Innovative Approaches to Enhance Data Quality in Official Statistics: A Case Study on Online Job Advertisement Data

Anca Maria Nagy[1], Eliane, Gotuzzo[2], Fernando Reis[3]

[1]Sogeti, Luxembourg, anca.kiss-nagy@sogeti.com

[2]Sogeti, Luxembourg, eliane.gotuzzo@sogeti.com

[3]Eurostat, Luxembourg, fernando.reis@ec.europa.eu

Abstract

The rapid proliferation of online job advertisements (OJA) has shown in a new era of non-traditional data sources, offering the potential to augment official statistics with real-time labour market insights. However, unlocking the potential of OJA data demands careful attention and emphasis on data quality. This paper delves into a methodology designed to assess and improve the quality of OJA data, with a particular focus on the occupation variable, within the context of official statistics.

The aim of this work is two-fold: to establish a robust quality monitoring procedure and to create a gold standard for OJA data. We leverage natural language processing (NLP) to scrutinize job descriptions and structured fields, utilizing both machine intelligence and human annotation. Importantly, our methodology introduces the incorporation of large language models, for quality assessment of labelled data and of occupation classification. This addition expands our ability to evaluate and classify occupations for specific samples that may pose challenges for classifiers or human annotators.

Notably, this study unfolds the interplay between human and machine intelligence, highlighting the potential of a combined approach in enhancing data quality. Human annotators contribute their domain expertise to the refinement of classifiers, resulting in a gold standard for OJA data. We calculate evaluation metrics, such as accuracy, to assess the quality of the labelled data, while machine learning models are trained on this human-labelled data for automated classification.

The results from this study, focusing on a selection of European countries, point to the need for improvements in occupation classification within OJA data. By providing an analysis of labelled OJA datasets, we offer insights into the accuracy and reliability of classifiers. We envision future iterations of this methodology to extend to more countries, encompass additional variables, and explore the development of refined ontologies.

This paper exemplifies the innovative approaches that official statistics agencies can employ to foster data quality in an era of emerging new data sources. It underscores the value of collaborative human-machine efforts and the pivotal role of advanced language models in enhancing data quality. By presenting this work, we hope to contribute to the ongoing discourse on innovation and research in official statistics, shedding light on novel strategies to ensure the highest quality data for informed decision-making and policy formulation.

Keywords: Official Statistics, Online Job Advertisements (OJA), Data Quality, Large Language Models, Human Annotation

1. Introduction

1.1 Aim of the study

1.1.1 Quality in official statistics

The use of alternative data sources, such as web-scraped or crowd-sourced data, in official statistics has gained traction in recent years. Studies have explored the potential of leveraging social media, web searches, and online platforms for augmenting traditional statistical methodologies. Within this landscape, online job advertisements have emerged as a particularly promising data source for labour market analysis Eurostat (2021), ESCO (2022), Smith et al (2020), Kiss-Nagy et al (2022).

Through digital platforms, job vacancies are increasingly being advertised online, presenting a wealth of real-time information about labour market dynamics. Traditional methods of data collection, although reliable, often suffer from delays and resource-intensive processes. In contrast, OJA data offer the potential of timelier insights into employment trends, job demand, and skill requirements. However, the integration of OJA data into official statistics frameworks necessitates careful consideration of data quality issues. Unlike structured surveys or administrative records, OJA data come with their own set of challenges, including variations in terminology, data incompleteness, and ambiguity in job descriptions. To realize the full potential of OJA data, it is imperative to develop innovative methodologies that address these quality concerns.

1.1.2 Human and LLM (Large Language Model) data annotation

In this paper, we present a novel approach to enhance the quality of OJA data within the context of official statistics. Our methodology focuses on the occupation variable, a critical component for understanding labour market dynamics. By leveraging a combination of natural language processing (NLP), human annotation (human labelling), machine learning techniques, and large language models (LLMs), we aim to establish a robust quality monitoring procedure and create a gold standard for OJA data.

We have incorporated large language models (ChatGPT-4), for the quality assessment of classifiers and of the human labelled data performed on the OJA sample. This addition expands our ability to evaluate and classify occupations for specific samples that may pose challenges for classifiers or human annotators. We intend to check to which extend the LLMs could replace human experts in the collection of labelled data for our OJA data collection.

1.1.3 Prompt engineering

An important aspect to consider when using LLMs is prompt engineering: a way to design the prompt message to the LLM, a sort of 'social skills' in the communication with the LLM. Despite employing identical tools and models, the outcomes can vastly differ based on the adeptness of the prompts used. It is crucial in influencing the quality of outputs produced by the LLMs, as the prompts essentially act as instructions or guidelines for the model to follow when generating text. Some of the reasons why prompt engineering is important are:

- *Control over Output*: Prompt engineering allows users to guide the model's output towards the desired direction.
- *Task Specificity*: Depending on the task at hand, prompt engineering enables users to tailor the model's output to suit specific requirements.
- *Mitigating Bias and Toxicity*: Well-designed prompts can help mitigate bias and toxicity in the model's outputs. By providing clear guidelines and examples, prompt engineering can steer the model away from generating inappropriate content.
- *Improving Coherence and Consistency*: Through prompt engineering, users can encourage the model to produce more coherent and consistent text. By providing relevant context and constraints within the prompt, users can guide the model towards generating consistent and desired outputs.
- *Optimizing Performance*: Effective prompt engineering can also optimize the performance of LLMs in specific applications. By experimenting with different prompts and fine-tuning their parameters, users can achieve better results in terms of accuracy, fluency, and relevance.

1.2 OJA use-case

Online job advertisements (OJA) refer to advertisements published on the World Wide Web revealing an employer’s interest in recruiting workers with certain characteristics for performing certain work. OJA include data on the characteristics of the job (e.g., occupation and location), characteristics of the employer (e.g., economic activity) and job requirements (e.g., education and skills). The data used in this paper is the Web Intelligence Hub's Online Job Advertisement (OJA) database, developed by Cedefop and Eurostat in Cedefop (2020). This dataset covers over one hundred million ads posted since July 2018 in more than three hundred web sources including job search engines and public employment services’ websites, together covering the labour markets of all EU countries and the UK. We focus our analysis on the period of October 2021 – January 2022.

The OJA data is classified using an information extraction process, defined through a set of processing pipelines. A pipeline refers to a part of information extraction dealing with a specific piece of content (attribute or variable) and a specific language. Each pipeline is tailored to a specific attribute (like occupation, skill, or salary) and language, creating a scalable system where each language or attribute added multiplies the number of pipelines. The pipelines are designed to be executed independently, allowing for selective re-running if necessary. The information extraction involves several stages:

1. Language Detection: Identifies the advertisement's language to select the appropriate pipeline.
2. Pre-processing: Addresses data issues like noise and uniformity across languages.
3. Ontology-Based Models: Classifies ads using a hierarchical approach, starting with exact matches in metadata and progressing to more general searches if needed.
4. Machine-Learning Classifier: Utilized if ontological methods fail, employing algorithms like Naïve Bayes trained on substantial datasets.

This classification framework combines ontologies and machine learning, providing flexibility and enhancing the accuracy of classification through both structured and unstructured data.

2. Methods to improve quality of OJA data: An OJA gold standard

2.1 Use of OJA-NLP dataflow

In our study we use the so-called Natural Language Processing (NLP) dataflow, that includes all relevant information accessible in online job advertisements (OJA), including the complete job description and supplementary text extracted from structured fields like job title and salary details. The primary objective is to conduct a thorough quality analysis and enhance the accuracy of OJA data by refining OJA classification algorithms.

For quality control purposes, the OJA data collection incorporates a dataflow enriched with additional information on OJAs – the OJA-NLP dataflow. This enhanced dataflow facilitates the evaluation of the accuracy of automatic classifiers, which extract skills, occupations, and other statistical variables from the natural language text in job descriptions and structured fields on job portals. By comparing the text in job ads with the outcomes generated by the classifiers, statisticians and users can assess the accuracy of the results.

2.2 Data annotation

2.2.1 Methods to get annotated data

High-quality annotated data is fundamental for training and refining machine learning models, enabling them to perform with higher accuracy, and to obtain accurate estimates of the precision of automatic classifiers. However, creating a robust dataset with accurate annotations is a significant challenge due to the time, cost, and expertise required. Given the importance of annotated data, it is imperative to explore various methods of data collection that balance cost, scale, and quality. Methods for data annotation include:

- *Crowdsourcing*: Use platforms like Amazon Mechanical Turk or Scale AI to distribute annotation tasks.
- *In-House Teams*: Establish a team of annotators, either full-time or contractors, for direct supervision.
- *Academic Collaboration*: Partner with academic institutions to engage students or researchers in annotation.
- *Subject Matter Experts*: Enlist experts in the relevant field for high-quality annotations.
- *Semi-Automated Tools*: Employ tools combining machine learning with human input for faster annotation.
- *Collaborative Platforms*: Use platforms like Labelbox or Prodigy for real-time teamwork among annotators.
- *Outsourcing Services*: Outsource annotation tasks to specialized service providers.
- *Active Learning*: Implement techniques to select the most informative data points for annotation.

These methods offer diverse approaches to effectively annotate data, enhancing its quality for various applications, but they are quite expensive, or they are limited to the use of most common languages.

2.2.2 Human annotation

We first performed human data annotation on our selected sample, which involves labelling raw data through human classification, a critical process for enriching linguistic data collection with annotations for quality assurance. The annotated corpus, comprising a single set of data annotated with the same specifications, serves as a vital input in the OJA data classification process. Each OJA variable is associated with an ontology, which consists of a list of keywords for specific classification categories.

Ensuring the quality of annotated data improves the overall quality of the annotated corpus, which is utilized to train, validate, and test machine-learning algorithms or estimate the precision of classifiers. Various measurements, such as accuracy and consistency, are employed to evaluate the quality of annotated data. Standard methods, including gold standards, consensus labelling, and auditing, are utilized to ensure high-quality annotated data.

The so called “gold standard”, or "ground truth," serves as the best benchmark for evaluating classification outcomes. It is established through human labelling of a predefined randomly selected sample, adhering to desirable requirements such as uncertainty, diversity, and random sampling. The importance of human errors in training data varies depending on the use case, with supervised learning models becoming more accurate with increased quality labelled data. Active learning assists in determining which data to sample for human annotation.

Doccano was used as a tool to gather annotations, providing features for text classification, sequence labelling, and sequence-to-sequence tasks. As an open-source tool for text human annotation, Doccano allows the collection of labelled data by creating projects, uploading data, and initiating the annotation process. In the context of the OJA data, Doccano is employed within the human-in-the-loop evaluation procedure.

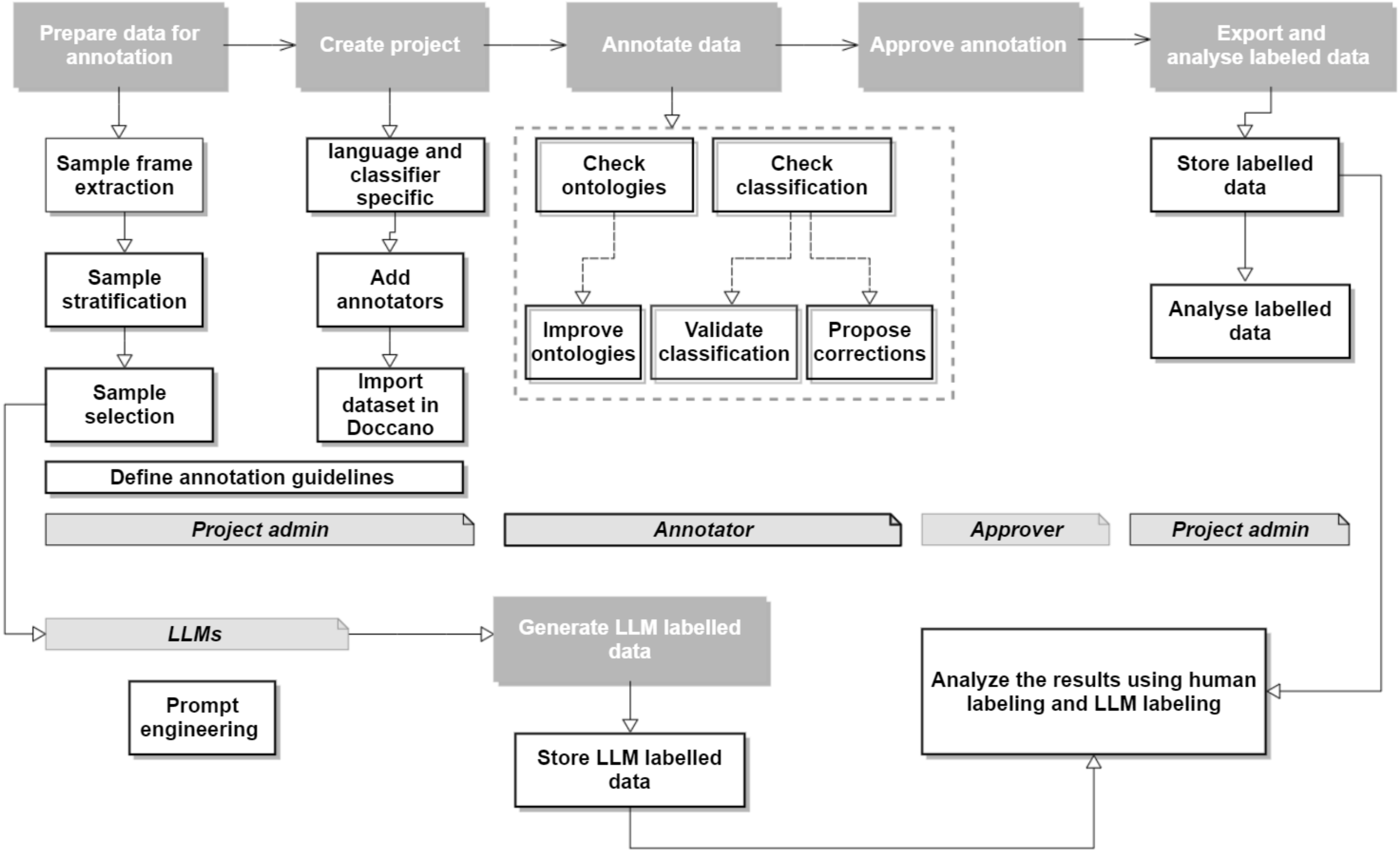


Figure 1: Workflow for the collection of labelled data for OJA using human experts and LLM

2.2.3 Human annotation of OJA data sample

2.1.3.1 OJA sampling

The gold standard sampling for OJA utilizes the OJA-NLP dataflow. We focus on analysing the occupation classifier, creating a stratified evaluation dataset. This dataset includes raw data from a sample of ads, encompassing full job descriptions, stratified by classification outcomes (up to ISCO-08 level 4) and combinations of country and language. It also includes job description tokens and dictionary terms matched by the algorithm. The stratification ensures representation across classification outcomes. The sample excludes OJAs no longer available online to avoid any negative effects on website owners. Future evaluations will occur annually, leading to:

- Evaluation metrics for classification algorithms (e.g., accuracy rate).
- Suggestions for improving keyword sets (i.e., "ontologies").
- Growing human-labelled data for ML model training.
- Complementing human-labelled data with LLM generated labelled data, faster to obtain and nowadays more accurate, in combination with human check.

2.1.3.2 Definition of specific metadata labels for OJA annotation

The annotation sought is the classification of the occupation into ISCO classes at 4 digits level (or lower levels if not possible at 4 digits). The labels include several ISCO classes and “metadata” labels designed to capture other situations. The metadata labels defined for this human annotation are as follows: “*Correct*”, “*Incorrect*”, “*Comment*”, “*No reference to occupation in the description*”, “*Impossible to classify at 4th level*”, “*Wrong language*”, “*Not a job ad*”, “*Job description missing*”, “*Multiple ISCO labels*”, “*Misspelling*”.

Annotators from the Web Intelligence Network, comprising labour market statistics experts and web intelligence specialists, performed annotation using Doccano. Each project within Doccano corresponds to a country-language pair, with specific labels created for annotation. Annotators have been assigned to these projects, and annotated data has been collected and analysed. The results of the first round of labelling exercise for the OJA gold standard samples showed a need for improvement of the classifier for the occupation variable, as presented in Nagy et al., 2023.

Future rounds of human labelling will help achieve a gold standard for OJA data. The extension of this methodology for more countries and for additional variables (e.g., classifiers) and considering suggestions for improving ontologies will contribute to a higher quality of OJA data.

Table 1: Analysis of human labelled OJA datasets

Country	**AT**		**BG**		**IT**		**PL**		**RO**		**SI**	
OJA label	**n***	**Aw**** (%)	**n**	**Aw** (%)	**n**	**Aw** (%)	**n**	**Aw** (%)	**n**	**Aw** (%)	**n**	**Aw** (%)
Correct	161	49.87	147	47.43	185	51.87	220	0.56	226	0.64	143	58.54
Incorrect	223	50.13	177	45.58	203	46.59	177	0.43	151	0.36	177	39.96
Comment	40	11.07	0	0	8	0.73	0	0	26	0.08	0	0
No reference to occupation in description	20	3. 67	1	0.01	1	0.01	0	0	14	0.04	1	0.01
Impossible to classify at 4th level	16	4.79	0	0	1	0.00	0	0	11	0.04	2	0.37
Wrong language	2	0.04	28	1.75	0	0	0	0	6	0.01	0	0
Not a job ad	3	0.88	7	0.10	11	2.46	1	0.00	10	0.02	6	0.84
Job description missing	82	16.06	5	0.64	0	0.00	0	0	1	0.00	1	2.43
Multiple ISCO labels	45	9.97	0	0	5	0.74	0	0	13	0.06	0	0
Total ads labelled		403		355		397		398		380		329

* n – number of times a label was used by the annotator to annotate the OJA sample selected for a specific country. The percentage sum of Aw is not 100% because of multiple labelling of the same ad (for example, the labels "*Incorrect*", "*Comment*" and "*Multiple ISCO labels*" are often used together. In addition, the sum of "Correct" and "Incorrect" <100% suggests that some ads were not labelled at all.

** Aw – weighted accuracy rate (in %): weighted on the proportion of job ads in each occupation compared to the total job ads.

3. Use of LLM to label data

3.1 Leveraging Generative Pre-Trained Transformers (GPT) for Cost-effective Text Annotation and Classification

Text annotation and classification tasks traditionally rely on human annotators, which can be both time-consuming and costly. However, the launch of LLMs such as GPT-3 and its subsequent iterations has offered promising alternatives that merit exploration.

3.1.1 GPT-3: Bridging the Gap Between Human and Machine Annotation

Initial studies, as referenced by Doe et al. (2022), suggested that while GPT-3 was indeed helpful for text annotation tasks, its performance fell short of surpassing human annotators. Despite this, GPT-3 demonstrated considerable utility, highlighting its potential as a viable complement to human annotators, particularly in scenarios where time or resources are limited.

3.1.2 GPT-3.5: Towards Parity with Human Annotators

Building upon the foundation laid by GPT-3, subsequent iterations such as GPT-3.5 have made significant strides towards achieving parity with human annotators. Research by Smith et al. (2023) indicates that GPT-3.5 yields result comparable to those obtained through crowdsourced annotation, underlining its potential as an efficient and reliable annotator when provided with adequate guidance and exemplars.

3.1.3 GPT-4: Surpassing Human Benchmarks

Recent developments, as highlighted by Roe et al. (2024), suggest that GPT-4 has surpassed human annotators in specific benchmark tasks, signifying a notable milestone in the evolution of LLMs for text annotation. Furthermore, the cost-effectiveness of employing GPT-4 for annotation tasks appears promising, with per-annotation costs lower than traditional human annotators and even surpassing the efficiency of services such as Mechanical Turk.

3.1.4 The Potential of LLMs for Text Classification

Beyond annotation tasks, LLMs hold immense potential for enhancing the efficiency of text classification. Research conducted by Brown et al. (2023) and the once previously cited, illustrate the capability of LLMs to drastically increase the efficiency of text classification, paving the way for streamlined workflows and enhanced productivity in various domains. The landscape of text annotation and classification is undergoing a transformative shift with the emergence of LLMs such as GPT-3 and its successors. While challenges persist, recent advancements indicate significant developments towards cost-effective and efficient solutions leveraging these powerful models. As research continues to unfold, the potential of LLMs to revolutionize text annotation and classification workflows remains an area for exploration and innovation. However, it is essential to note that human annotation and quality control remain crucial for achieving high-quality results. While LLMs can expedite the data labelling process, human oversight is still necessary to ensure accuracy and reliability in the annotated data.

3.2 Generating LLM labelled data for OJA sample

There are several ways to use LLMs to generate annotated data (Figure 3), including:

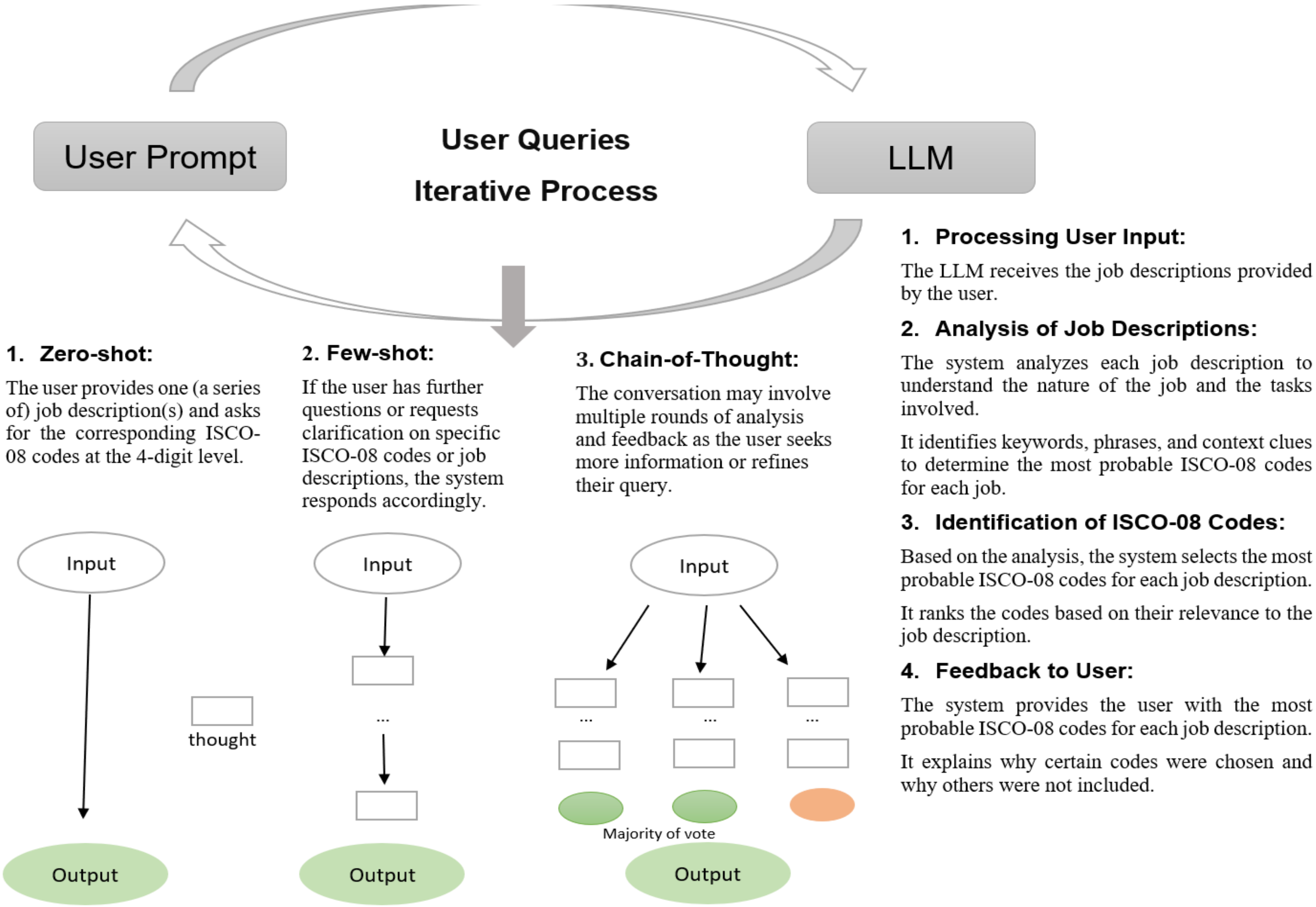
- *Zero-shot prompting*: The user provides one (a series of) job description(s) and asks for the corresponding ISCO-08 codes at the 4-digit level.
- *Few-shot prompting*: The user provides examples of good responses or additional information useful for the LLM response. The user may have further questions or requests clarification on specific ISCO-08 codes or job descriptions, the system responds accordingly.
- *Chain-of-Though (CoT)*: The LLM breaks down complex problems into a series of intermediate reasoning steps before providing the final answer. The conversation may involve multiple rounds of analysis and feedback as the user seeks more information or refines their query, minimizing the chances of the model leaping to illogical conclusions.
- *Prompt-chaining* (or C*hain-of-Prompts CoP*): The user performs task decomposition, meaning that the overall task is divided into more manageable subtasks, each handled by an individual prompt.
- *Tree-of-thought (ToT)*: more advanced than the CoT (individual workflow for each prompt), as it combines information from several prompts to get the final answer.

3.3 Assess the human labelled data using LLMs for OJA data

3.3.1 Collection of LLM labelled data for OJA sample

We have incorporated LLMs - large language models (ChatGPT-4), for the quality assessment of classifiers and of the human labelled data performed on the OJA sample. This addition expands our ability to evaluate and classify occupations for specific samples that may pose challenges for classifiers or human annotators. Our aim is to compare the ISCO-08 classifications from different sources:

- The ISCO-08 code predicted by an automatic (machine learning, ontology based) classifier.
- The judgment made by a human expert.
- The classification proposed by LLM (assume may be another model or system and re-checked by human – the user).



1. **Zero-shot:**

The user provides one (a series of) job description(s) and asks for the corresponding ISCO-08 codes at the 4-digit level.

2. **Few-shot:**

If the user has further questions or requests clarification on specific ISCO-08 codes or job descriptions, the system responds accordingly.

3. **Chain-of-Thought:**

The conversation may involve multiple rounds of analysis and feedback as the user seeks more information or refines their query.

1. **Processing User Input:**

The LLM receives the job descriptions provided by the user.

2. **Analysis of Job Descriptions:**

The system analyzes each job description to understand the nature of the job and the tasks involved.

It identifies keywords, phrases, and context clues to determine the most probable ISCO-08 codes for each job.

3. **Identification of ISCO-08 Codes:**

Based on the analysis, the system selects the most probable ISCO-08 codes for each job description.

It ranks the codes based on their relevance to the job description.

4. **Feedback to User:**

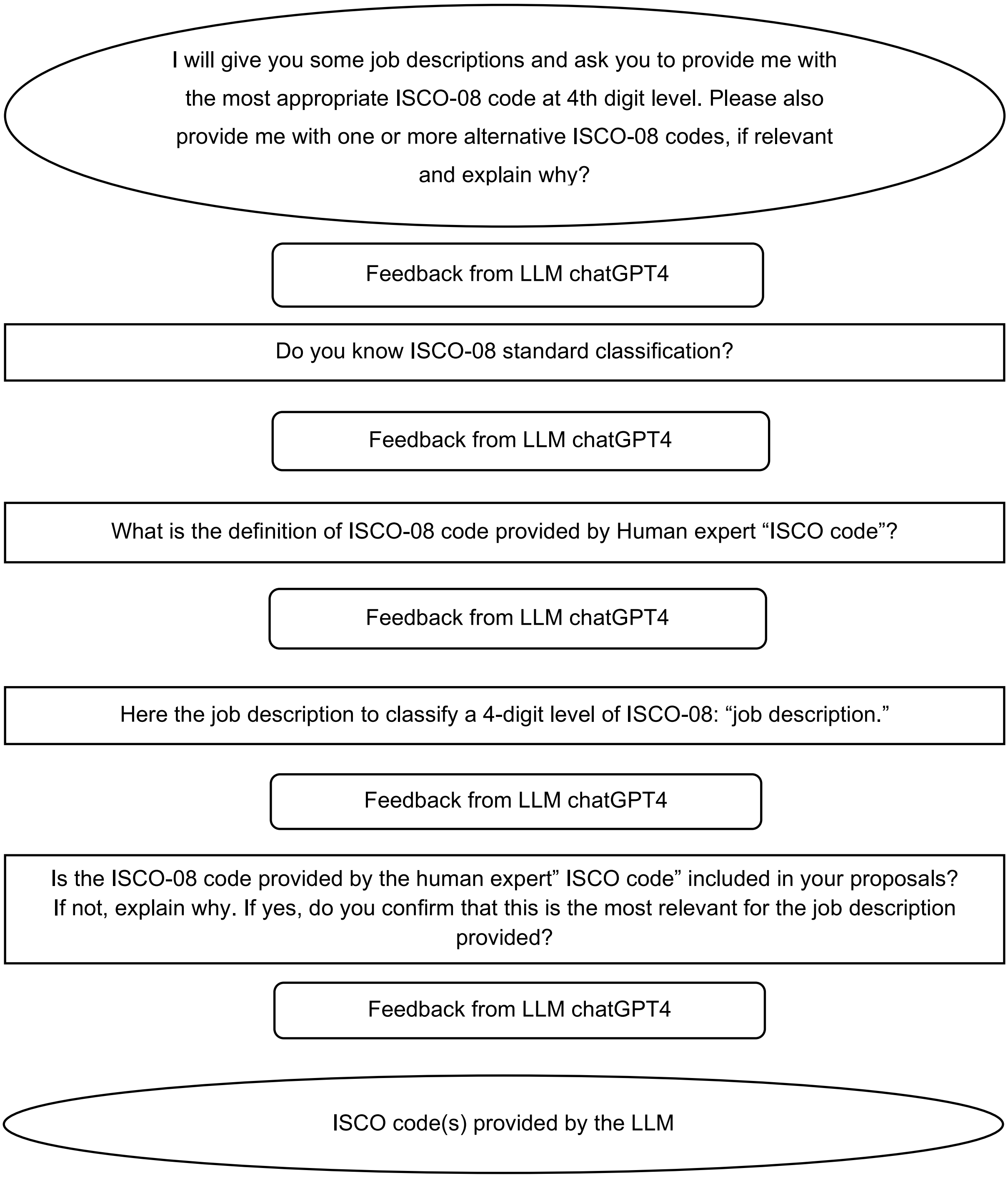
The system provides the user with the most probable ISCO-08 codes for each job description.

It explains why certain codes were chosen and why others were not included.

Figure 3: Workflow of collecting LLM labelled data

The goal is to assess the agreement or discrepancy between these sources of labelled data on a given set of job ads, which includes verifying the language of the job ad and whether the text describes a job.

The approach to prompting employed in this exercise adopts the "few-shot prompting" method of interacting with the LLM. Please find below an example of the prompting used:



I will give you some job descriptions and ask you to provide me with the most appropriate ISCO-08 code at 4th digit level. Please also provide me with one or more alternative ISCO-08 codes, if relevant and explain why?

Feedback from LLM chatGPT4

Do you know ISCO-08 standard classification?

Feedback from LLM chatGPT4

What is the definition of ISCO-08 code provided by Human expert “ISCO code”?

Feedback from LLM chatGPT4

Here the job description to classify a 4-digit level of ISCO-08: “job description.”

Feedback from LLM chatGPT4

Is the ISCO-08 code provided by the human expert” ISCO code” included in your proposals? If not, explain why. If yes, do you confirm that this is the most relevant for the job description provided?

Feedback from LLM chatGPT4

ISCO code(s) provided by the LLM

We have analysed the agreement between human experts and LLM classifications for the OJA sample. More specifically, we have first calculated the agreement rate. This indicates how often the human expert's classification matches the LLM's classification. Second, we have performed identification of discrepancies: the code groups the data by both human and LLM classifications where they do not match and counts these occurrences, helping identify the most common mismatches. This analysis will help understand the consistency between human and LLM in classifying job ads, highlighting areas where they align or differ significantly.

Table 2: Analysis of labelled OJA dataset: Human expert and LLM for '*ro*' sample

Country (RO)	**Human expert labelled data**	**LLM generated labelled data**
OJA metadata label	**n**	
*Correct **	226*	18*
*Incorrect **	151*	310*
Impossible to classify at 4th level	11	0
Wrong language	6	8
Not a job ad	10	38
Job description missing	1	3
Multiple ISCO4D labels	13	36
Total ads labelled	380	328**

* '*Correct*' and '*Incorrect*' attributes are given in comparison with the OJA classifier that we want to assess

**In collecting labelled data using LLMs, we have excluded from the initial OJA sample (labelled by the human expert) the cases where '*job description missing*', '*wrong language*', '*not a job ad'*

Table 2 presents the results of a labelling exercise comparing human expert-labelled data with data labelled by a Large Language Model (LLM) for a sample of online job advertisements (OJA) from Romania (RO). When comparing human expert-labelled data with LLM-generated labelled data, several observations can be made:

1. *Correct and Incorrect Labels*: the substantial difference suggests a disparity in the accuracy of labelling between the human expert and the LLM (ChatGPT-4). However, it is important to consider that the human expert's judgments may have been influenced by their awareness of the OJA classifier's results. This potential bias could have led to the human expert being more conservative in labelling ads as correct, while the LLM, not being aware of the OJA classifier's results, may have provided more varied classifications.
2. *Multiple ISCO4D Labels*: Both the human expert and the LLM (ChatGPT-4) encountered cases where multiple ISCO4D labels were assigned. This suggests that

both the human expert and the LLM faced challenges in accurately classifying certain ads with multiple job categories.

3. *Potential Bias in Human Expert Judgments*: The hypothesis suggests that the human expert's judgments may have been influenced by their awareness of the OJA classifier's results when labelling the sample. This potential bias could have impacted the accuracy and consistency of the human expert-labelled data. In contrast, the LLM classification was not affected by this bias since the result of the OJA classifier was not provided in the prompting.

3.3.2 Agreement rate

Agreement rates at ISCO 4- digit, ISCO 3- digit, ISCO 2- digit and ISCO 1- digit are presented in the table 3 below.

Table 3: Agreement rate at ISCO-08 at respectively 4-, 3- and 2-digit levels between human expert, OJA classifier and LLM classification

Agreement rate	**ISCO-08 4D**	**ISCO-08 3D**	**ISCO-08 2D**	**ISCO-08 1D**
Human expert – LLM (chatGTP-4)	9.5 %	25.93 %	45.83 %	62.5 %
OJA classifier – Human expert	58.71 %	63.76 %	66.05 %	70.64 %
OJA classifier – LLM (chatGPT-4)	6.7 %	20.68 %	35.86 %	53.59 %

OJA Classifier vs. Human Expert

4-Digit Level (58.71%): This moderate to high agreement rate suggests that the OJA classifier does a reasonably good job at matching the human expert’s classification at a very detailed level. It implies good training and tuning of the classifier based on human-labelled data. However, it is crucial to note that the human expert's judgments may have been influenced by their awareness of the OJA classifier's results when labelling the sample. This potential bias did not affect the LLM classification, as the result of the OJA classifier was not provided in the prompting.

Human Expert vs. LLM (ChatGPT-4)

4-Digit Level (9.5%): This low agreement rate suggests that there is a significant disparity in specific job classifications between the human expert and the LLM. The detailed 4-digit classification requires precise understanding and categorization, which might be challenging for the LLM to match human-level granularity and contextual interpretation. However, the human expert's judgments may have been influenced by their awareness of the OJA classifier's results. This potential bias could have led to the human expert being more conservative in labelling ads as correct, while the LLM, not being aware of the OJA classifier's results, may have provided more varied classifications.

OJA Classifier vs. LLM (ChatGPT-4)

4-Digit Level (6.7%): The low agreement rate here is striking and indicates that the OJA classifier and LLM significantly diverge in their detailed job classifications. This could be due to different training data, model capabilities, or inherent biases in each approach.

3-Digit Level (20.68%), 2-Digit Level (35.86%), and 1-Digit Level (53.59%): These increasing rates suggest that as the classification becomes more general, the LLM starts to better align with the classifications derived from the OJA classifier, though still lagging the alignment seen between human experts and the OJA classifier.

Here below in Figure 4, you can find a visual comparison between the OJA classifier, the LLM classification and additional proposals and the human expert investigation.

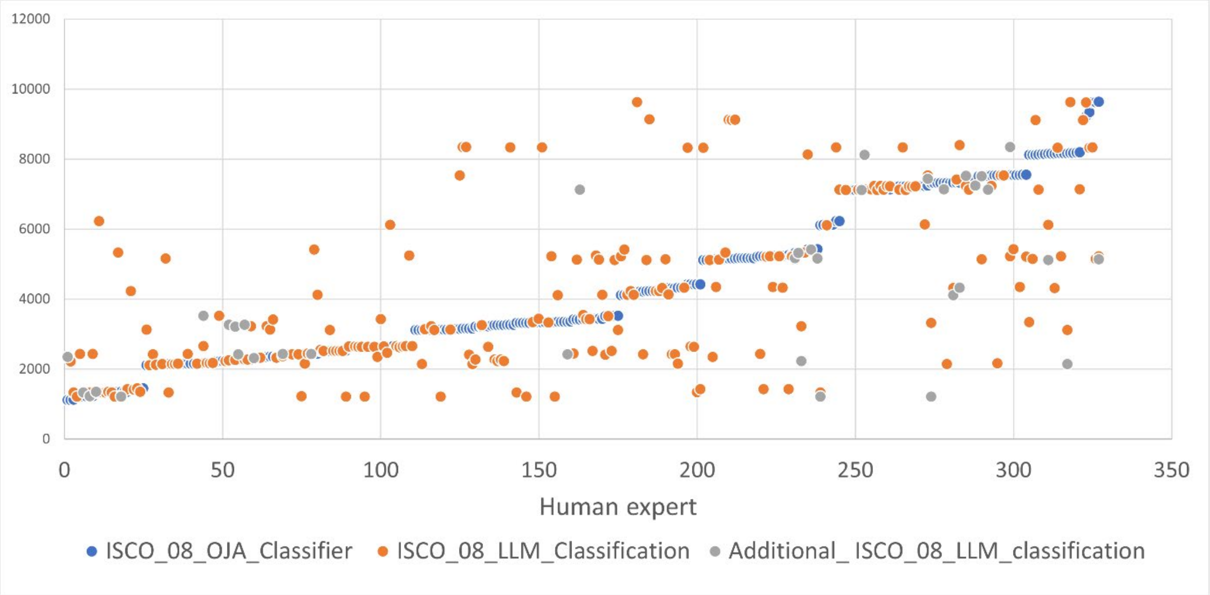


Figure 4: Comparison between OJA classifier (blue), and LLM (with additional proposals when applicable) by human expert opinion

4. Results and Conclusions

In this paper, we introduce and test an innovative method to improve the quality of Online Job Advertisement (OJA) data, crucial for analysing labour market trends. Our approach integrates natural language processing, human annotation, machine learning, and large language models (LLMs) like GPT-4, to develop a comprehensive quality monitoring system and establish a gold standard for OJA data. We also explore the potential of LLMs to supplement or replace human experts in labelling data, enhancing our capacity to assess and categorize occupations effectively.

We have analysed the agreement between human experts and LLM classifications for the OJA sample. The pattern observed across all comparisons (OJA classifier, human expert and LLM) shows increasing agreement rates with decreasing granularity of the ISCO codes, which is expected in classification systems. The LLM's lower performance, particularly at more detailed levels, could point to limitations in its training or the inherent complexity of processing nuanced job descriptions compared to a specialized OJA classifier. This suggests a need for further investigation of more advanced prompting techniques of the LLM on specific job classification tasks, such as chain-of-thought (CoT) or tree-of-thought (ToT), or possibly integrating human oversight to ensure accuracy at more detailed levels. Nevertheless, it is also important to note and further investigate the fact that the human expert's judgments may have been influenced by their awareness of the OJA classifier's results when labelling the sample. This potential bias did not affect the LLM classification, as the result of the OJA classifier was not provided in the prompting.

These results highlight the new strategies that official statistics agencies can adopt to enhance data quality in the face of new and emerging data sources. It emphasizes the importance of combining human expertise with machine capabilities and the critical role that advanced language models play in improving data quality. Through sharing this work, we aim to contribute to the ongoing discussion about innovation and research within official statistics, employing innovative methods to secure the highest quality data for informed policy- and decision-making.

References

Eurostat. (2021). Competition in Urban Hiring Markets: Evidence from Online Job Advertisements. Eurostat working papers.

ESCO (2022). The Use of Online Job Sites for Measuring Skills and Labour Market Trends: A Review Oleksii Romanko and Mary O'Mahony ESCoE Technical Report No. 2022-19 May 2022.

Smith, J., Johnson, E., & Lee, D. (2020). The use of alternative data sources for labour market analysis: A focus on online job advertisements.

Kiss-Nagy, A., Marconi, G., Paulino, R., Bitoulas, A., Reis, F., (2022). Path to a quality framework for Online Job Advertisement data, European Conference on Quality in Official Statistics (Q2022), Vilnius, Lithuania.

Cedefop. (2020). Cedefop and Eurostat formalise joint approach to online job advertisement data. Retrieved from Cedefop website.

Wissler, L., Almashraee, M., Monett, D., Paschke, A. (2014). The Gold Standard in Corpus Annotation. 10.13140/2.1.4316.3523.

Monarch R. (Munro) (2021), Human-in-the-Loop Machine Learning, Active learning and annotation for human-centered AI, Simon and Schuster.

Lawn, M., & Nóvoa, A. (2013). The European Educational Space: New Fabrications. *Sisyphus – Journal of Education*, *1*(1), 11-17. https://doi.org/10.25749/sis.2827.

Nagy, A.-M., Reis, F. (2023) Development of an OJA gold standard to classify occupation, Conference on New Techniques and Technologies for Statistics (NTTS) – Brussels, 6-10 March 2023.

Doe, J., Smith, A., & Brown, C. (2022). "Want To Reduce Labeling Cost? GPT-3 Can Help." arXiv preprint arXiv:2108.13487.

Smith, A., Roe, B., & Jones, D. (2023). "Making Large Language Models to Be Better Crowdsourced Annotators." arXiv preprint arXiv:2303.16854.

Roe, B., Smith, A., & Taylor, E. (2024). "Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark." arXiv preprint arXiv:2304.03279.

Brown, C., Taylor, E., & Johnson, F. (2023). "ChatGPT just outperformed Mechanical Turk workers on text annotation tasks." arXiv preprint arXiv:2303.15056.