Using Large Language Models to measure and classify occupations in surveys*

Patrick Sturgis[1], Tom Robinson[1], Laura Fung[1], and Caroline Roberts[2]

[1]Department of Methodology, London School of Economics
[2]Department of Sociology, University of Lausanne

Abstract

We present the results of a new approach to measuring the occupations of respondents in surveys using Large Language Models (LLM). Occupation is a notoriously difficult variable to measure accurately due to the very large number of occupations and the technical ways they are described in standard classifications. These features of occupational classification systems mean that respondents cannot feasibly pick their occupation from a list, even with dynamic text prediction. The measurement and classification stages are therefore usually not conducted simultaneously, with coding of open responses about job title and tasks implemented in a subsequent stage of 'office coding'. In our new approach, an LLM integrated in the questionnaire scripting is used to code the job title response to the occupational classification within the interview. Where the job title does not contain sufficient information to be coded with confidence, the LLM probes for further relevant detail on job tasks, industry, qualifications, and so on. The approach has the potential to reduce respondent burden, lower costs, and yield more timely and accurate data. We evaluate the methodology by comparing the LLM-coded data to codes applied by human coders in a field experiment using the Verian Public Voice online probability panel.

1 Introduction

Accurate classification of occupations is critical to a wide range of theoretical and policy-focused research in the social sciences, including, *inter alia*, the study of socio-economic stratification and social mobility (Erikson & Goldthorpe, 2010), labour market polarisation (Acemoglu & Autor, 2011; Goos & Manning, 2007), and sex inequality (Jacobs,

*Paper presented at the European Survey Research Association (ESRA) conference, Utrecht, 30 June-4 July 2025

1989). Occupation is the fundamental building block for widely used measures of social class such as the UK National Statistics Socio-Economic Classification (Rose, 2003) and related measures of socio-economic status (Chan, 2004). However, the measurement of occupation in surveys is a notoriously challenging task (Elias, 1997). Because there are a very large number of occupations in a modern economy and the technical ways they are often described in classification systems, it is not effective to use a measurement approach which relies on respondents selecting from a pre-determined list, whether fixed or dynamic. For this reason, occupations are typically measured through open-ended questions, commonly asking respondents to state their job title, describe their main duties and tasks, and sometimes to provide additional context such as the type of organisation or industry in which they work (United Nations Department of Economic and Social Affairs, 2025). These responses must then be mapped to detailed classification schemes which contain occupational categories organised hierarchically from broad major groups down to highly specific and very numerous unit groups. Achieving accurate and consistent coding of these open responses is challenging because respondents often provide brief, ambiguous, or incomplete descriptions of their job tasks, leaving considerable scope for subjective interpretation and, therefore, coding errors. This problem is particularly acute for self-completion surveys where there is no interviewer to motivate the respondent to provide sufficiently detailed and relevant answers (Kochar et al., 2025)

Given these factors and the low rates of inter-rater reliability found in survey coding tasks generally (Kalton & Stowell, 1979), it is no surprise that studies have consistently found substantial inter-rater variability in occupation coding. Inter-coder agreement rates typically range from around 60% to 80% at detailed levels of classification, increasing as codes are aggregated to broader categories (Belloni et al., 2016; Conrad et al., 2016; Elias, 1997). Reliability is consistently found to be higher when responses contain clear, specific job titles with unambiguous task descriptions, whereas vague or abstract descriptions, or those containing general terms such as "administrator" or "services," are associated with greater coder disagreement and higher rates of referrals for additional information (Conrad et al., 2016; Elias, 1997). Notably, increasing the length or detail of responses does not always improve reliability; indeed, longer descriptions may introduce additional ambiguity or conflicting information, paradoxically reducing agreement among coders except in cases involving inherently complex or unfamiliar occupations (Belloni et al., 2016; Conrad et al., 2016).

Coder characteristics, particularly experience and training, are also important, expert coders generally achieve higher agreement than novices, and ongoing feedback or quality improvement systems can further enhance reliability (Elias, 1997). However, even among experts, subjective interpretation and the use of informal coding rules can produce systematic differences, especially in borderline cases or when multiple codes might plausibly apply (Conrad et al., 2016; Elias, 1997). Sparse or ambiguous responses can also result in coders being unable to apply an occupation code at all. The prevalence of such unclassifi-

able or missing occupation data is generally low in large-scale surveys but the proportion of cases that require referral or cannot be coded at all can reach 10–20% in some contexts (Conrad et al., 2016; Schierholz et al., 2018). Both respondent and job characteristics systematically influence coding reliability, with higher education, self-employment, foreign birth, and certain occupational groups being associated with increased coding error (Belloni et al., 2016; Peycheva et al., 2021).

To address the limitations and high cost of human coding, researchers have developed automated and semi-automated tools to assist in the coding process. Following Kochar et al. (2025), these approaches can be broadly grouped into three categories. First, semi-automated tools for post-survey coding, such as CASCOT (Elias et al., 2014), use predictive models to suggest occupation codes based on textual similarity and keyword matching. These rule-based tools advanced the field by introducing certainty scores and semi-automatic workflows, enabling efficient triage of cases for human review, augmented with ancillary variables (such as industry or education) to boost coding specificity and reduce manual workload (Belloni et al., 2016). However, rule-based and dictionary-driven methods are constrained by their dependence on the quality and coverage of the underlying dictionaries, often struggling with ambiguous or novel job descriptions.

Second, entirely closed-question approaches offer respondents fixed lists of occupations to choose from directly, removing the ambiguity inherent in open-ended responses. However, these methods frequently encounter usability challenges and significant respondent burden due to the difficulty respondents face in interpreting and selecting from extensive occupational lists (Tijdens, 2015). For this reason, they are mostly used when occupations are aggregated to higher level groupings, although this raises the challenge of respondents understanding the labels of the aggregated occupation groups and where their job sits within them (Kochar et al., 2025). Additionally, the rigidity of closed-question approaches often results in reduced specificity and accuracy of occupational data.

Third, some approaches use algorithms that present respondents with candidate occupation codes derived from their initial open-text answers, allowing respondents themselves to select the most appropriate code from a shortlist (Gweon et al., 2017; Peycheva et al., 2021). Schierholz et al. (2018), for example, implemented a supervised learning algorithm within interviewer-administered surveys, providing immediate occupation code suggestions that respondents could verify. Although this approach reduces coding burden, like the use of closed-questions, it relies heavily on the accuracy of the shortlisting algorithm and on respondents accurately identifying their correct occupational category from the suggestions provided, which can be challenging given the size and technical complexity of most occupational classifications. The result is that many responses still require office-coding as well as low rates of coder reliability (Schierholz et al., 2018)

More recent research has turned to fully automated approaches using machine learning and large language models (LLMs). Schierholz and Schonlau (2021), for example, conducted a benchmark comparison of seven occupation coding algorithms, showing that

supervised learning yields only modest accuracy gains over dictionary-based coding, with results highly sensitive to dataset variation and constrained by the quality of training data. Safikhani et al. (2023) used transformer-based models (BERT and GPT-3) to improve coding accuracy via hierarchical fine-tuning and digit-level prediction, achieving significantly higher performance than earlier methods. However, reliability at more detailed classification level was quite poor using this approach, with BERT achieving agreement rates of only 68%, and GPT-3 even lower at 57% for four digit unit-groups (Safikhani et al., 2023). Of particular relevance to our concerns here, both of these studies used the transformer model for coding only, rather than also employing it to improve the quality of responses obtained from respondents.

In this paper, we build on these earlier automation efforts by further integrating LLMs into the measurement of occupation. The key innovation is that, in addition to coding the survey responses to the occupational classification, our tool uses the LLM to dynamically generate tailored follow-up questions when initial coding confidence is low. These tailored follow-up questions specifically address the ambiguities or omissions in respondents' initial answers, closely replicating interviewer-style probing. By adapting interactions in real-time to clarify incomplete or ambiguous responses and integrating the coding of responses within the interview, the method has the potential to improve coding accuracy, reduce respondent burden, and lower coding costs compared to existing automated and semi-automated approaches.

2 Methodology

Overview Figure 1 presents a schematic of our pipeline for using LLMs to measure and classify occupations of survey respondents. Broadly, the pipeline comprises two interrelated components, a classifier and dynamic probing of survey responses. For the classifier, we build on work by the Office for National Statistics which uses retrieval-augmented generation (RAG), a natural language processing strategy where a generative model is presented with a pre-calculated shortlist of options from which to choose (Lewis et al., 2020). This RAG system works in three stages. First, we create a static, numerical representation of *all* SOC codes which we can then compare against respondent's own provided information. Each code is represented as a vector–called an embedding–and stored in a database that can be queried. Second, from the preliminary information collected from a subject (typically their job title), we "retrieve" from this database a list of SOC codes that are most similar to how the respondent described their occupation. Finally, we send a set of instructions to an LLM model (the prompt), asking it to either choose the correct occupation *from this shortlist* or generate a followup question that would help it to identify the correct code in a subsequent step. We discuss these options in more detail below.

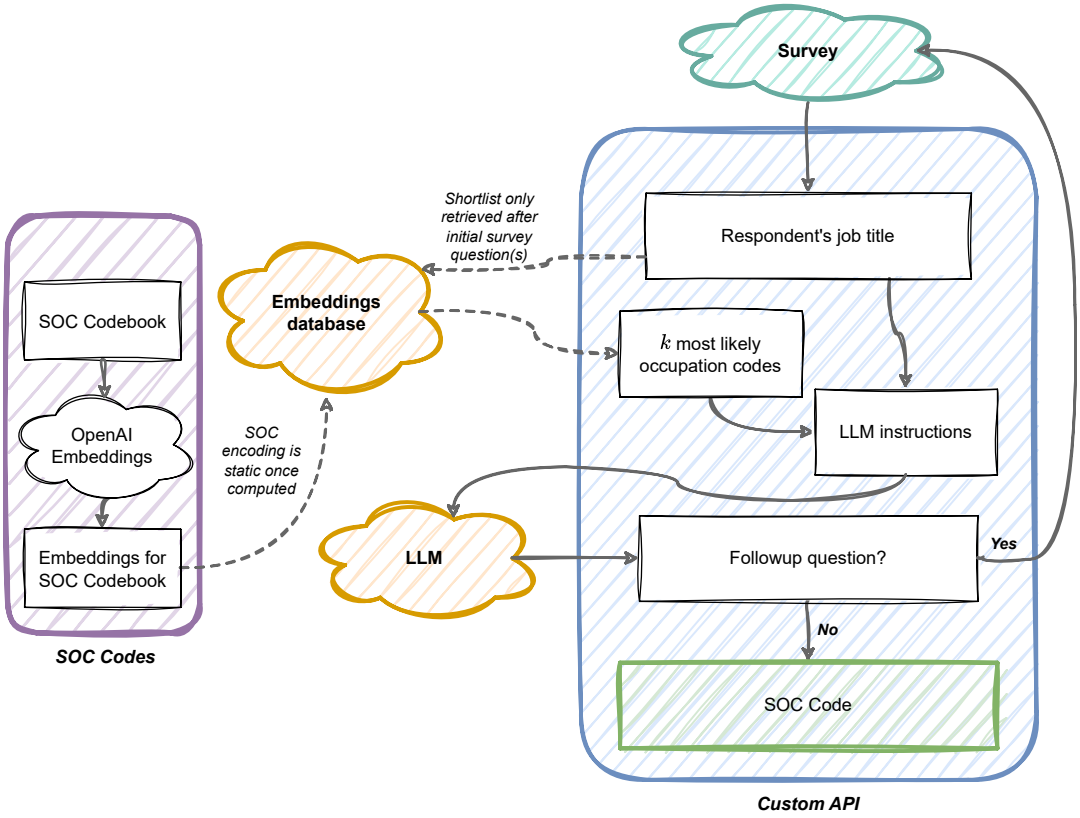

Figure 1: Schematic of occupation classification using a retrieval-augmented LLM pipeline

Shortlisting SOC codes The retrieval step serves to narrow the focus of the LLM at the point it makes a decision over a classification or followup question. This step has several advantages: it limits the amount of information we have to send the LLM, thus reducing the economic costs of integrating LLMs in survey research[1]; shorter prompts also yield quicker response times from the LLM and thus make the integration more seamless for respondents; and, substantively, it limits the extent to which LLMs can go "off topic", hallucinate, or focus on extraneous detail in the SOC codebook but which is irrelevant given the subject's pre-provided information.

To generate this shortlist, we use an embeddings-based approach. For every SOC code (and description) provided by the Office for National Statistics, we use OpenAI's pre-trained embeddings model to represent that entry as a 1024-length vector of numbers. This set of vectors is calculated once and then stored in an online database. Embeddings models are trained such that words or sentences that are more similar conceptually should have vectors that are also more similar.[2] Therefore, once a respondent has provided their job title, we can embed that title into the same 1024 space, and find the $k = 50$ closest

[1]The effective cost of sending the full SOC list to an OpenAI o4 model is XXX (correct as of XX June 2025).

[2]A canonical example is to think of the terms "King", "Man", and "Woman". Suppose $\overrightarrow{\text{King}}$ is the vector representation (i.e. word embedding) of "King". If the embeddings model is well-trained, then calculating $\overrightarrow{\text{King}} - \overrightarrow{\text{Man}} + \overrightarrow{\text{Woman}}$ should yield a word embedding vector very similar to $\overrightarrow{\text{Queen}}$.

vectors by calculating the cosine similarity between the job title embedding and every SOC code embedding in the database.[3]

Classification In the simplest version of our method, given the information provided by the subject (and a retrieval step to shortlist potential SOC codes), we prompt the LLM to choose the most likely code from this shortlist. In our testing, we found that guiding the reasoning of the LLM in a set of steps helped improve both the accuracy and reliability of the codings. We therefore ask the LLM to:

1. Identify a shortlist of three codes from the 50 provided codes that could be correct
2. Identify whether or not the information provided is adequate to choose amongst those three codes
3. Pick one of those codes that they think is most likely to be correct (regardless of whether they need more information)
4. Come up with an explanation for why they chose that code.

We also provide the model with a summary of the hierarchy of SOC codes and five examples of cases where similar sounding titles have different SOC codes, alongside the reasons for their differences. A full version of this prompt is available in Appendix Section XX.

Dynamic followups To integrate our model directly into surveys, allowing the LLM not only to classify respondent's occupations but ask questions in order to improve the accuracy of these classifications, we build a feedback system into our query where, if the LLM deems they do not have enough information, instead of returning a SOC code they return a question that can be fed directly into the survey itself. In turn, the respondent's answer to this question is fed back into the information sent to the LLM (along with all previous responses). This process can be repeated a finite number of times (set by the surveyor/researcher), or until the model returns a SOC code.[4]

In our early testing, we found that LLMs often erred on the side of caution, asking followup questions where we might expect a human coder to be able to decide on the SOC code. For example, it was quite common for the model to ask what subjects university professors taught, even though all university professors would be classified under the same SOC code (2311). In an effort to limit this behaviour, our prompt includes prescriptive information on what types of questions the LLM can ask, focused on *"the industry of the organization the subject works for; the sorts of tasks the respondents performs in their*

[3]The cosine similarity can be estimated as $\frac{\sum_{i=1}^{1024} \overrightarrow{A_i} \times \overrightarrow{B_i}}{\sum_{i=1}^{1024} \overrightarrow{A_i^2} \times \sum_{i=1}^{1024} \overrightarrow{B_i^2}}$, where $\overrightarrow{A_i}$ stands for the ith element of the embedding vector representing concept A.

[4]We prompt the LLM to return different special characters based on the type of response it provides (i.e. classification or followup), allowing the survey flow to route questions automatically.

role; if the respondent's job requires any specific qualifications; whether the respondent has any supervisory or managerial responsibilities". We also allow the model to ask followup questions where the respondent's answer to a previous question was unclear. As in the classification-only version, we provide the same information on the SOC schema and examples of differences between similar-sounding job titles.

LLM instances and balancing latency in dynamic surveys For classification only workflows, which can be performed "offline" (i.e. not while the survey is in progress), our strategy has been to use more advanced reasoning models that have been shown to have considerable advantages in providing reliable and accurate classifications [TOM TO FIND CITES]. These models, however, have the downside of being slow – it may take anywhere from 5 to 30 seconds for the model to process the prompt and return a SOC code. Given that we can parallelise this step (i.e. query each subject's code at the same time), there are negligible costs to this slower reasoning.

However, for dynamic use where our method is returning followup questions directly to the respondent, these sorts of delays could lead to overly high attrition as the survey appears to stall. Hence, in the dynamic use-case, we have balanced the predictive fidelity of the model with its latency–the time it takes to return a fully generated message. At the time of writing, our experiments suggested that an OpenAI **...** model struck the best balance, and led to latencies of around 2-3 seconds on average – although not instant, this was in the realm of what we suspected a survey respondent might consider normal loading times.

3 Results

TBC

4 Discussion

TBC

References

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of labor economics* (pp. 1043–1171, Vol. 4). Elsevier. https://doi.org/10.1016/S0169-7218(11)02410-5

Belloni, M., Brugiavini, A., Meschi, E., & Tijdens, K. (2016). Measuring and detecting errors in occupational coding: An analysis of SHARE data. *Journal of Official Statistics*, *32*(4), 917–945. https://doi.org/10.1515/jos-2016-0049

Chan, T. W. (2004). Is there a status order in contemporary british society?: Evidence from the occupational structure of friendship. *European Sociological Review*, *20*(5), 383–401. https://doi.org/10.1093/esr/jch033

Conrad, F. G., Couper, M. P., & Sakshaug, J. W. (2016). Classifying open-ended reports: Factors affecting the reliability of occupation codes. *Journal of Official Statistics*, *32*(1), 75–92. https://doi.org/10.1515/jos-2016-0003

Elias, P. (1997, January 1). *Occupational classification (ISCO-88): Concepts, methods, reliability, validity and cross-national comparability* (OECD Labour Market and Social Policy Occasional Papers No. 20) (Series: OECD Labour Market and Social Policy Occasional Papers Volume: 20). https://doi.org/10.1787/304441717388

Elias, P., Birch, M., & Ellison, R. (2014). *CASCOT international version 5, user guide*. Institute for Employment Research, University of Warwick.

Erikson, R., & Goldthorpe, J. H. (2010). Has social mobility in britain decreased? reconciling divergent findings on income and class mobility: Has social mobility in britain decreased? *The British Journal of Sociology*, *61*(2), 211–230. https://doi.org/10.1111/j.1468-4446.2010.01310.x

Goos, M., & Manning, A. (2007). Lousy and lovely jobs: The rising polarization of work in britain. *Review of Economics and Statistics*, *89*(1), 118–133. https://doi.org/10.1162/rest.89.1.118

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three methods for occupation coding based on statistical learning [Publisher: SAGE Publications]. *Journal of Official Statistics*, *33*(1), 101–122. https://doi.org/10.1515/jos-2017-0006

Jacobs, J. A. (1989). Long-term trends in occupational segregation by sex. *American Journal of Sociology*, *95*(1), 160–173. https://doi.org/10.1086/229217

Kalton, G., & Stowell, R. (1979). A study of coder variability [Publisher: Wiley for the Royal Statistical Society]. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(3), pp. 276–289. http://www.jstor.org/stable/2347199

Kochar, S., Brown, M., & Calderwood, L. (2025). *Occupation coding in selfcompletion surveys: Evidence review*. Survey Futures.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 9459–9474, Vol. 33). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Peycheva, D. N., Sakshaug, J. W., & Calderwood, L. (2021). Occupation coding during the interview in a web-first sequential mixed-mode survey [Publisher: SAGE Publica-

tions]. *Journal of Official Statistics*, *37*(4), 981–1007. https://doi.org/10.2478/jos-2021-0042

Rose, D. (Ed.). (2003). *A researcher's guide to the national statistics socio-economic classification*. SAGE.

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*(1), 6. https://doi.org/10.1007/s44163-023-00050-y

Schierholz, M., Gensicke, M., Tschersich, N., & Kreuter, F. (2018). Occupation coding during the interview. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181*(2), 379–407. https://doi.org/10.1111/rssa.12297

Schierholz, M., & Schonlau, M. (2021). Machine learning for occupation coding—a comparison study. *Journal of Survey Statistics and Methodology*, *9*(5), 1013–1034. https://doi.org/10.1093/jssam/smaa023

Tijdens, K. (2015). Self-identification of occupation in web surveys: Requirements for search trees and look-up tables. *Survey Methods: Insights from the Field (SMIF)*. https://doi.org/10.13094/SMIF-2015-00008

United Nations Department of Economic and Social Affairs. (2025, February 12). *Energy statistics pocketbook 2025*. United Nations. https://doi.org/10.18356/9789211069280

5 Appendix