Working paper ZMB-25058 June 2025

Transforming Zambia's Labour Force Survey using artificial intelligence

Tyler Rossow
Tobias Edison
Rory Hardie
Harry Mayne

IGC

DIRECTED BY LSE UNIVERSITY OF OXFORD

FUNDED BY UKaid from the British people

Transforming Zambia's Labour Force Survey Using Artificial Intelligence

Tyler Rossow, Tobias Edison, Rory Hardie, and Harry Mayne

June 2025

Abstract

We document large language models (LLMs) as a potential tool to improve the classification of responses to Zambia's Labour Force Survey, a household-based sample survey on the country's labour force characteristics. In collaboration with the Zambia Statistics Agency, we assess the ability of GPT-4 Turbo to classify descriptive survey responses into occupational and industry codes. The dataset contains 1,059 observations from the 2023 Labour Force Survey. Respondents' verbal descriptions of their occupation and industry are used to assign four-digit codes from the International Standard Classification of Occupations (ISCO) and the International Standard Industrial Classification (ISIC). Our results indicate that, for most digits, GPT-4 Turbo outperforms survey enumerators by a statistically significant ($p < 0.01$) margin. With a typical Labour Force Survey of 10,400 households, our estimates suggest that the Zambia Statistics Agency could save up to 130 working days annually whilst improving classification accuracy by over 8 percentage points.

Contents

1 Introduction

Zambia's Labour Force Survey (LFS) is a household-based sample survey administered by the Zambia Statistics Agency (ZamStats) in partnership with the Ministry of Labour and Social Security (MLSS). The primary objective of the LFS is to determine the size of the labour force and analyse patterns across age, gender, industry, sector of employment, and education. Since 2017, the quarterly LFS has been the basis for nationally representative statistics on Zambia's labour market, underpinning policies on unemployment, job creation, and informality.

The survey involves enumerators, hired individuals conducting the survey on the ground, using electronic questionnaires to collect information during face-to-face interviews with respondents. The core component of the LFS questionnaire is the 'Characteristics of the Main Job' module, which seeks to identify an individual's occupation and industry. The key questions asked are:

- Occupation: *'In his/her main job/business, what kind of work does (NAME) usually do? (Write occupation title, if any, and main duties and tasks)'*
- Industry: *'In (NAME) workplace what kind of business/activity is mainly carried out? (Write name of establishment, if any, and main activity, goods, or services)'*

The LFS enumerators use the descriptions provided by the respondents to assign four-digit codes for the International Standard Classification of Occupations (ISCO) and the International Standard Industrial Classification (ISIC). The primary difference between the two codes is that ISCO codes classify *jobs* based on the duties required, while ISIC codes classify *industries* based on the economic activities performed. In both cases, the level of granularity increases with each digit, i.e. the first digit provides a more general description than all four digits together. Table 1 explains each digit and its level of detail for the ISCO-08 and ISIC Rev. 4, the editions used by ZamStats. An example from the dataset – a subsistence crop farmer (ISCO occupation) growing cereals (ISIC industry) – is included for illustration. (Note that in this case the third and fourth digits for ISCO-08 capture identical information, but this is not always the case).

Table 1: ISCO and ISIC Codes Explained

LFS Descriptions	Number of Digits	ISCO-08	ISIC Rev. 4
Occupational Title: Subsistence Crop Farmer	One	Major Group *(e.g., 6 – Skilled Agricultural, Forestry, and Fishery Workers)*	N/A
Main Tasks and Duties: Subsistence Crop Farmer Growing of Maize	Two	Sub-major Group *(e.g., 63 – Subsistence Farmers, Fishers, Hunters, and Gatherers)*	Division (*e.g., 01 – Crop and Animal Production, Hunting, and Related Service Activities)*
Main Activity, Goods, or Services: Subsistence Crop Farmer Growing Crops for Sale	Three	Minor Group *(e.g., 631 – Subsistence Crop Farmers)*	Group *(e.g., 011 – Growing of Non-Perennial Crops)*
	Four	Unit Group *(e.g., 6310 – Subsistence Crop Farmers)*	Class *(e.g., 0111 – Growing of Cereals (Except Rice), Leguminous Crops, and Oil Seeds)*

2 Policy Demand

The process of converting a description of an occupation or industry into the ISCO and ISIC codes is not straightforward. Challenges arise due to the complexity of the classification systems and the time constraints faced by enumerators assigning codes after interviews. ZamStats officials overseeing the LFS raised these concerns during discussions with the Zambia Evidence Lab (ZEL). To assess the survey process, we spent one week monitoring the survey process across two enumeration areas (EAs). During monitoring and subsequent discussions with the LFS team, two issues stood out:

- **Complex Classifications**— Each digit of the ISCO and ISIC codes represents an added level of specificity. Accurate coding requires enumerators to spend significant time searching the relevant codebook on their tablets. With PDF copies of the ISCO-08 and ISIC Rev. 4 codebooks spanning 433 and 306 pages, respectively, enumerators are unlikely to recall each section of the codebook nor have time to parse the entire document.
- **Incomplete Descriptions**— When respondents provide short or incomplete descriptions of their occupation or industry, the enumerators often lack sufficient detail to assign the four-digit ISCO and ISIC codes accurately. For instance, the example in Table 1 highlights that enumerators will often report the same information for the title, description, and main activities recorded in response to the bulleted questions in the Context section.

These challenges place considerable pressure on enumerators to complete ISIC and ISCO codes accurately and on time. As an alternative, we proposed using large language models (LLMs) – artificial intelligence (AI) trained to answer natural language questions – to classify the responses. Given that the classification of responses is a time-intensive and highly standardized task, LLMs hold the potential to both save time and improve accuracy.

3 Methodology

3.1 Dataset

We use a random sample of 1,059 observations from the 2023 LFS, which consists of 10,400 households. The dataset includes the occupation and industry descriptions, the original ISCO and ISIC codes from the enumerators, and codes generated by expert officials from the ZamStats LFS team. These expert officials, who oversee the survey process for the entire country, have greater training and experience compared to enumerators and were not under the time pressure that enumerators face. The ZamStats codes serve as the ground truth against which enumerator and LLM classifications are compared.

While all 1,059 observations are classified by the LLM, the sample size for comparison is 1,002. Because there are 57 duplicate observations, where the recorded responses are identical, it is not possible to identify the comparable classifications for these responses between the ground truth, enumerator, and LLM datasets.

3.2 Task

The LLM is prompted to assign ISCO and ISIC codes for each observation. In the dataset, there are three columns where respondents described their job title, job description, and the main activities they conduct. Using the example in Table 1, the entries would be:

- Occupational Title (D1_TITLE): Subsistence Crop Farmer
- Main Tasks and Duties (D1_DESC): Subsistence Crop Farmer Growing of Maize
- Main Activity, Goods, or Services (D1_MAIN_ACTIVITIES): Subsistence Crop Farmer Growing Crops for Sale

3.3 Implementation

We evaluate GPT-4 Turbo (OpenAI, 2024) using OpenAI's Application Programming Interface (API). This model is provided with a prompt instructing it to classify batches of twenty dataset instances at a time. We find that twenty instances per prompt helps manage rate limits and improve runtime stability. The model is instructed to return a structured output of the twenty predicted classes, which we process in Python. We do this across all 1,059 instances in the dataset. In this context, the model is not given the book of potential classes and must rely on its preexisting knowledge of ISCO and ISIC codes.

The prompt, available in the appendix, instructs the model to focus on accuracy and consistency. Should the LLM fail to provide an answer, which we observe in some cases, the script allows up to ten retries. We select GPT-4 Turbo because it offers strong performance while remaining relatively inexpensive compared to newer models. This makes it a practical long-term option for countries with limited government budgets.

We anticipate that performance could be substantially higher using newer, state-of-the-art LLMs, albeit at a much greater cost. In addition, we do not explore more advantageous prompting strategies such as few-shot prompting, where the LLM is shown a small number of correct classifications within the prompt. In our case, the model is never provided with examples of successful completions, meaning that it receives no information besides the dataset and the one-paragraph prompt. Nonetheless, our initial results provide strong reasons to be optimistic about the capabilities of LLMs to classify survey responses.

3.4 Scoring

The problem of evaluating the capabilities of LLMs has received significant attention in the academic literature (Pangakis and Wolken, 2024; Shankar et al., 2024). A core element of this challenge is how to score the LLM-generated classifications against the ground truth. While human oversight can be used, it is typically costly and can be biased, concerns that are alleviated but not resolved by expert, trained judgement (Biderman et al., 2024).

In this case, we use exact match to verify the accuracy of the LLM's response. For each response, we test if the generated code perfectly matches the ground truth answer. To assess performance on individual digits, we use cumulative accuracy: the model must correctly predict digit n before it can receive credit for digit $n+1$. This allows for partial correctness but ensures that higher-digit accuracy depends on accuracy in preceding digits.

By using exact match, we avoid introducing human judgment into the scoring process; classifications either match the ground truth or they do not. However, this approach depends on the availability of high-quality ground truth labels, as discussed in Assumption 2.1 below.

3.5 Assumptions

For our methodology to assess accuracy on a nationally representative sample of Zambia's labour force, we require two assumptions:

Assumption 1.1: The survey responses are a nationally representative sample of Zambia's labour force.

The assumption of national representativeness embeds two assumptions necessary for our methodology to be internally valid to the LFS and externally valid to other surveys.

Assumption 1.1a: The 1,059 ground truth examples are a random sample from the full LFS dataset.

From the 10,400 responses in the 2023 LFS, the ZamStats team randomly sampled from the responses of individual cases for those who were categorised as employed. Random sampling is a necessary assumption to test if LLMs can accurately classify LFS surveys. If the sample is not random, the model's performance will not reflect performance in the wider LFS survey.

Assumption 1.1b: The full LFS dataset is nationally representative.

Further details on the sample process are available from the LFS survey results (ZamStats, 2024). ZamStats uses a Split-Panel Design that selects a random sample of 520 enumeration areas (EAs) each quarter while ensuring that each EA is surveyed once annually. ZamStats ensures representativeness at national and sub-national levels using sampling weights for each EA.

Assumption 2.1: ZamStats expert officials provide a baseline against which enumerators and LLMs can be compared.

Given that the LFS team classifying responses oversees the survey for the entire country, it is reasonable to assume that their codes represent the highest quality classifications available in a Zambian context.

4 Results

Our results suggest that LLMs outperform human enumerators whilst significantly reducing the time required for classification.

4.1 ISCO Classification

The ISCO results are displayed in Figure 1. With statistically significant results ($p < 0.01$) bolded, we find:

- **80.2%** accuracy between the LLM and ground truth on the first digit and **73.5%** accuracy between enumerators and ground truth on the first digit.
- 65.0% accuracy between the LLM and ground truth on the first two digits and 63.9% accuracy between enumerators and ground truth on the first two digits.
- **41.6%** accuracy between the LLM and ground truth on the first three digits and **53.6%** accuracy between enumerators and ground truth on the first three digits.
- **27.1%** accuracy between the LLM and ground truth on all four digits and **48.5%** accuracy between enumerators and ground truth on all four digits.

It is unsurprising that the LLM performs better on the first digits but weaker on the third and fourth digits. LLMs will generally be less prone to human error on the more basic first and second digits, where the consistency to avoid simple mistakes is crucial. On the more specific final digits, which require greater country and interview context, LLMs are likelier to struggle.

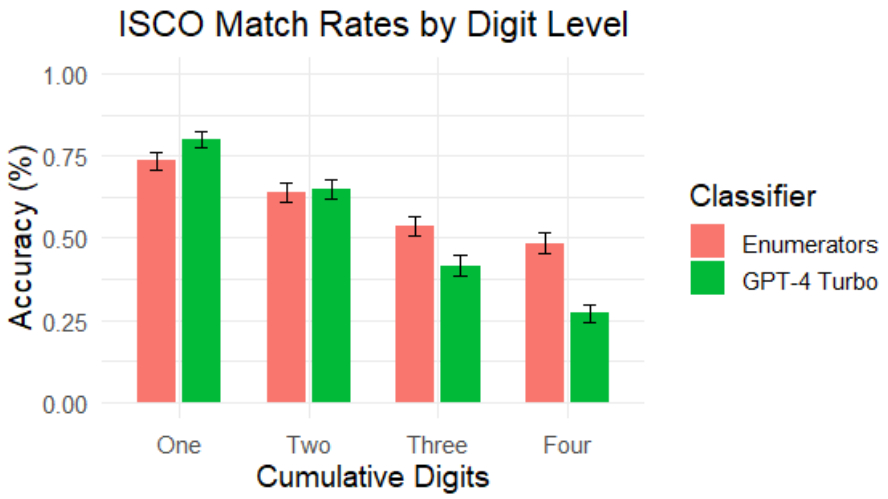
An example of the LLM avoiding simple errors can be seen in its approach to the example from Table 1. Recall that this respondent was listed as the following[1]:

- Occupational Title: Subsistence Crop Farmer
- Main Tasks and Duties: Subsistence Crop Farmer Growing of Maize
- Main Activity, Goods, or Services: Subsistence Crop Farmer Growing Crops for Sale

[1]A more comprehensive list of examples, including this one, can be found in Appendix Table 1. In this table, the first four rows document cases where the LLM matched the master, but the enumerators did not. The fifth row contains an example where all three groups differ. Finally, the sixth row showcases where the enumerators match the master codes, but the LLM does not.

In this case, the enumerator makes a simple mistake, classifying the ISCO codes as 0110, Armed Forces Occupations. Because 0110 is the first code on a long list of occupations, time-constrained enumerators will occasionally input unrelated occupations as armed forces to avoid perusing the entire list. The ground truth and LLM assign 6310 and 6221, respectively, both identifying the Industry Major Group as Skilled Agricultural, Forestry and Fishery Workers.

Figure 1



This chart compares the accuracy of the enumerators and GPT-4 Turbo against ground truth across all ISCO digits. Digits are cumulative, meaning that accuracy on digit *n* is a prerequisite for accuracy on digit *n+1*. For this reason, accuracy declines as cumulative digits increase with both methods. All error bars are presented at the 95% confidence level.

4.2 ISIC Classification

The ISIC results are displayed in Figure 2. Here, the LLM outperforms the enumerators **across all categories**[2]. With statistically significant results ($p < 0.05$) bolded, we find:

- **85.1%** accuracy between the LLM and ground truth on the first two digits and **77.0%** accuracy between enumerators and ground truth on the first two digits.
- **66.7%** accuracy between the LLM and ground truth on the first three digits and **58.3%** accuracy between enumerators and ground truth on the first three digits.
- **54.1%** accuracy between the LLM and ground truth on all four digits and **47.1%** accuracy between enumerators and ground truth on all four digits.

A successful example of the LLM's ISIC classification abilities comes from the first example in Appendix Table 2[3]. The respondent describes their work as:

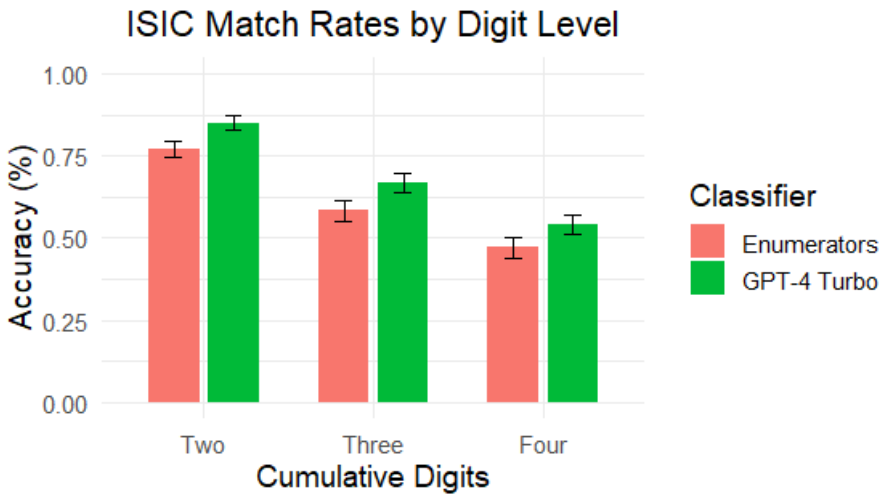
- Occupational Title: Beer Brewer
- Main Tasks and Duties: Brewing Local Beer
- Main Activity, Goods, or Services: Brewing Local Beer

The LLM assigns 1103 and the ground truth is 1101. Both codes identify the industry section as Manufacturing, and the four-digit codes both indicate a manufacturer of alcohol. On the other hand, the enumerator assigns code 9609, Other Service Activities. The four-digit code 9609 describes personal service activities not illustrated elsewhere in the codebook, such as Turkish baths, shoe shining, and photo booths. Because the enumerator faces time constraints, he or she may not have realised from the codebook that a separate section for manufacturers of alcohol existed. The LLM, which faces no time constraint, matches the ground truth to the first three digits.

[2] The first digit is omitted as it communicates information on a respondent's industry only in conjunction with the second digit.

[3] A more comprehensive list of examples, including this one, can be found in Appendix Table 2. In this table, the first four rows document cases where the LLM matched the master, but the enumerators did not. The fifth row contains an example where all three groups differ. Finally, the sixth row showcases where the enumerators match the master codes, but the LLM does not.

Figure 2



This chart compares the accuracy of the enumerators and GPT-4 Turbo against ground truth across relevant ISIC digits. The first digit is omitted as it communicates information on a respondent's industry only in conjunction with the second digit. Digits are cumulative, meaning that accuracy on digit n is a prerequisite for accuracy on digit *n+1*. For this reason, accuracy declines as cumulative digits increase with both methods. All error bars are presented at the 95% confidence level.

4.3 Test of Statistical Significance

Following these results, we test the differences for statistical significance using a two-proportion Z-test. This is consistent with the methodology suggested in Miller (2024), which advises practitioners to test language model evaluations for statistical significance and report standard errors. The two-proportion Z-test compares the accuracy differences between the LLM and enumerators. For the test, we use the following hypotheses:

- Null Hypothesis (H_0): LLM – enumerators = 0
- Alternative Hypothesis (H_1): LLM – enumerators $\neq$ 0

Failure to reject the null hypothesis indicates that the difference in accuracy between the LLM and enumerators is not statistically significant and may be due to random variation.

The two-proportion Z-test relies on three assumptions, namely that:

Assumption 1.2: The population samples are random and independent from each other.

This assumption is comparable to Assumption 1.1 above, and the support for it relies on similar arguments.

Assumption 2.2: The data is categorical, e.g. possible results are pass/fail or yes/no.

This assumption holds, since the LLM and enumerator codes either match the ground truth, or they do not.

Assumption 3.2: Sample sizes are large enough to ensure sample proportions are normally distributed.

As a rule of thumb, $n\widehat{p} > 10$ and $n(1\text{-}\widehat{p}) > 10$, where n is the sample size and $\widehat{p}$ is the sample proportion.

With n = 1002, the sample size is large enough for this assumption to hold in all cases.

After performing the test, we find that the LLM outperforms humans with statistically significant differences on the first ISCO digit and all ISIC digits. This means that with at least 95% confidence, the variations in performance between enumerators and the LLM are not due to random chance. Table 2 reports the accuracy differences between the LLM and enumerators for each digit, as well as p-values. Because the first two digits of an ISIC code jointly determine an industry section, no first digit results are reported.

Table 2: Tests of Statistical Significance

Cumulative Digits	**LLM Less Enumerator Accuracy (ISCO)**	**LLM Less Enumerator Accuracy (ISIC)**
One	0.067***	N/A
	(0.000)	
Two	0.011	0.081***
	(0.607)	(0.000)
Three	-0.120***	0.084***
	(0.000)	(0.000)
Four	-0.214***	0.070**
	(0.000)	(0.002)

*p-values are in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$*

5 Implications

These results, while statistically significant, are only meaningful from a policy perspective insofar as they materially affect the survey process and results. From ZEL's consultations with ZamStats, there are three primary implications of great importance to policymakers.

5.1 Time Savings

Officials from the ZamStats LFS team estimate that each classification takes around one minute for staff in Lusaka, and up to ten minutes for enumerators. With 10,400 responses annually, this can become a significant burden that trades off with other job responsibilities. Using conservative estimates of one minute per code for the LFS staff in Lusaka and three minutes for enumerators, we estimate over 43 working days saved for the team in Lusaka and 130 working days saved for enumerators, assuming a typical 8-hour workday.

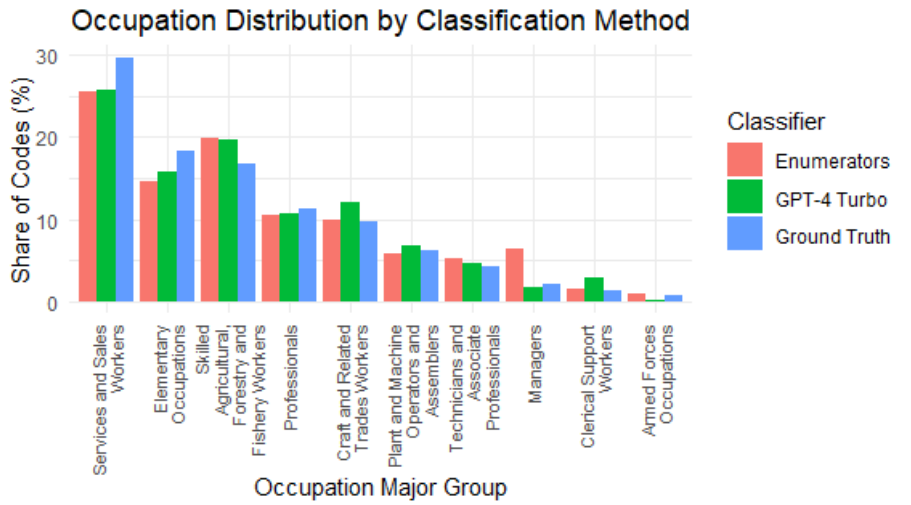
Furthermore, this has the potential to increase the policy relevance of the official statistics by reducing the time to dissemination. While all LFS data collection for 2023 was finished by December 2023, the annual report for the survey was not produced until November 2024. By freeing labour power to be used on other outputs, automating classifications has the potential to speed up the statistical production process. Given that governments often conduct fiscal and monetary stabilization policies in response to labour market trends, increasing the speed at which reports are produced can help policymakers identify and rapidly respond to labour market shocks.

5.2 Accuracy

Accuracy gains can be forecast by comparing enumerators against the LLM (see Table 2). On the first digit of ISCO, the accuracy gain is nearly 7 percentage points; on the first two digits of ISIC, it is over 8 percentage points. In a typical sample of 10,400 households with one employed person on average per household, LLM classification would offer policymakers 842 additional accurate first two-digit classifications on ISCO and 697 additional accurate first-digit classifications on ISIC.

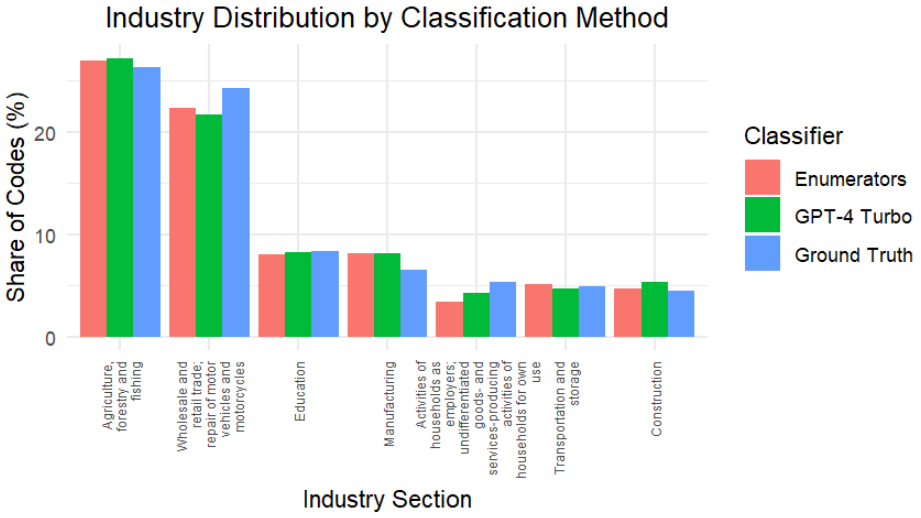
In larger samples, accuracy gains also have the potential to reshape our understanding of Zambia's labour market composition, which is crucial for any policy relating to sectoral targeting or structural transformation. Figures 3 and 4 below show how frequently the ground truth, enumerators, and LLM assign various occupation major codes from ISCO and industry sections from ISIC. The insight is that, in addition to improving accuracy, LLM classification will alter the *distribution* of labour market categories. With a larger sample, where there are many cases of each occupation major code and industry section, it would also be possible to obtain more precise estimates of how the labour composition estimates vary using different classification techniques.

Figure 3



This chart compares the distribution of occupation major groups (ISCO first digit), between the enumerators, GPT-4 Turbo, and ground truth. There are ten occupation major groups total. For each classifier, the share of codes will sum to 100%.

Figure 4



This chart compares the distribution of industry sections (ISIC first two digits), between the enumerators, GPT-4 Turbo, and ground truth. While there are twenty-one total industry sections, we limit the graph to the seven sections where the share of codes exceeds three percent for all three classifiers. As such, the share of codes will not sum to 100%.

5.3 Fresh Applications

After refining the LFS classification process, a similar approach could likely be used for the Census, the Living Conditions Monitoring Survey (LCMS), and the Demographic and Health Survey (DHS), which follow comparable classification methods. ZamStats officials have already expressed interest to ZEL in adopting the same process for the Census, highlighting the potential for transformative change.

6 Limitations

Our initial results, while encouraging, suffer from possible limitations that will be addressed as this working paper develops. Below, we outline four limitations as well as possible mitigation steps.

6.1 Implementation

Any proposed solution will need to be implemented by ZamStats as the official statistical producer in Zambia. As ZEL's collaboration with ZamStats continues, important decisions will need to be made about whether responses should be classified after the survey or as part of the survey process.

Enumerators would need to record responses in detail to provide the LLM with as much information as possible to classify responses after the survey process. From there, ZamStats could run the code in Python, possibly via a Graphical User Interface (GUI).

Alternatively, response classification could be built into the survey process. Currently, enumerators input the responses into a tablet and are expected to select occupational and industry codes from a long list of available choices. With automated classifications, the manual selection of a code could be replaced by an LLM tool that classifies responses on the tablet in real time.

A more advanced method, currently being explored by academic researchers at the Department of Methodology at the London School of Economics, involves dynamically generating survey questions using LLMs. Instead of relying on a fixed set of questions, the LLM would iteratively generate questions based on previous responses. For straightforward cases (e.g., "accountant"), the model may require only one or two questions. In more complex cases, it would continue generating targeted questions until it gathers enough information for an accurate classification. Given rural connectivity challenges in Zambia, this approach would also require consideration of internet access and whether an open-source offline LLM could be used.

6.2 Model Selection

While GPT-4 Turbo is relatively cheap, it is now an older legacy model (OpenAI, 2024). Newer models released after the time of the initial experiment, such as GPT-4.1 mini, may offer a superior blend of performance and cost efficiency (OpenAI, 2025). Other possibilities worth exploring include the free tier of Gemini models to minimise cost (Google, 2025), as well as frontier models such as o4-mini (OpenAI, 2025) and Claude Opus 4 (Anthropic, 2025) to optimise performance.

6.3 Prompting

GPT-4 Turbo possesses extensive text classification abilities, enabling it to outperform human enumerators in many cases (Kostina et al., 2025). However, our current methodology uses *zero-shot prompting*, where the LLM receives natural language instruction describing the task, but no examples of successful completions (Brown et al., 2020).

Alternatives to zero-shot prompting are one-shot or few-shot prompting, where an LLM is provided one or a few *instances* of successful completions of the task (Li et al., 2023). Research has shown that LLMs perform *in-context learning*, meaning that they perform increasingly better when more correct instances are included in the prompt (Brown et al., 2020).

LLMs also benefit from prompt engineering. Chain-of-thought prompting, where an LLM is instructed to "think step-by-step" when solving the task, has been shown to improve model performance in multiple reasoning contexts (Wei et al., 2022). In the LFS context, prompt engineering might also involve instructing the LLM to process each digit sequentially, rather than returning a final four-digit output, perhaps with examples of this procedure included. It may also be worth processing instances one by one, rather than in chunks of twenty, to improve runtime stability.

Lastly, research has shown that small changes in prompt formatting, such as casing, can result in large differences in model performance for semantically equivalent prompts (Sclar et al., 2024). It may be worth employing tools such as FormatSpread, which evaluate prompt formats, to determine the best prompt format to optimise performance (Sclar et al., 2024).

6.4 Small Dataset

With a dataset of 1,059 observations, a larger sample is needed to verify the results. While our current sample is large enough to obtain statistical significance, the number of instances for certain occupational and industry categories is small. By using a larger dataset, it is possible to test for statistical significance at the category level and identify areas where LLMs are weakest.

Furthermore, the smaller dataset makes few-shot learning more difficult as there is a limited set of successful completions that can be inputted alongside natural language instructions.

To increase data availability, we plan to explore using the LFS from other years, recently shared by ZamStats. While this will greatly increase the number of instances, addressing labelling problems will be paramount as only enumerator codes are available for this dataset. With several potential labelling techniques available (Allam et al., 2025), it will be essential to identify the optimal technique to avoid humans needing to manually classify thousands of responses.

7 Conclusion

In this working paper, we demonstrate that LLMs can outperform human enumerators in classifying responses to Zambia's Labour Force Survey. With 1,059 observations from the 2023 LFS, we compare enumerator and GPT-4 Turbo classifications against ground truth for ISCO and ISIC codes. The LLM outperforms human enumerators by nearly 7 percentage points on the most general ISCO group and over 8 percentage points on the most general ISIC group. With a typical full survey consisting of 10,400 households, the Zambia Statistics Agency could save up to 130 working days annually and bolster classification accuracy by over 800 responses. These improvements could accelerate the production of official statistics, improving macroeconomic stabilization policies. They should also enhance understanding of Zambia's labour force composition, supporting policies on labour market targeting and structural transformation.

Additionally, there is significant room for further improvements to our approach, including the small sample size, zero-shot prompting, model selection, and lack of current implementation steps. In each area, we outline possible steps to address the limitations, including additional LFS data, few-shot learning, frontier models, and LLM integration into ZamStats tablets. Once complete, this current project could serve as a basis for work on the Census, Living Conditions Monitoring Survey (LCMS), and Demographic and Health Survey (DHS), which use similar classification procedures. It might also inform projects in IGC countries such as Ethiopia, Rwanda, and Uganda, which conduct their own labour force surveys. We look forward to continuing to explore these and other avenues in the emerging field of using LLMs in the statistical production process.

8 References

Allam, H., Makubvure, L., Gyamfi, B., Graham, K.N., and Akinwolere, K. (2025). Text Classification: How Machine Learning Is Revolutionizing Text Categorization. *Information, 16(2).* https://doi.org/10.3390/info16020130

Anthropic. (2025). *Claude Opus 4.* https://www.anthropic.com/claude/opus

Biderman, S., Schoelkopf, H., Sutawika, L., Gao, L., Tow, J., Abbasi, B., Aji, A.F., Ammanamanchi, P.S., Black, S., Clive, J., DiPofi, A., Etxaniz, J., Fattori, B., Forde, J.Z., Foster, C., Hsu, J., Jaiswal, M., Lee, W.Y., Li, H., Lovering, C., Muennighoff, N., Pavlick, E., Phang, J., Skowron, A., Tan, S., Tang, X., Wang, K.A., Winata, G.I., Yvon, F., and Zou, A. (2024). Lessons from the Trenches on Reproducible Evaluation of Language Models. *arXiv.* https://doi.org/10.48550/arXiv.2405.14782

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv.* https://doi.org/10.48550/arXiv.2005.14165

Google. (2025). *Gemini models.* https://ai.google.dev/gemini-api/docs/models

International Labour Organization. (2012). *International Standard Classification of Occupations ISCO-08 Volume 1: Structure, group definitions and correspondence tables.* https://webapps.ilo.org/ilostat-files/ISCO/newdocs-08-2021/ISCO-08/ISCO-08%20EN%20Vol%201.pdf

Kostina, A., Dikaiakos, M., Stefanidis, D., and Pallis, G. (2025). Large Language Models For Text Classification: Case Study And Comprehensive Review. *arXiv.* https://doi.org/10.48550/arXiv.2501.08457

Li, Z., Zhu, H., Lu, Z., and Yin, M. (2023). Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. *arXiv.* https://doi.org/10.48550/arXiv.2310.07849

Miller, Evan. (2024). Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. *arXiv.* https://doi.org/10.48550/arXiv.2411.00640

OpenAI. (2024). *GPT-4 Turbo.* https://platform.openai.com/docs/models/gpt-4-turbo

OpenAI. (2025). *GPT-4.1 mini.* https://platform.openai.com/docs/models/gpt-4.1-mini

OpenAI. (2025). *o4-mini*. https://platform.openai.com/docs/models/o4-mini

Pangakis, N., and Wolken, S. (2024). Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels. *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science*, 113-131. https://doi.org/10.48550/arXiv.2406.17633

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2024). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *arXiv*. https://doi.org/10.48550/arXiv.2310.11324

Shankar, S., Zamfirescu-Pereira, J.D., Hartmann, B., Parameswaran, A., and Arawjo, I. (2024). Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, 13,* 1-14. https://doi.org/10.1145/3654777.3676450

United Nations. (2008). *International Standard Industrial Classification of All Economic Activities (ISIC), Rev. 4*. Statistical Papers, Series M, No. 4, Rev. 4. https://unstats.un.org/unsd/publication/seriesm/seriesm_4rev4e.pdf

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2201.11903

Zambia Statistics Agency. (2024). *2023 Labour Force Survey (LFS)*. https://www.mlss.gov.zm/wp-content/uploads/2024/11/2023-Labour-Force-Survey-Report-03112024-1.pdf

Assumptions and background for the two-proportion Z-test is taken from: https://web02.gonzaga.edu/faculty/rayr/ma121/Section9.2.pdf and https://www.qualitygurus.com/two-proportions-z-test-or-two-sample-z-test-for-proportions/.

9 Appendix

9.1 LLM Prompt

The following prompt instructs the LLM to classify survey responses into ISCO and ISIC codes based on the occupational title, main tasks and duties, and main activity, goods, or services.

prompt = (

"Classify the following Zambia Labour Force Survey responses into ISCO and ISIC categories (International Standard Classification of Occupations and International Standard Industrial Classification of All Economic Activities)."

"Your response MUST be a strict JSON array with keys: 'Unique_ID', 'D1_TITLE', 'D1_DESC', 'D2_MAIN_ACTIVITIES', 'ISCO_CODE_AI' and 'ISIC_CODE_AI'."

"DO NOT include any text before or after the JSON array."

"In your response for ISCO and ISIC Category columns, give the result at the 4 digit number - nothing else (no text, for example)."

"Ensure the columns 'Unique_ID', 'D1_TITLE', 'D1_DESC', 'D2_MAIN_ACTIVITIES' are filled out and kept from the original table."

"Accuracy is very important. Your ISCO and ISIC classifications should surpass the Zambian LFS enumerators."

"Try to make the responses as consistent as possible."

"The consistency of the responses is crucial to avoid JSON parsing errors."

)

9.2 Appendix Table 1: ISCO Comparisons

Occupational Title	**Main Tasks and Duties**	**Main Activity, Goods, or Services**	**Ground Truth Classification [Occupation Major Group]**	**LLM Classification [Occupation Major Group]**	**Enumerator Classification [Occupation Major Group]**
Subsistence Crop Farmer	Subsistence Crop Farmer Growing of Maize	Subsistence Crop Farmer Growing Crops for Sale	6310 [Skilled Agricultural, Forestry and Fishery Workers]	6221 [Skilled Agricultural, Forestry and Fishery Workers]	0110 [Armed Forces Occupations]
Retail Charcoal Seller	Selling Charcoal	Selling Charcoal	5249 [Services and Sales Workers]	5220 [Services and Sales Workers]	1420 [Managers]
Consultant	Consultation and Networking	Consultancy and Networking	2511 [Professionals]	2419 [Professionals]	4221 [Clerical Support Workers]
IT Specialist	Configuring Internet, Computer Maintenance	Banking Services	2511 [Professionals]	2523 [Professionals]	1330 [Managers]
Business Lady	Brewing and Selling Traditional Beer (Kachasu) at Mpulungu Market	Brewing and Selling of Traditional Beer (Kachasu) at the Market	5249 [Services and Sales Workers]	7431 [Craft and Related Trades Workers]	3339 [Technicians and Associate Professionals]
Council Police	Inforce Law for Local Government	Collecting of Tax and Cleaning of Town	3359 [Technicians and Associate Professionals]	5414 [Services and Sales Workers]	3355 [Technicians and Associate Professionals]

9.3 Appendix Table 2: ISIC Comparisons

Occupational Title	**Main Tasks and Duties**	**Main Activity, Goods, or Services**	**Ground Truth Classification [Industry Section]**	**LLM Classification [Industry Section]**	**Enumerator Classification [Industry Section]**
Beer Brewer	Brewing Local Beer	Brewing Local Beer	1101 [Manufacturing]	1103 [Manufacturing]	9609 [Other service activities]
Direct Sales Assistant	Direct Sales of Airtel Products	Sales of Phones and Accessories	4773 [Wholesale and retail trade; repair of motor vehicles and motorcycles]	4742 [Wholesale and retail trade; repair of motor vehicles and motorcycles]	6209 [Information and Communication]
Education Director	Education Planning	Certification of Accountancy	8550 [Education]	8542 [Education]	6920 [Professional, scientific, and technical activities]
Soldier	Protecting the Nation	Protecting the Country	8422 [Public administration and defence; compulsory social security]	8422 [Public administration and defence; compulsory social security]	9609 [Other service activities]
Building Care Taker	Looking After Somebody's Building	Looking After a Building	4100 [Construction]	6820 [Real estate activities]	9700 [Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use]
Driver (Soldier)	Driving Army Vehicles	Driving	4922 [Transportation and storage]	8422 [Public administration and defence; compulsory social security]	4923 [Transportation and storage]

theigc.org