

International Journal of Population Data Science

Journal Website: www.ijpds.org



Swansea University
Prifysgol Abertawe

Use of Large Language Models (LLM) for Occupational Classification: An Example of Early Learning & Child Care (ELCC) Workers

Maria Ahmed^{1,*}, Kristyn Frank², Anne Guèvremont², Dafna Kohen², and Joanne Yoon^{3†}

Submission History

Submitted:	07/10/2025
Accepted:	29/07/2026
Published:	19/10/2026

¹ Department of Sociology,
University of Western Ontario,
London, Ontario, Canada

² Health Analysis and Modelling
Division, Statistics Canada,
Canada

³ Artificial Intelligence Methods
Division, Statistics Canada,
Canada

Abstract

Introduction

The use of Artificial Intelligence (AI) for occupational classification has drawn the attention of statistical agencies because manual coding is time-consuming and expensive. In Canada, the national “\$10-a-day” Early Learning and Child Care (ELCC) program has increased the need for more detailed information on the ELCC workforce.

Objectives

Given the successful use of Machine Learning (ML) for occupational coding in open-ended Census write-in responses, this study explored the feasibility of creating an in-context ML classification protocol with a particular focus on Canada’s growing ELCC sector.

Methods

Using Canada’s 2021 Census of Population, write-in responses ($n = 57,135$) were classified using a large language model (LLM) for few-shot learning classification, an ML paradigm where a model is instructed to perform a task with a limited number of examples. Data were processed with Meta AI’s Llama 2 model.

Results

Existing National Occupational Classification (NOC) and North American Industrial Classification System (NAICS) categories are not sufficiently specific for the identification of a group of ELCC workers that is meaningful for policymakers. Through subject-matter expertise and iterative LLM classification reviews, NOC occupational groups were refined into several subgroups to better classify jobs relevant to the sector. Assigning workers to these refined subgroups would otherwise have been an inefficient and onerous undertaking through typical manual classification methods.

Conclusions

This study demonstrates the utility of LLM classification as an efficient and time saving tool. However, the tool cannot offer improvements in isolation and needs to be complemented by subject-matter expert input. The use of LLM resulted in a more granular and meaningful occupational classification for ELCC workers.

Keywords

Large Language Models, Census of Population, Canada, Child care, Classification, Workforce

*Corresponding Author:

Email Address: Kristyn.frank@statcan.gc.ca (Kristyn Frank)

†Authors are listed alphabetically

Introduction

Artificial Intelligence (AI) has had a substantial impact on generating efficiencies in many public sectors, including healthcare, education, and intragovernmental institutions [1, 2]. Consequently, interest in how National Statistical Organisations (NSOs) can leverage AI through Machine Learning (ML) and Large Language Models (LLM) to improve applications and processes in official statistics [3, 4] has increased. For example, NSOs have used AI to improve manual data processing and classification tasks such as automating the production of the Consumer Prices Index [5] and coding of open-ended survey questions [6].

The use of AI for occupational coding has drawn the attention of statistical agencies due to the time and cost intensive nature of manual and computer-assisted classification [7]. AI methodologies have evolved to automate these tasks [7, 8]. While these methods have shown promise in increasing efficiency, more recent LLM strategies, such as zero-shot data extraction with community-fine-tuned large language models (CLLMs) or pre-trained natural language processing (NLP) models such as the Bidirectional Encoder Representations from Transformers (BERT), showcase the ability to automate occupational coding with little, if any, labelled data [9, 10]. These developments situate occupational coding within the broader literature on automated text classification, where ML and NLP methods are widely used to categorise large collections of textual data, including topic modelling and supervised text classification [11–13].

One area of interest for NSOs is the classification of occupational data from surveys to produce accurate and timely labour market estimates [14]. With the advent of the Canada-wide “\$10-a-day” Early Learning and Child Care (ELCC) system, the federal government established bilateral agreements with each province and territory, aiming to create 250,000 new child care spaces by March 2026 [15]. As both dual-earning and lone-parent families have increased, demand for child care spaces has risen [16], leading to greater need for ELCC workers. Canada’s ELCC workforce consists of “formal” providers working in centres or in private homes, and “informal” providers such as relatives, nannies, and babysitters [16]. For policy and planning, this signals a need for accurate, timely estimates of the ELCC labour force to better understand demand for different worker types.

The ELCC workforce is typically defined by two occupational groups in Canada’s National Occupational Classification (NOC): Early Childhood Educators and Assistants (NOC 42202) and Home Child Care Providers (NOC 44100) [16–19]. While these groups have changed over time [20–22], the categories remain broad and encompass a wide range of roles. As such, they do not capture distinctions meaningful for policy and workforce planning, such as differences between formal and informal providers or workers in centre- versus home-based care. This gap has motivated efforts to develop more granular classifications of ELCC workers. For example, Guèvremont [18] used Canadian Census and Labour Force Survey (LFS) data to identify six subcategories of ELCC workers based on industry, self-employment status, and place of work, distinguishing roles previously grouped together (e.g., early childhood educators in kindergarten classrooms, nannies, and workers providing

drop-in care in commercial establishments). However, key policy-relevant distinctions, such as whether care is provided in licensed centres or homes, cannot be fully identified by existing classification systems.

Generally, NSOs use responses to national labour force surveys or the Census to manually code thousands of records according to standard classification systems such as the NOC or the North American Industry Classification System (NAICS) [8, 14, 23]. Statistics Canada has increasingly automated this process using ML to classify occupational codes from the LFS and the Census using fastText, a supervised text classification model trained on approximately 1.5 million previously coded records [14]. This workflow maps open-ended responses onto predefined occupational classification codes. While this approach achieves high accuracy with large, labelled datasets, it has limitations: it relies on substantial training data and is optimised for pre-existing categories.

Building on the successful use of ML for coding write-in responses, this study develops and evaluates an ML classification protocol focused on Canada’s ELCC sector, where more granular, policy-relevant information is needed. This sector was selected because existing methods cannot differentiate its nuanced occupational settings and emerging workforce roles, distinctions important for workforce planning and policy development [16, 17]. The ELCC sector therefore provides a useful case study for domains where occupational roles span formal and informal settings, job titles and descriptions vary substantially, and classification systems risk misclassification [7, 11]. The objective of this study was to explore the feasibility of using an LLM with an in-context learning approach to categorise census write-in responses into more granular ELCC occupational groupings. Rather than comparing model performance, this study makes a unique contribution by examining an iterative classification protocol that integrates subject-matter expertise to refine occupational categories where training data are limited, classification schemes are evolving, and distinctions between formal and informal roles are difficult to capture.

Methods

This study used the 2021 Census Response Database (RDB), an internal unweighted dataset that includes write-in responses to the long-form questionnaire of the 2021 Census of Population, a national census conducted every five years. The questionnaire was administered to 25% of Canadian households and included questions about age, gender, sex, languages spoken, household relationships, and detailed questions about respondents’ labour force characteristics, work activities, and occupations. The data were processed through Large Language Model Meta AI 2 (Llama 2) [24], a high-performing open-source model developed by Meta AI.

Variables

The occupation information used in this study included both write-in responses and coded variables. For employed or self-employed persons in the current or previous year, three write-in questions were used to categorise workers’ occupation and industry using the NOC and NAICS: “What kind of business,

industry or service was this?”, “What was this person’s work or occupation?”, and “In this work, what were this person’s main activities?”

In the NOC, **Early Childhood Educators and Assistants** (NOC 42202) are workers who “plan, organize and implement programs for children between the ages of infancy and 12 years” and “are employed in child care centres, daycare centres, kindergartens, agencies for exceptional children and other settings where early childhood education services are provided” [19]. **Home Child Care Providers** (NOC 44100) are individuals who “care for children on an ongoing or short-term basis” and “are employed primarily within their homes or in the children’s home” [19]. These categories encompass all ELCC workers but capture a broader group than those of interest to policy makers and researchers.

Another way to identify ELCC workers in the Census is by industry using the NAICS, developed by the statistical agencies of Canada, Mexico, and the United States. The 2021 Census utilised the 2022 NAICS. The daycare services industry (NAICS 6244) “comprises establishments primarily engaged in providing daycare services for infants or children. These establishments may care for older children when they are not in school and may also offer pre-kindergarten educational programs.” However, this classification alone is inadequate, as it includes individuals who do not directly work with children (e.g., cleaners, cooks) and excludes ELCC workers who may be classified in other industries (e.g., private households). Nonetheless, NAICS remains useful for identifying the *types of industries* employing ELCC workers, such as child care centres or schools.

Other important attributes for classifying the nature of ELCC workers include employment status (employed or self-employed), whether they work at home, and education [18]. Beach [25] recommended classifying child care workers as working at a usual place of work or at home, with centre-based care represented by those working at a usual place of work and home-based care by working at home. Such additional information can improve classification accuracy.

Analytic Strategy

This study used a rule-based filter alongside an LLM text classifier. LLMs are AI systems designed to process and generate human-like text. They are pre-trained on massive data sets to comprehend context and language nuances with the goal of generating coherent and contextually relevant responses [26]. LLMs inherently capture intricate patterns and semantic relationships in data to tackle various natural language tasks well, which was the ideal methodology for this study.

A set of rules largely based on previous work [18] were used to filter the possible subgroups for the Census write-in responses using the known structure of ELCC workers. These rules used Census variables such as occupation, industry, and location of work. Other categories were introduced iteratively to further refine subgroups based on expert subject-matter input. The LLM text classifier then assigned the write-in responses to the most appropriate subgroup category. Figure 1 illustrates the process flow.

Since LLMs outperform traditional ML models in few-shot classification due to extensive pre-training on large, diverse

datasets, this methodology was selected. In this paradigm, known as in-context learning, the model performs a task (classification) using only instructions and a small number of examples, far fewer than traditional models require. Tasks are performed by prompting the LLM instead of retraining or tuning it. It is preferred over traditional approaches, especially by subject-matter experts, because it enables models to rapidly grasp nuances and generalise to new cases without extensive data labelling.

This study used open-source LLM model Llama 2 which, at its release, outperformed open-source chat models on most benchmarks tested and in human evaluations for helpfulness and safety. Llama 2 was a suitable substitute for closed source models [24] and was used in a protected environment, accessed through Application Programming Interface (API) calls within the internal Statistics Canada network, ensuring control over data access and no risk of unauthorised exposure associated with cloud-based solutions. The LLM was hosted on a dedicated graphics processing unit (GPU) computer with 256 GB of memory and dual Quadro GPU, supported by a 1 TB solid state drive for system operations. The classification program ran on a separate computing environment with two central processing unit (CPU) processors and 8 GB of memory, which was sufficient to interact with the Llama 2 API. Individuals’ data were not accessible to Meta AI or incorporated into external training datasets.

To classify job descriptions into meaningful occupational subgroups, prompts included job titles and representative write-in responses describing main work activities for each subgroup (see Supplementary Appendix 1). These prompts guided the model by linking job roles with appropriate examples from the write-in responses. The validity of the LLM’s predictions was assessed through manual verification by subject-matter experts using samples of 250 responses per category, randomly selected across six iterations (4-6 categories each). Prompts were supplied in English, while example and input write-ins were provided in English and French, Canada’s two official languages. This highlights a key advantage of LLMs; their extensive multilingual pre-training enables classification across languages without additional model tuning.

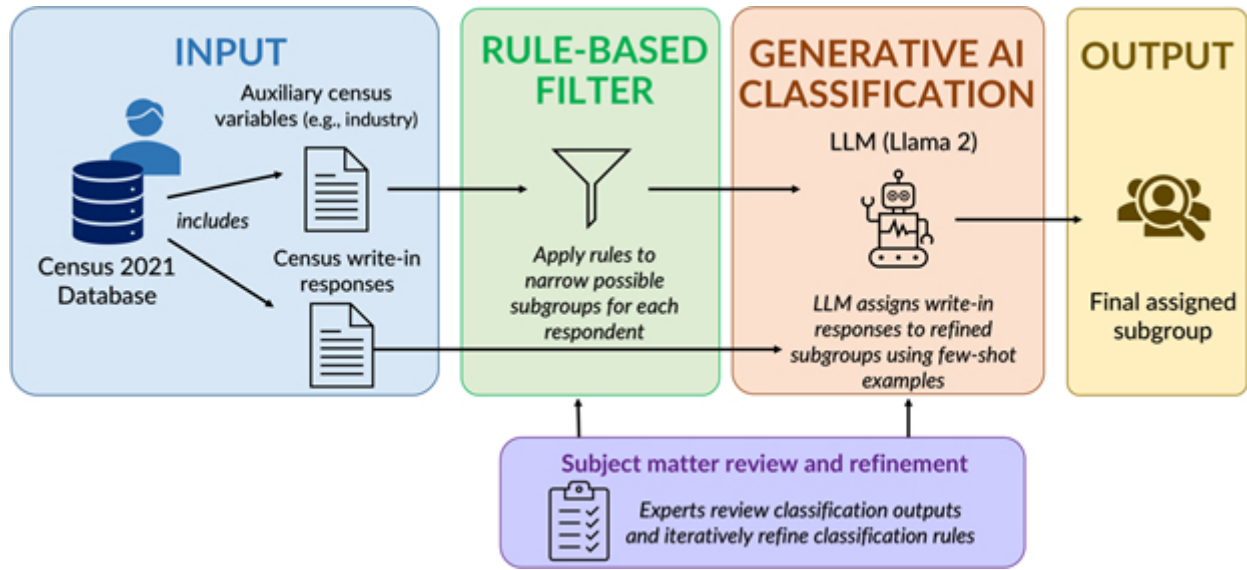
False positives, where write-in responses did not align with the expected subgroup, were addressed through expert review and category refinement. False negatives arose when overly strict parameters (e.g., education criteria), caused the model to exclude valid entries. The guiding prompts were then revised iteratively, with each round reducing misclassifications. For example, ambiguity was reduced by adding more labelled examples to the prompt or creating new subgroups to capture distinct write-in responses.

Initial analyses included NOC categories 42202 and 44100; however, Early Childhood Educators and Assistants (NOC 42202) were used as a policy-relevant example to focus the results. Results were based on unweighted data (n=57,135); therefore, they should be considered exploratory and interpreted with caution.

Results

Rule-based filters were initially applied to sort each respondent into a possible subgroup, based on a combination of

Figure 1: Process Flow for Rule-Based and AI-Assisted Classification of ELCC Worker Subgroups



Note: Census 2021 Database refers to the 2021 Canadian Census Response Database, a confidential data set which is not publicly available.

occupation (NOC), industry (NAICS), and other census information (e.g., education level, work from home status, self-employment status). Based on subject-matter expertise and prior Census-based work [18], four subgroups were proposed within the NOC group “Early Childhood Educators and Assistants”:

- (1) Early Childhood Educators (ECEs) in Daycare Centres: Worked in the daycare services industry (NAICS 6244), did not work at home, were paid employees, and had a college diploma or higher. Example job titles in the write-in responses included “Early Childhood Educator” or “ECE”, “Registered Early Childhood Educator” or “RECE”, “Daycare worker”, and “Child care worker”.
- (2) Early Childhood Educator Assistants in Daycare Centres: These workers had a similar profile to ECEs but were expected to have lower levels of education (i.e., did not have a college diploma or higher). Example job titles included “Early childhood assistant”, “Early childhood worker”, “Daycare worker”, and “Child care worker.”
- (3) Early Childhood Educators in kindergarten classrooms: These workers were similar to (1) ECEs but worked in schools (NAICS 6111) rather than the daycare services industry. Example job titles included “Early Childhood Educator” or “ECE”, “Registered Early Childhood Educator” or “RECE.”
- (4) School-based child care providers (i.e., workers in a child care centre located in a school OR a before- and after-school program): These workers were similar to (3) ECEs in kindergarten classrooms but were expected to have lower education levels and job titles like “Extended day program worker” or “After school camp counselor.”

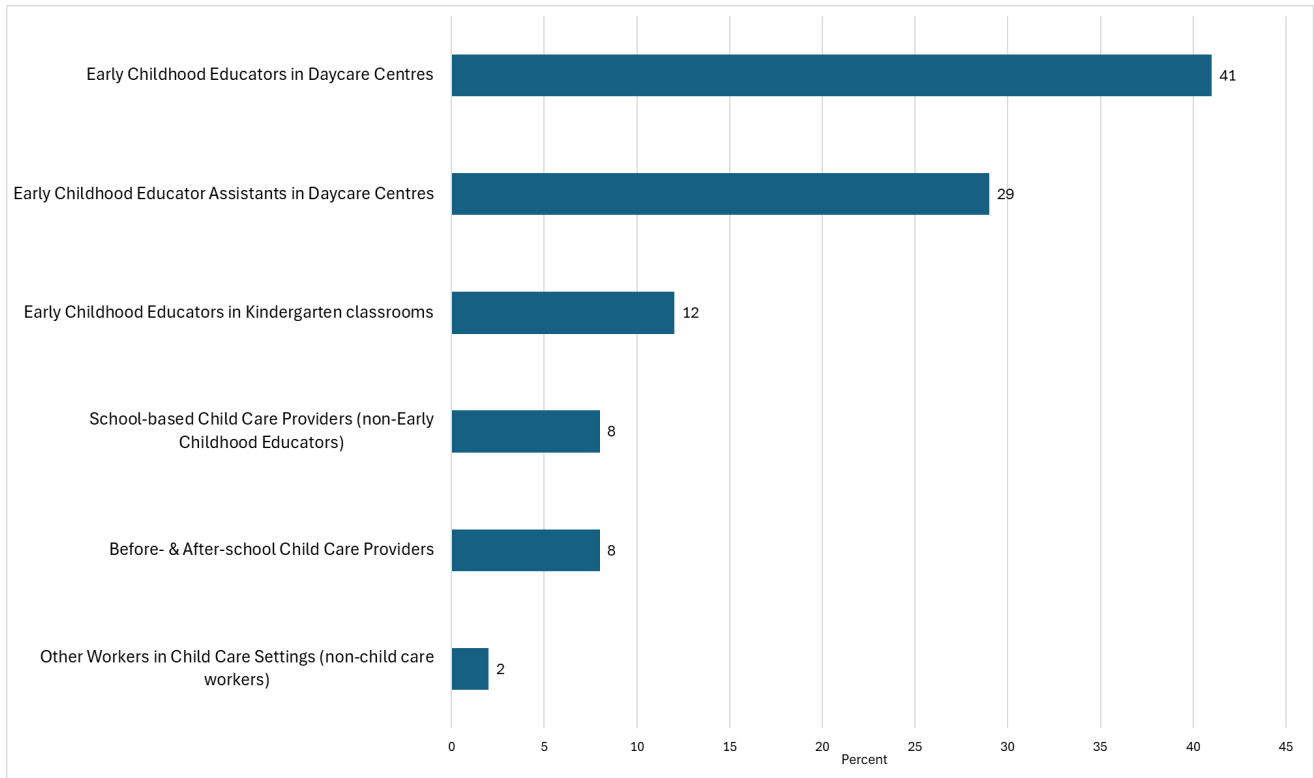
Following this, an LLM text classifier categorised Census write-in responses to the best subgroup. The LLM processed the

results in about 2 seconds per entry, compared to manual coding by a human of about 30 seconds per entry. For the 57,135 responses, this would equate to 30 hours for the LLM compared to about 500 hours for manual coding (about 12 weeks of 40 hours/week of work for one human coder). However, expert review is still needed for the LLM, requiring another 20 hours.

Subject-matter experts conducted a collective manual review of a sample of about 250 responses in each category. Differences in categorisation were resolved by consensus. Initial results were relatively accurate (92% accuracy compared to manual classification by a subject-matter expert), but subject-matter review revealed responses (8%) that could be better categorised. For example, many workers in the “Daycare Centre” category had job titles or activities that indicated they provided before- and after-school care. Some responses for occupation were “after school worker” and “before and after school staff” and/or with a type of business of “after-school program” or “after school education program”. As these job categories are not early childhood educators, a new subgroup for before- and after-school child care providers was created. Job titles and activities suggesting management positions (e.g., “director”, “supervisor”) or other non-child care activities (e.g., “cleaning”, “kitchen help”) indicated many were likely not working directly with children. Therefore, another subgroup was created: “Other workers in child care settings (non-child care).”

The manual review also examined the classification of misspelled words and compared the accuracy of English and French entries. Misspelled entries appeared to be classified correctly. French text entries were classified with similar or higher accuracy (89% to 98% accuracy depending on subgroup) compared to English text entries (81% to 95% accuracy depending on subgroup). Five further iterations were conducted, each supplying the LLM with additional labelled examples, creating new subgroups and/or including other variables to refine rule-based filters. Each iteration involved

Figure 2: Final Distribution of Derived Subgroups for Write-In Responses from Workers Coded to Early Childhood Educators and Assistants NOC Group 42202



Source: 2021 Canadian Census Response Database; the proportion of workers in each sub-category does not represent the actual percentage of these workers in Canada.

manual verification by subject-matter experts with a random sample of about 250 cases per subgroup.

Field of study and weekly work hours were also tested as rule-based filters. Field of study was considered to identify workers trained in early childhood education; however, Census data report only the highest level of education, limiting information about respondents with higher education in unrelated disciplines. For example, immigrants, who comprise a notable share of Canada’s ELCC workforce [16], often work in jobs for which they are overqualified due to difficulties with foreign credential recognition [27]. Manual verification revealed several misclassifications after including field of study, leading to its removal from the rule-based filters. Similarly, weekly hours were considered to identify before- and after-school child care workers, based on the assumption they may work fewer hours than other ELCC workers. Manual verification again identified several misclassifications, and this variable was also excluded. After five iterations, the resulting final categorisations are summarised in Figure 2.

The accuracy of the final categorisations was compared to manual classification by subject-matter experts. The overall accuracy decreased compared to the original classification (82% vs. 92%). However, for the original four categories, the accuracy improved and ranged from 96% to 99% in the final classification. The accuracy of the two new categories was poor: 21% accuracy for “Before- and after- school child care providers” and 65% accuracy for “Other workers in child care settings (non-child care workers)”. Low accuracy was likely due to vague write-in responses. For example, many

of the miscategorised responses for occupation were “child care worker” or “child care provider”, with the main activity of “daycare” or “taking care of children”. Many similar responses were categorised correctly as ECEAs, either in daycare centres or schools. Although the accuracy of the new categories was poor, the exercise was stopped after five iterations due to project scope and time constraints. Accuracy may have been improved with further iterations by providing more training examples for the model.

While the data are unweighted and not representative of the actual percentage of ELCC workers in Canada, the resulting proportions offer useful insights into the distribution of the ELCC workforce. Most ELCC workers in NOC 42202 were daycare centre workers, with the highest proportion being ECEs and a lower proportion being ECE assistants (Figure 2). The remaining workers were ECEs in kindergarten classrooms, before- and after-school child care providers, school-based child care providers, with a very small percentage classified as other workers in child care services.

Discussion

This study demonstrates how an LLM can be applied, using an in-context learning approach, to classify open-ended Census write-in responses into more granular occupational categories. Further refinement of ELCC worker categories – such as identifying those employed in child care centres – was particularly important given current policy priorities.

A key finding is that the LLM could not derive appropriate classifications without the input of subject-matter expertise, which is essential for defining and refining more meaningful occupation subclassifications. However, this study highlights the potential to build upon the current NOC by introducing new subcategories to address information gaps in priority areas related to the Canada-wide ELCC system.

Although Statistics Canada has used ML, such as fastText [14], to accurately assign NOC and NAICS codes, this study demonstrates the potential of combining subject-matter expertise with the efficiency of LLM-based classification that does not require large, labelled training datasets. Because the LLM was not trained on domain-specific data, its performance in classifying nuanced occupational responses may be limited; however, the semi-structured approach and expert review likely helped mitigate this. This approach enables more detailed occupational classification that would otherwise be misclassified using existing NOC codes alone. The large number of respondents and unstructured nature of open-ended survey responses supports the classification of large quantities of open-ended data into meaningful categories. This, in turn, yields more meaningful occupational information for policy planning, including training and recruitment.

Another efficiency identified is the ability of LLMs to “understand” and apply instructions in multiple languages (in this case, English and French). As LLMs are pre-trained on large multilingual datasets, they can recognise and process many languages simultaneously – an advantage for classification tasks. French entries were classified with similar or higher accuracy than English entries, possibly due to Quebec’s more established, province-wide child care system (since 1997), relative to the more recent Canada-wide ELCC program. Additionally, the LLM seemed to handle common misspellings effectively.

This study also provides a methodology for applying few-shot learning classification within official statistics for labour-intensive manual classification projects. As the model uses a limited number of examples to infer how to perform the classification task, this approach provides significant time and cost savings compared to traditional methods. It captures nuances without the burden of labelling large amounts of examples or using greater server resources for data input and processing. In this study, a sample of 57,135 total responses were analysed (with 250 responses in each of 6 categories used for manual verification); this is relatively small considering the vast amounts of training data required for other predictive ML algorithms [9]. For example, fastText training data included more than 1.5 million records and a further 200,000 records were retained as a validation dataset [14].

The primary limitation was the LLM’s reliance on substantial subject-matter expert input, both preceding and during text classification. Compared to traditional attribute-based classification methods [18], the time and expertise required was substantial. While reviewing multiple iterations improved understanding of how the LLM applied the rules, it was time consuming. This limitation may be an artefact of the area of study (i.e., ELCC workers) where responses vary widely. However, this approach may benefit other fields, such as dentistry, nursing, or other healthcare professions, where job titles and descriptions also vary widely and support roles are often reported alongside primary occupations.

Some variables considered (field of study, weekly work hours) did not help to improve classification. In the census, field of study reflects respondents’ highest level of education, which may not capture ELCC training, though it may be more useful in industries where highest degree aligns with the required disciplinary background. Weekly work hours varied within groups, limiting its usefulness; this variable may be more informative in industries where hours of work differ across worker subcategories. Researchers should be familiar with their data and recognise the data needed for nuanced categories may be unavailable.

This study leveraged existing subclasses and examples to better guide the LLM. Although initially time consuming, zero-shot classification using instructions and categories should significantly reduce subject-matter experts’ time in future applications by not having to create examples. Future work could also compare different models to enable more rigorous evaluation of classification performance. Additionally, an inductive approach without rule-based filters would allow the model to infer the categories. This may be a promising direction for researchers and limit the amount of input required. Newer models may be able to improve classification accuracy for difficult categories; the very low accuracy for the “Before- and after- school child care providers” group was a limitation, but due to Statistics Canada’s closed environment, this study was limited to the use of the Llama 2 model.

Conclusion

This study demonstrates the ability and necessity to bridge subject-matter expertise with the efficiencies of LLM classification to produce more nuanced occupational classification. This approach improves classification granularity while balancing analytic resources by limiting the burden of manually classifying large volumes of data. The LLM alone was insufficient and depended on expert input; however, it can assist subject-matter experts in testing and iteratively refining new classification categories. Within the context of ELCC workers in Canada, this study proposes a more granular categorisation that addresses information gaps. Future studies could evaluate whether this LLM protocol is effective for other occupational groups, and in other policy contexts.

Ethics Statement

This study did not require ethical approval, as it was conducted in accordance with the *Statistics Act* [28].

Statement on Conflicts of Interest

The authors declare that they have no competing interests relevant to the content of this article.

Data Availability Statement

The dataset used in this study is not publicly available as it is internal to Statistics Canada and based on unweighted data.

Census of Population data are available through Public Use Microdata Files or Statistics Canada Research Data Centres.

AI Disclosure Statement

The authors used ChatGPT (OpenAI) for minor language editing to improve sentence clarity and to assist with editing and improving an existing figure. All changes were reviewed and verified by the authors.

References

1. Abbas SW, Hamid M, Alkanhel R, Abdallah HA. Official statistics and big data processing with artificial intelligence: capacity indicators for public sector organizations. *Systems*. 2023;11(8):424. <https://doi.org/10.3390/systems11080424>
2. Faisle D. The potential of artificial intelligence for the SDGs and official statistics. PARIS21 Working Paper [Internet]. 2024 Apr [cited 2025 Sep 9]. Available from: https://www.paris21.org/sites/default/files/related_documents/2024-04/the-potential-of-ai-for-the-sdgs-and-official-stats_working-paper.pdf.
3. Kloos K. A new generic method to improve machine learning applications in official statistics. *Statistical Journal of the IAOS*. 2021;37(4):1181–96. <https://doi.org/10.3233/SJI-210885>
4. Molladavoudi S, Bedard ME. Data science and technology trends for official statistics: opportunities and challenges. In: Liebowitz J, editor. *Pivoting government through digital transformation*. New York: Auerbach Publications; 2023. <https://doi.org/10.1201/9781003369783>
5. Martindale H, Rowland E, Flower T, Clews G. Semi-supervised machine learning with word embedding for classification in price statistics. *Data & Policy*. 2020;2:e12. <https://doi.org/10.1017/dap.2020.13>
6. Rezaei Ghahroodi Z, Ranji H, Rezaei A. Using machine learning classification algorithms in official statistics. *Journal of Statistical Sciences*. 2021;15(1):119–46.
7. Schierholz M, Schonlau M. Machine learning for occupation coding—A comparison study. *Journal of Survey Statistics and Methodology*. 2021;9(5):1013–34. <https://doi.org/10.1093/jssam/smaa023>
8. Takahashi K, Taki H, Tanabe S, Li W. An automatic coding system with a three-grade confidence level corresponding to the national/international occupation and industry standard—open to the public on the web. In: *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (IC3K 2014)*. SciTePress; 2014. p. 369–75. <https://doi.org/10.5220/0005131703690375>
9. Gweon H, Schonlau M. Automated classification for open-ended questions with BERT. *Journal of Survey Statistics and Methodology*. 2024;12(2):493–504. <https://doi.org/10.1093/jssam/smad015>
10. Kazi N, Kahanda I, Rupassara SI, Kindt JW. Zero-shot information extraction with community-fine-tuned large language models from open-ended interview transcripts. In: *Proceedings of the 2023 International Conference on Machine Learning and Applications (ICMLA)*. IEEE; 2023. p. 932–37. <https://doi.org/10.1109/ICMLA58977.2023.00138>
11. Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys*. 2002;34(1):1–47. <https://doi.org/10.1145/505282.505283>
12. Griffiths TL, Steyvers M. Finding scientific topics. *Proceedings of the National Academy of Sciences*. 2004;101(Suppl 1):5228–5235. <https://doi.org/10.1073/pnas.0307752101>
13. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown DE. Text classification algorithms: A survey. *Information*. 2019;10(4):150. <https://doi.org/10.3390/info10040150>
14. Evans J, Oyarzun J. Need for speed: using fastText (machine learning) to code the labour force survey. In: *Proceedings of the Statistics Canada Symposium 2021* [Internet]. Ottawa: Statistics Canada; 2021 [cited 2025 Sep 9]. Available from: <https://www150.statcan.gc.ca/n1/pub/11-522-x/2021001/article/00013-eng.pdf>.
15. Employment and Social Development Canada. Toward \$10-a-day: early learning and child care [Internet]. 2022 [cited 2025 Jan 3]. Available from: <https://www.canada.ca/en/employment-social-development/campaigns/child-care.html>.
16. Uppal S, Savage K. Child care workers in Canada. Statistics Canada [Internet]. 2021 Jun 25 [cited 2025 Jan 3]. Available from: <https://www150.statcan.gc.ca/n1/en/pub/75-006-x/2021001/article/00005-eng.pdf>.
17. Frank K, Arim R. Indigenous and non-Indigenous early learning and child care workers in Canada. *Economic and Social Reports*. 2021;1(8). <https://doi.org/10.25318/36280001202100800005-eng>
18. Guèvremont A. Workforce indicators: can we report on sub-groups of child care workers? Ottawa: Statistics Canada; 2021 Mar. [unpublished report].
19. Statistics Canada. National Occupational Classification (NOC) 2021 version 1.0 [Internet]. Ottawa: Statistics Canada; 2023 [cited 2025 Sep 9]. Available from: <https://www.statcan.gc.ca/en/subjects/standard/noc/2021/indexV1>.
20. Statistics Canada. Standard Occupational Classification 1980 [Internet]. Ottawa: Statistics Canada; 1981 [cited 2025 Sep 9]. Available from:

- https://publications.gc.ca/collections/collection_2017/statcan/CS12-565-1981-eng.pdf.
21. Statistics Canada. Occupation according to the 1991 Standard Occupational Classification [Internet]. Ottawa: Statistics Canada; 1995 [cited 2025 Sep 9]. Available from: https://publications.gc.ca/collections/collection_2012/statcan/rh-hc/CS92-344-1995-eng.pdf.
 22. Statistics Canada. National Occupational Classification (NOC) 2011 [Internet]. Ottawa: Statistics Canada; 2018 [cited 2025 Jan 23]. Available from: <https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=122372&CVD=122376&CPV=4411&CST=01012011&CLV=4&MLV=4f>.
 23. Stelmack A. Integrating machine learning into coding of the 2021 Canadian census using fastText. In: Proceedings of the Statistics Canada Symposium 2021 [Internet]. Ottawa: Statistics Canada; 2021 [cited 2025 Sep 9]. Available from: <https://www150.statcan.gc.ca/n1/pub/11-522-x/2021001/article/00010-eng.pdf>.
 24. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. arXiv [Internet]. 2023 [cited 2025 Mar 7]. Available from: <https://arxiv.org/abs/2307.09288>.
 25. Beach J. Overview of child care wages 2000-2010 [Internet]. Ottawa: Child Care Human Resources Sector Council; 2013 [cited 2025 Sep 3]. Available from: <https://www.ccscc-cssge.ca/sites/default/files/uploads/Projects-Pubs-Docs/EN%20Pub%20Chart/Final%20Wages%20Paper.pdf>.
 26. Alto V. Modern generative AI with ChatGPT and OpenAI models: leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT-3 and GPT-4. Birmingham: Packt Publishing; 2023.
 27. Mahboubi P, Zhang T. Harnessing immigrant talent: reducing overqualification and strengthening the immigration system [Internet]. Toronto: C.D. Howe Institute; 2024 Dec [cited 2025 Sep 3]. <https://doi.org/10.2139/ssrn.5106385>
 28. Canada. Statistics Act [Internet]. Ottawa: Government of Canada; 1985 [cited 2025 Sep 9]. Available from: <https://laws-lois.justice.gc.ca/eng/acts/s-19/fulltext.html>.

