crosswalks by LLMs. Furthermore, we analyze the impact of model combinations across two interactive rounds, highlighting the importance of model selection and consistency. This work contributes to the growing field of NLP applications for domain-specific knowledge mapping and demonstrates the potential of LLMs in advancing crosswalk methodologies.

1 Introduction

Crosswalks are structured mappings that connect one classification system to another, enabling data to be compared or integrated across different contexts. These mappings are essential in numerous domains, from harmonizing occupational codes across time or countries (Rémen et al., 2018) to aligning taxonomies in biology (Cheng et al., 2017) or mapping educational milestones between frameworks (Subramaniam et al., 2013). While the contexts vary, the underlying challenge remains the same: translating between systems that often reflect different conceptual frameworks, levels of granularity, or terminologies.

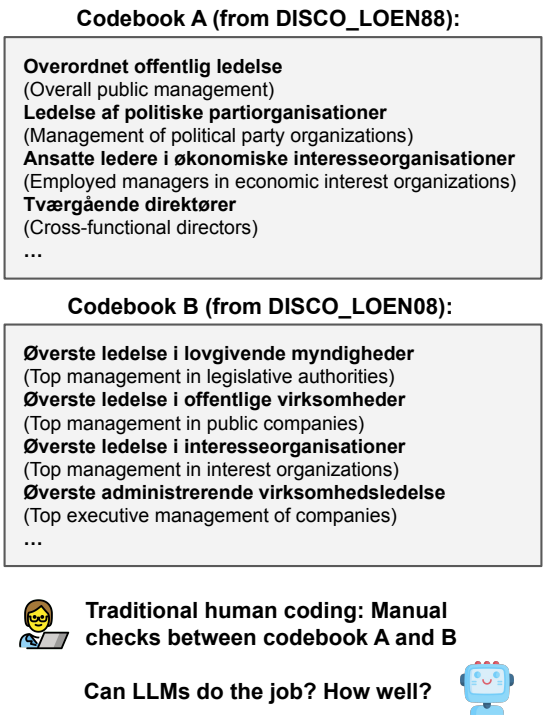

Figure 1: An example of crosswalks between two codebooks from the Danish occupation data. Translations are in commas. Traditionally, crosswalks are created manually by humans. Can LLMs assist in this process?

In the context of occupational classifications, for instance, crosswalks allow researchers to analyze labor market trends across time or national boundaries despite differences in coding systems. Figure 1 gives an example of crosswalks based on Danish occupation data. However, creating these mappings is a complex and labor-intensive process (Rémen et al., 2018). Large Language Models (LLMs) offer a promising avenue for addressing this challenge. Yet, their use in creating crosswalks raises essential questions: How can LLMs reliably infer mappings

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 392–399
April 30 - May 1, 2025

between systems with limited contextual overlap? What are the best strategies for prompting LLMs to elicit meaningful, interpretable outputs? And how do we ensure that the outputs of LLMs align with domain-specific requirements while remaining accessible to human users?

This paper explores the potential of LLMs to assist in creating crosswalks, using Danish occupational classifications from two different time points as a case study (Statistics Denmark, 2025b,a). Our aim is not to fully automate crosswalk creation but to develop an assisted workflow that combines the efficiency of LLMs with the judgment of human experts. Using a curated two-round judgment framework, we compare the performance of different LLMs to evaluate their strengths and limitations in supporting this task. Our empirical findings indicate that, despite certain limitations, the interactive LLM-based crosswalking process outperforms an embedding-based baseline. Through this work, we contribute to the growing field of NLP applications in social science research, showing how LLMs can be effectively integrated into complex domain-specific knowledge-mapping tasks.

2 Background

Much of the work at the intersection of NLP and Computational Social Science (CSS) focuses on labeling texts from social science domains to systematically analyze patterns, opinions, or topics (Chae and Davidson, 2023; Ziems et al., 2024). Occupational coding, a critical task in labor market research and social science, is an excellent use case to explore if and how large language models can enhance methodological approaches in these fields (Liu et al., 2022; Safikhani et al., 2023; Laughlin et al., 2024; Kononykhina et al., 2025).

Occupational codes are standardized labels assigned to jobs based on their duties, responsibilities, and required skills. However, occupational coding is a particularly complicated task because job descriptions can be context-dependent, and often ambiguous (Schierholz and Schonlau, 2020). Adding to this complexity, different countries and time periods often use distinct occupational classification schemes, each tailored to specific economic, social, or policy contexts. For instance, the International Standard Classification of Occupations (ILO, 2025) may differ significantly from national systems like the U.S. Standard Occupational Classification (BLS, 2025), necessitating the development of crosswalks to translate codes from one system to another. Crosswalks like the one by (Rémen et al., 2018) establish equivalencies between two occupational classification schemes allowing data coded in one system or country (US vs. Canada) to be translated into another. This process is essential for enabling international comparisons, historical analyses, and the integration of datasets that rely on different coding standards.

These crosswalks are typically created manually by domain experts who possess deep knowledge of the classification schemes in question. For example, Humlum (2021) developed a detailed crosswalk for Denmark's DISCO classifications. While such manually created crosswalks are highly accurate and tailored to specific needs, they are also exceptionally time-intensive and resource-intensive to produce, as they require establishing mappings between several hundred codes in each classification scheme. Therefore, there is growing interest in exploring whether LLMs can assist in the creation of crosswalks. Similar efforts have been made in other domains, such as healthcare and biomedical research, where tools like MapperGPT use large language models to refine and align entity mappings (Matentzoglu et al., 2023).

3 Method

We propose a two-round prompt-based framework to conduct the crosswalks for the occupation codes. The basic idea of the crosswalk is to find the possible matching code from codebook B for every code in codebook A. Figure 2 illustrates the basic workflow of our framework. The first round is about prompting the models to do similarity checks with certain degrees across all codes in both codebooks. Based on the results from the first round, the second round is about selecting the final candidate matching code from another codebook. This search is done for every code in one of the codebooks. The workflow is detailed as follows:

Round 1: Similarity Check across Codes. We begin with two codebooks (A and B) to work on, where codebook A contains a codeset of unique code names (Code A 1, Code A 2, ...), and codebook B contains a codeset of unique code names (Code B 1, Code B 2, ...). The task of the crosswalk is to map the codes from A to the codes from B. Therefore, in the initial step, we construct code pairs for each code from codebook A to every code from codebook B.

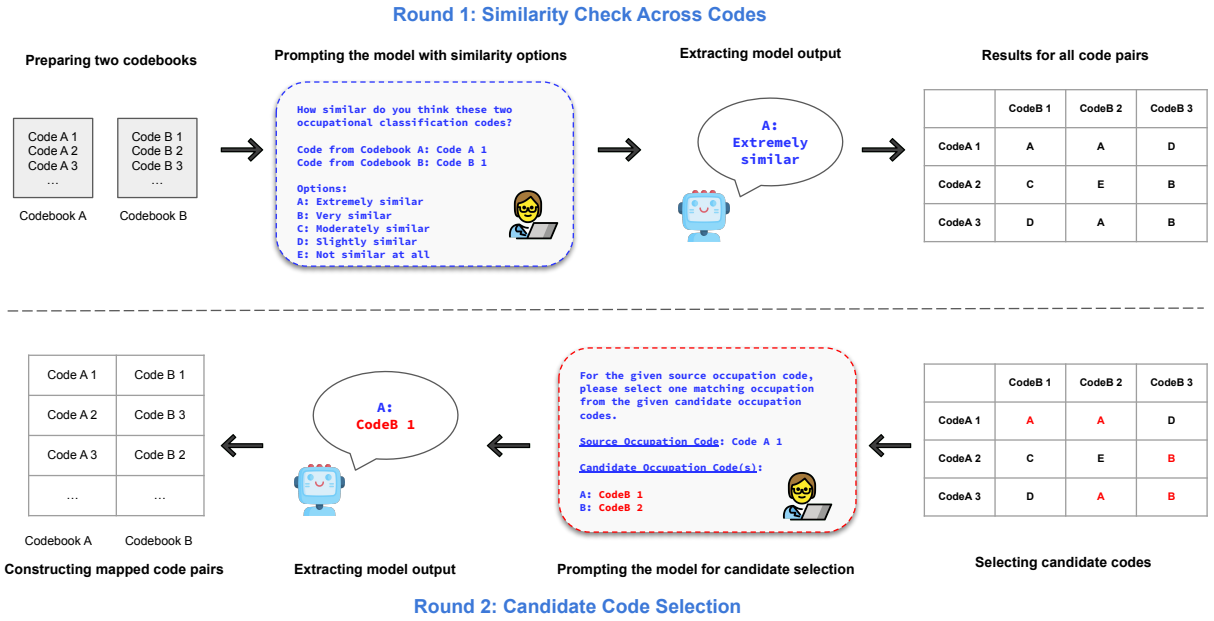

Figure 2: Our two-round prompt-based framework to conduct the crosswalks for the occupation codes using zero-shot LLM prompting.

For each code pair, we prompt the LLM with a question asking for the similarity and options indicating different similarity polarities at 5 scales (from A to E indicating extremely similar towards not similar at all). This scale is commonly used for survey questionnaires, due to its structured design, which presents respondents with predefined answer options, reducing ambiguity and ensuring consistency in responses; as well as its format, which facilitates faster decision-making by guiding participants through a clear set of choices, minimizing cognitive load and improving response accuracy (Likert, 1932; Groves, 2011). This setup has also been recently increasingly introduced in LLM evaluation, to assess the opinions, knowledge, and behaviors embedded in LLM models (e.g., Hendrycks et al., 2021b,a; Huang et al., 2023; Santurkar et al., 2023; Sravanthi et al., 2024; Ma et al., 2024, 2025).

The response of the LLM is then extracted using a string matching method using `RegEx` to map the responses to the 5 scale points. After all code pairs have been evaluated, we save the results in a table representing the similarities between the codes in matrix format.

Round 2: Candidate Code Selection. With the similarity results for all code pairs collected from the first round, the task of the second round is to find one final code partner for each code of codebook A. As the results from the first round are distributed across the five scale points A-E, we select the potential code matches by taking the codes rated with "A. Extremely similar" or "B. Very similar" to be the candidates for final selection. In case there are no A or B results, we consider that this source code does not have a matching code in codebook B.

We then prompt the LLM with the source code from codebook A and the candidate codes from codebook B (i.e., those that have a similarity result of "Extremely similar" or "Very similar"). We ask the LLM to select the code from the candidate codes B that matches A best, and extract the model output. In the end, we construct the final codebook for the mapped code pairs.

4 Experimental Setups

Data - The Danish Occupation Codes. We use the 6-digit, level 5 granularity of DISCO-LOEN[1] 88 and 08 from Statistics Denmark as codebook A and B respectively to test our framework. It is standard practice for crosswalks to be produced at the most granular level of a hierarchical code system to utilize the specificity of description. Mapping code pairs at lower levels of granularities can be aggregated to produce associations between codes in higher-level granularities (e.g. level 5 to level 4), but the reverse is not true.

[1] https://www.dst.dk/da/Statistik/dokumentation/nomenklaturer/disco-loen

Round 1

```
How similar do you think these two occupational classification codes are in terms of tasks, skills, and responsibilities? Please respond with only the option letter out of "A", "B", "C", "D", or "E".
Code1: {code1}
Code2: {code2}
Options:
A Extremely similar
B Very similar
C Moderately similar
D Slightly similar
E Not similar at all
```

Round 2

```
Below are a source occupation title (Occupation 1) and a number of potential matching occupation titles (Occupation 2).
For the given source occupation title, please select one matching occupation from the given Occupation 2 options.
Please select only one occupation option from Occupation 2, corresponding to the most similar matching occupation based on tasks, skills, and responsibilities.
Please only return the option letter of the selected match (A, B, C, ...) and don't say any other extra thing.
Source Occupation Title (Occupation 1): {code1}
Potential Matching Occupation Title(s) (Occupation 2): {options}
```

Figure 3: Prompts for the 2 rounds.

Ground-Truth Data. There are existing attempts at generating crosswalks between them for comparison. This includes a partial Many-to-1 crosswalk published by Statistics Denmark. The latter contains 332 code pairs linking DISCO-LOEN88 codes to 332 DISCO-LOEN08 code deemed equivalent by Statistics Denmark. Notably, as shown in Table 1, this crosswalk does not provide correspondences for all 570 and 559 level 5 codes in each codebook, leaving researchers to develop their own correspondences for the remaining codes, as conducted by Humlum (2021).

	Version 88	Version 08	Mapped Code Pairs
Count	570	559	332

Table 1: Summary of counts of the unique occupation codes in the codebooks and in the code mapping. Version 88 denotes the DISCO-LOEN88 codes and version 08 the DISCO-LOEN08 codes.

We use the partial Statistics Denmark crosswalk as ground truth mapping code pairs to evaluate the performance of our framework. Under our framework, every pairwise combination of codes from codebook A and B are potential mapping code pairs.

Models. We choose four instruction-tuned open-weight LLMs for conducting the experiments: Llama-3.1 8B (AI@Meta, 2024), Mistral 7B (Jiang et al., 2023), Gemma-2 9B (Team, 2024a), Qwen-2.5 7B (Team, 2024b).

Prompt Design. We design the prompt based on similar instructions and options given to the human participants in real surveys. The prompts used for the two rounds are presented in Figure 3.

Baseline. We compare our LLM-based framework to the approach using embeddings to find the most similar code for the given code, as applied in Liu et al. (2022) and Kononykhina et al. (2025). Since the data is in Danish, we use the multilingual version of the sentence transformers (Reimers and Gurevych, 2020). Specifically, we use the model for paraphrasing (`paraphrase-multilingual-MiniLM-L12-v2`). The basic workflow is this: For each code in the source code, it calculates the cosine similarity of the embeddings of the source code and every target

code; the target code with the highest similarity score to the source code is then selected as the mapped code for the target code.

Evaluation Metrics. We use the weighted F1 score to evaluate the model performance of our approach compared to the baseline. Further, as we apply different LLMs in our framework, we are also interested in how those models agree with each other while doing the crosswalks. Therefore, in further analysis, we calculate the inter-annotator agreement metric Cohen's Kappa (κ) to investigate the agreement between different LLMs.

5 Results

Main Results. Table 2 presents the main results of our framework applied to four LLMs and the embedding model baseline. Among the models evaluated, Qwen2.5 achieved the highest F1 score of 70.01%, indicating its strong ability to identify correct crosswalk mappings. This suggests that Qwen2.5 is particularly effective at capturing the semantic relationships between occupational codes in the Danish context. Gemma2 and Llama3.1 also demonstrated solid performance, with F1 scores of 67.35% and 61.25%, respectively, reflecting their capability for the task.

	Baseline	Gemma2	Llama3.1	Mistral	Qwen2.5
F1	57.12	67.35	61.25	40.58	70.01

Table 2: Main results of model performance in F1 (%) compared to the baseline.

Mistral, however, achieved an F1 score of only 40.58%, showing limited effectiveness in this specific application. This result may reflect differences in the architecture or training data of the model, which could make it less suited for nuanced crosswalk mapping tasks in Danish. The multilingual embedding model baseline attained an F1 score of 57.12%, performing better than Mistral but falling short of the other three instruction-tuned LLMs. These results highlight the advantages of instruction-tuned models for complex semantic tasks compared to traditional embedding-based methods.

Agreement Analysis. We next analyze the agreement between the four LLMs based on their final outputs. Figure 4 presents the heatmap of Cohen's Kappa scores, which measure the level of agreement between each model pair. Overall, the models exhibit relatively low agreement, with all Kappa scores falling below 60%. The Qwen2.5 model shows the highest agreement with the other models, which can be attributed to its better performance, as indicated by the results in Table 2. In contrast, the Mistral model shows more variability in their outputs, which is reflected in their lower Kappa scores.

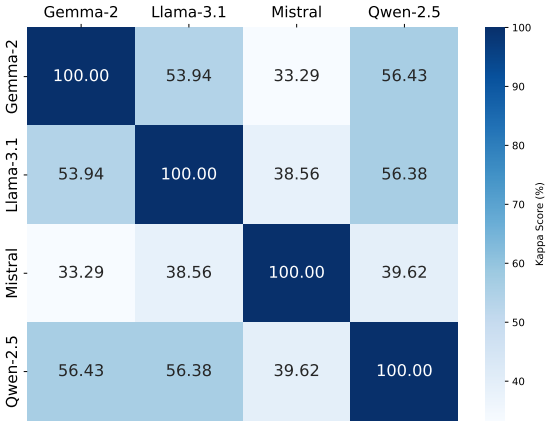

Figure 4: Kappa scores between models.

The Effect of Different Models in Two Rounds. Our framework operates in two rounds, where the results presented earlier assume that the same LLM is queried in both rounds. However, since the models are used independently in each round, we now investigate whether varying the models between rounds affects overall performance. Specifically, we explore whether swapping models leads to any significant changes in the results. The results, as shown in Figure 5, present the performance of different model combinations.

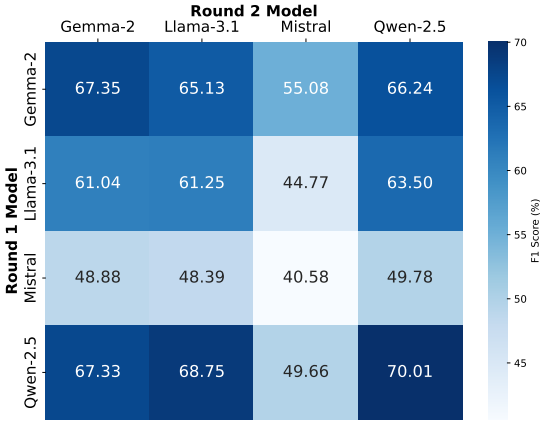

Figure 5: F1 results for experiments with different models in two rounds. X-axis: Round 1 models, Y-axis: Round 2 models.

Overall, the diagonal values in the table represent scenarios where the same model is used in both rounds, corresponding to the main results. These values generally indicate the highest or near-highest performance across rows and columns, suggesting that maintaining model consistency benefits performance. An exception is observed with Mistral, where using the same model in both rounds results in the worst performance. This reinforces Mistral's overall weaker effectiveness in the task, indicating that its predictions do not improve even when it has access to its own prior outputs.

Among the evaluated models, Qwen2.5 consistently outperforms others across different pairings, highlighting its robustness in identifying correct crosswalk mappings. Its closest competitor, Gemma2, also shows strong performance, particularly when paired with itself or with Qwen2.5. In contrast, Llama3.1 exhibits moderate performance, benefiting from combinations with stronger models but falling short of top-tier results.

These findings suggest that performance is optimized when stronger models like Qwen2.5 and Gemma2 are used consistently. Swapping models, especially involving Mistral, tends to reduce effectiveness, highlighting the importance of model selection.

6 Discussion & Conclusion

The results of this study demonstrate the potential of LLMs to assist in creating crosswalks for occupational classifications. Our findings highlight the advantages of instruction-tuned LLMs in handling semantic complexity and improving efficiency compared to traditional embedding-based approaches. Models like Qwen2.5 showed strong performance in aligning Danish occupational codes, emphasizing the value of instruction tuning and contextual understanding in these tasks.

However, the relatively low inter-model agreement underscores the variability in outputs across different LLMs, pointing to the importance of model selection and parameter tuning. This variability also highlights the need for integrating human expertise into the workflow to validate and refine LLM-generated mappings. The interactive, prompt-based framework we proposed aligns with the concept of human-in-the-loop workflows, where LLMs augment rather than replace expert judgment.

Additionally, our findings highlight the advantages of maintaining model consistency across rounds, especially for strong models like Qwen2.5. Swapping models, particularly when involving weaker ones like Mistral, leads to diminished results, emphasizing the need for robust and consistent modeling strategies.

Our findings also resonate with similar efforts in other domains, such as MapperGPT, which refines entity mappings in fields like healthcare and biomedical research (Matentzoglu et al., 2023). These parallels reinforce the versatility of LLMs in supporting knowledge-mapping tasks across diverse contexts, though domain-specific adaptations remain critical for success. Future work could explore how our two-step prompting framework can be extended beyond occupational classifications to other classification mapping tasks in fields such as finance, education, and public administration, where structured yet flexible mappings are essential for accurate data integration and interoperability.

7 Limitations

Despite the promising results, this study has several limitations. First, the reliance on Danish occupational codes limits the generalizability of our findings to other languages and classification systems. Future studies should investigate the performance of LLMs on crosswalks involving additional languages and classification schemes, such as ISCO and SOC.

Second, the use of multiple-choice questions to evaluate LLMs may introduce biases inherent to this format, such as response tendencies (Li et al., 2024; Pezeshkpour and Hruschka, 2024; Wang et al., 2024). Further exploration of alternative evaluation frameworks, such as open-ended prompting or pairwise ranking, could provide more robust insights into LLM performance.

8 Ethical Considerations

The use of LLMs for creating crosswalks must consider potential biases (e.g., regarding gender) in the models, which could lead to inaccurate or inequitable mappings, especially for underrepresented groups (Touileb et al., 2023; Nghiem et al., 2024; Sancheti et al., 2024). Ensuring human oversight is crucial to validate and refine LLM outputs, preventing the propagation of errors that may impact labor market analyses or policy decisions.

References

AI@Meta. 2024. Llama 3.1 model card.

U.S. Bureau of Labor Statistics BLS. 2025. Standard occupational classification (soc). Accessed: 2025-01-25.

Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. Open Science Foundation.

Yi-Yun Cheng, Nico Franz, Jodi Schneider, Shizhuo Yu, Thomas Rodenhausen, and Bertram Ludäscher. 2017. Agreeing to disagree: Reconciling conflicting taxonomic views using a logic-based approach. *Proceedings of the Association for Information Science and Technology*, 54(1):46–56.

Robert M. Groves. 2011. Three eras of survey research. *Public Opinion Quarterly*, 75(5):861–871.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Anders Humlum. 2021. Crosswalks between (d)isco88 and (d)isco08 occupational codes.

International Labour Organization ILO. 2025. Classification of occupations: Concepts and definitions. Accessed: 2025-01-25.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Olga Kononykhina, Malte Schierholz, and Frauke Kreuter. 2025. Can large language models advance occupational coding? evidence and methodological insights. Unpublished Manuscript.

Lynda Laughlin, Xi Song, Megan Wisniewski, and Jiahui Xu. 2024. From job descriptions to occupations: Using neural language models to code job data.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834, Torino, Italia. ELRA and ICCL.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Junhua Liu, Yung Chuen Ng, Zitong Gui, Trisha Singhal, Lucienne T. M. Blessing, Kristin L. Wood, and Kwan Hui Lim. 2022. Title2vec: A contextual job title embedding for occupational named entity recognition and other applications. *Journal of Big Data*, 9:99.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *Preprint*, arXiv:2502.12378.

Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. 2024. The potential and challenges of evaluating attitudes, opinions, and values in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8783–8805, Miami, Florida, USA. Association for Computational Linguistics.

Nicolas A. Matentzoglu, John Harry Caufield, Harshad B. Hegde, Justin T. Reese, Sierra A T Moxon, Hyeongsik Kim, Nomi L. Harris, Melissa A. Haendel, Christopher J. Mungall, and Robert Bosch. 2023. Mappergpt: Large language models for linking and mapping entities. *ArXiv*, abs/2310.03666.

Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé Iii. 2024. "you gotta be a doctor, lin" : An investigation of name-based bias of large language models in employment recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7268–7287, Miami, Florida, USA. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

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

Thomas Rémen, Lesley Richardson, Corinne Pilorget, Gilles Palmer, Jack Siemiatycki, and Jérôme Lavoué. 2018. Development of a coding and crosswalk tool for occupations and industries. *Annals of work exposures and health*, 62(7):796–807.

Parisa Safikhani, Hayastan Avetisyan, Dennis Föste-Eggers, and David Broneske. 2023. Automated occupation coding with hierarchical features: A data-centric approach to classification with pre-trained language models. *Discover Artificial Intelligence*, 3:6.

Abhilasha Sancheti, Haozhe An, and Rachel Rudinger. 2024. On the influence of gender and race in romantic relationship prediction from large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 479–494, Miami, Florida, USA. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Malte Schierholz and Matthias Schonlau. 2020. Machine learning for occupation coding—a comparison study. *Journal of Survey Statistics and Methodology*, 9(5):1013–1034.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Statistics Denmark. 2025a. DISCO-08 Classification - Statistics Denmark. https://www.dst.dk/en/Statistik/dokumentation/nomenklaturer/disco-loen. Accessed: 2025-01-24.

Statistics Denmark. 2025b. DISCO Classification - Statistics Denmark. https://www.dst.dk/en/Statistik/dokumentation/nomenklaturer/disco?id=ec4f3246-ea1a-4e8b-b229-f03c0dc680c6. Accessed: 2025-01-24.

Mega Subramaniam, June Ahn, Amanda Waugh, Natalie Greene Taylor, Allison Druin, Kenneth R Fleischmann, and Greg Walsh. 2013. Crosswalk between the" framework for k-12 science education" and" standards for the 21st-century learner": School librarians as the crucial link. *School Library Research*, 16.

Gemma Team. 2024a. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2403.05530.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2023. Measuring normative and descriptive biases in language models using census data. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2242–2248, Dubrovnik, Croatia. Association for Computational Linguistics.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.