EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Innovative Approach to Enhance Data Quality in Official statistics: OJA use case

Anca Maria Nagy, Eliane Gotuzzo

WIH methodology team, Sogeti Luxembourg

Fernando Reis

Eurostat, WIH methodology team

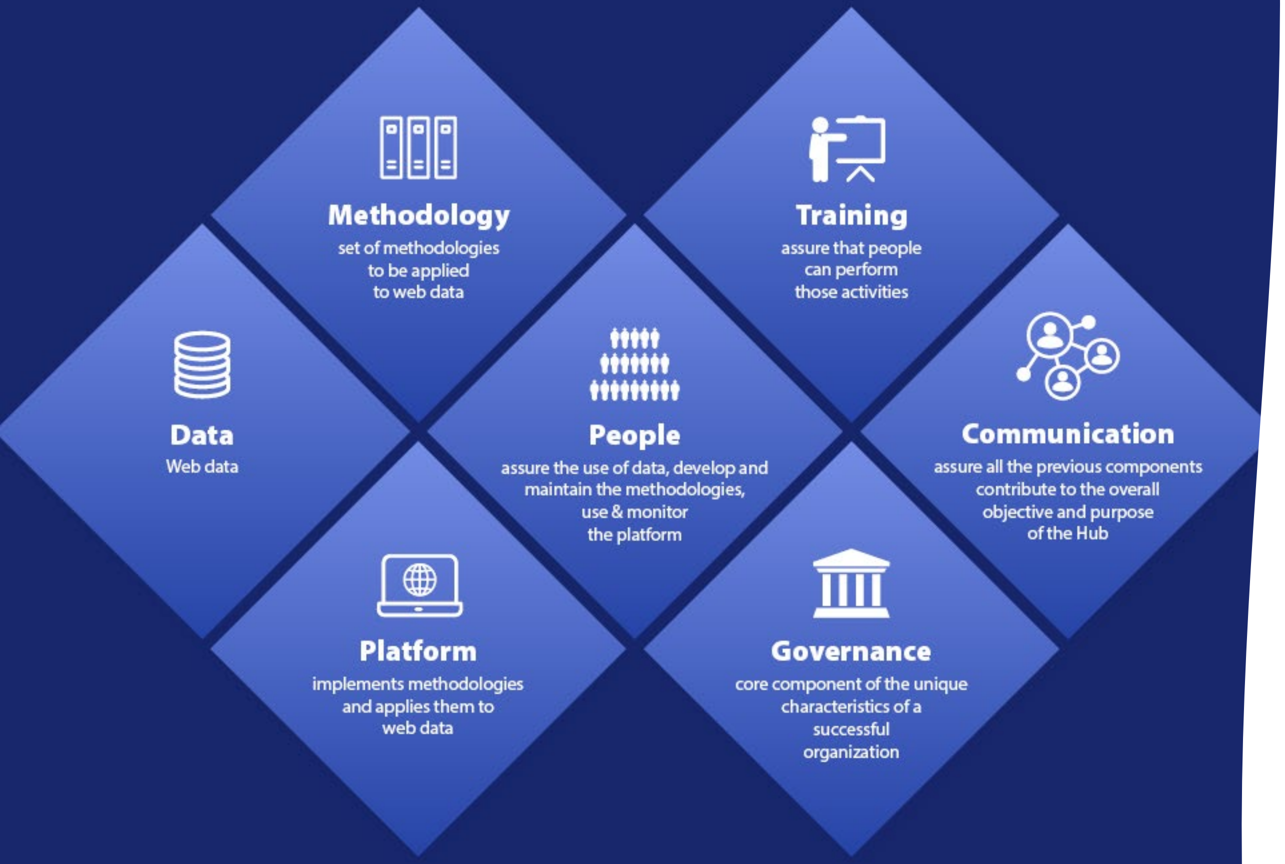
Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

Web
Intelligence
Hub

Web Intelligence Hub



Methodology
set of methodologies to be applied to web data

Training
assure that people can perform those activities

Data
Web data

People
assure the use of data, develop and maintain the methodologies, use & monitor the platform

Communication
assure all the previous components contribute to the overall objective and purpose of the Hub

Platform
implements methodologies and applies them to web data

Governance
core component of the unique characteristics of a successful organization

Web Intelligence Hub (WIH)

❑ The **WIH** is the **pillar** of TSS that provides the fundamental building blocks for harvesting information from the web to produce statistics

❑ **Mission**: *“a high-quality source of data extracted from web content, methodologies and algorithms, ready to be used to produce European and national official statistics”*

❑ **Collaborative effort:** Eurostat, NSIs, statistical authorities and partners

❑ **Community of experts**: Web Intelligence Network, CEDEFOP

❑ **WIH Platform**: technical components and services

❑ **Current use cases**:

- **Online Job Advertisements**,
- Online Based Enterprise Characteristics (OBEC),
- Multinational Enterprises (MNE)

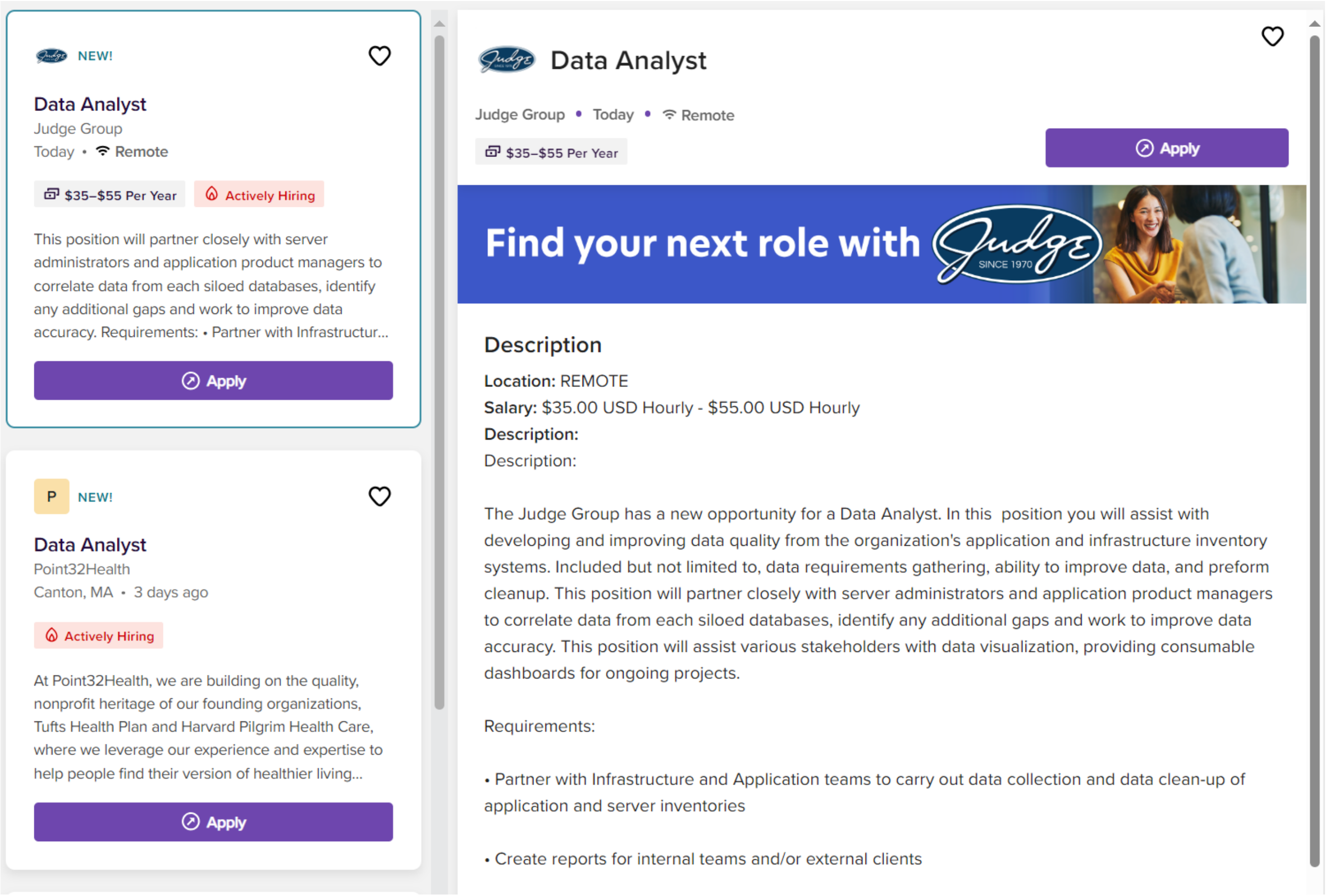
OJA data

Online Job Advertisements

- ❑ Web data source
- ❑ Advertisements published on the World Wide Web:
  - Reveal an employer’s interest in recruiting workers with certain characteristics for performing certain work
- ❑ 200 million ads
  - Posted in EU countries & UK - July 2018
  - Collected from more than 600 web sources (job search engines, public employment services’ websites…)
- ❑ Classified data (ISCO, ISCED, NACE, NUTS)
  - Language Detection
  - Pre-processing: noise detection, …
  - Ontology-Based Models
  - Machine-Learning Classifier

OJA-NLP Dataflow

- ❑ OJA classifiers only use the **job title** to classify the ‘*occupation*’ ISCO-08
- ❑ Explore richness of the information extracted from OJAs:
  - Full description of the job ads
  - Additional text from structured fields (raw text on job title, salary, etc.)



Data Analyst
Judge Group
Today • Remote
$35–$55 Per Year
Actively Hiring

This position will partner closely with server administrators and application product managers to correlate data from each siloed databases, identify any additional gaps and work to improve data accuracy. Requirements: • Partner with Infrastructur...

P

Data Analyst
Point32Health
Canton, MA • 3 days ago
Actively Hiring

At Point32Health, we are building on the quality, nonprofit heritage of our founding organizations, Tufts Health Plan and Harvard Pilgrim Health Care, where we leverage our experience and expertise to help people find their version of healthier living...

Data Analyst

Judge Group • Today • Remote
$35–$55 Per Year

Description

Location: REMOTE
Salary: $35.00 USD Hourly - $55.00 USD Hourly
Description:
Description:

The Judge Group has a new opportunity for a Data Analyst. In this position you will assist with developing and improving data quality from the organization's application and infrastructure inventory systems. Included but not limited to, data requirements gathering, ability to improve data, and preform cleanup. This position will partner closely with server administrators and application product managers to correlate data from each siloed databases, identify any additional gaps and work to improve data accuracy. This position will assist various stakeholders with data visualization, providing consumable dashboards for ongoing projects.

Requirements:

• Partner with Infrastructure and Application teams to carry out data collection and data clean-up of application and server inventories

• Create reports for internal teams and/or external clients

Build a gold standard

Labelling = Annotation

Analyze the quality of the OJA data production system

Evaluation of classifiers

Perform quality checks of the data classified

Measurement of the accuracy of the classifiers

Why to Collect labelled data?

Build a **gold standard**

Monitor the quality of automatic classification process

Benchmark Human annotators

Explore the possibility to complement human labelled data with LLM labelled data

Data labelling: Gold standard

Quality of labelled data

Gold standard:

sample carefully built by experts with a very high precision.

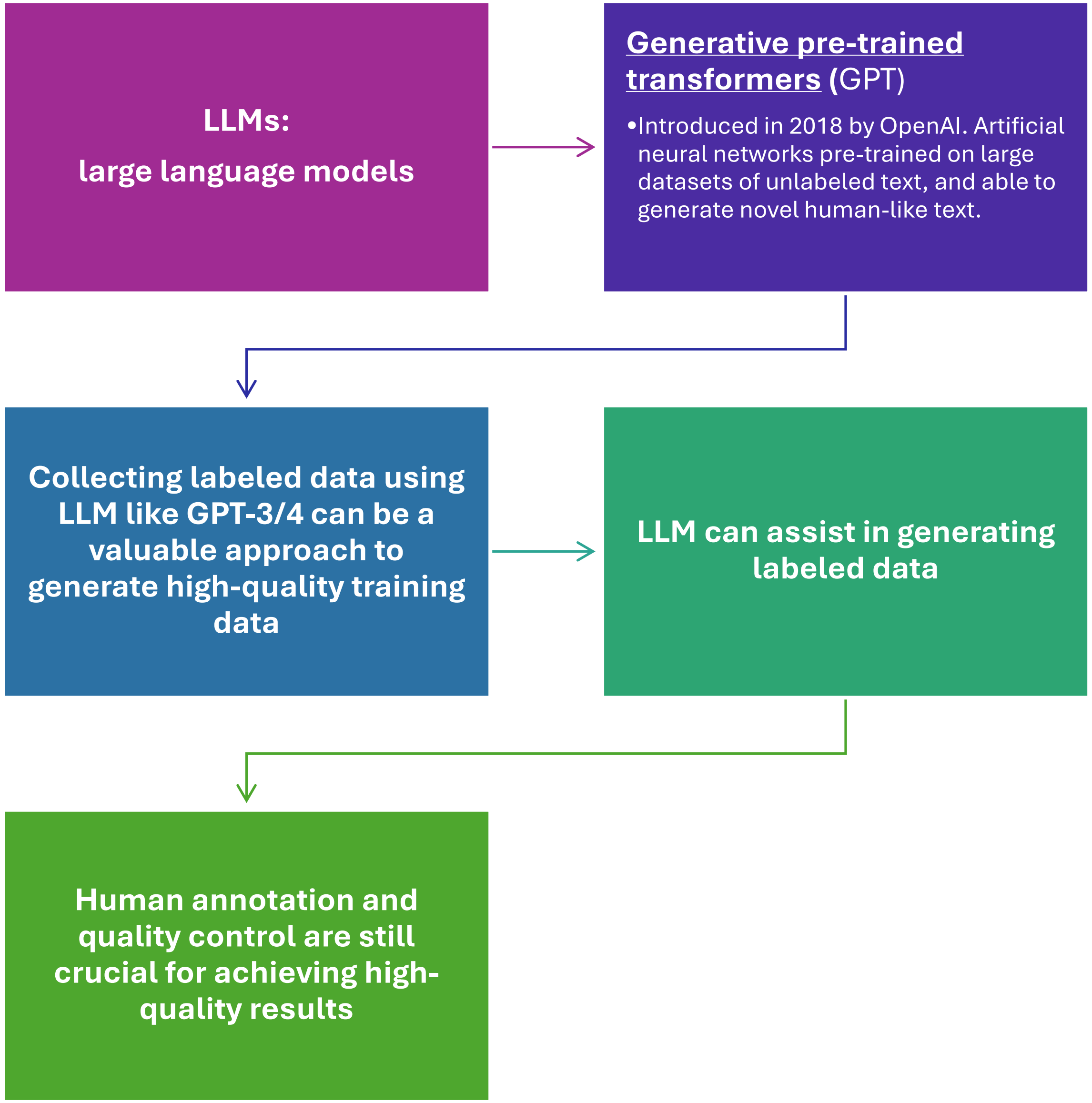
Expensive:

require the intensive teamwork of highly qualified experts.

Small in relative terms and are normally not be sufficient for training ML models.

Ideal for **benchmarking annotators** used to obtain other types of annotated data.

Use of LLMs to collect labelled data



LLMs:

large language models

Generative pre-trained transformers (GPT)

- Introduced in 2018 by OpenAI. Artificial neural networks pre-trained on large datasets of unlabeled text, and able to generate novel human-like text.

Collecting labeled data using LLM like GPT-3/4 can be a valuable approach to generate high-quality training data

LLM can assist in generating labeled data

Human annotation and quality control are still crucial for achieving high-quality results

Use of LLMs to collect labelled data: Performance - literature

Want To Reduce Labeling Cost? GPT-3 Can Help

GPT-3 was helpful but not better than humans (https://doi.org/10.48550/arXiv.2108.13487)

Making Large Language Models to Be Better Crowdsourced Annotators

GPT-3.5 is about on par w/ humans (https://doi.org/10.48550/arXiv.2303.16854)

Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark

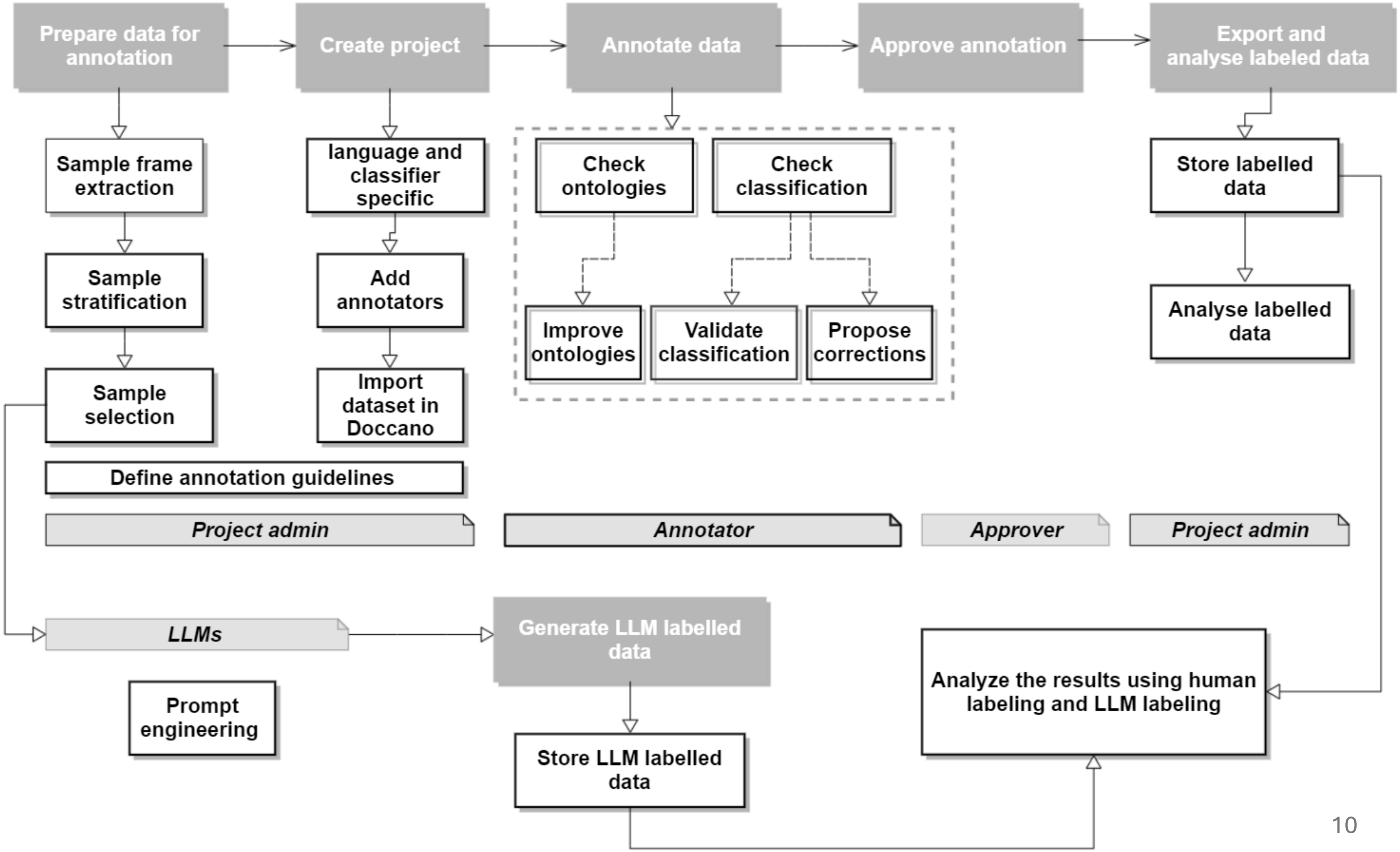
GPT-4 is better than $25/hr humans (arxiv.org/abs/2304.03279)

ChatGPT just outperformed Mechanical Turk workers on text annotation tasks

per-annotation cost of ChatGPT is less than $0.003 ≈ **20 times cheaper than MTurk**

potential of LLM to drastically increase the efficiency of text classification (https://arxiv.org/pdf/2303.15056v1.pdf)

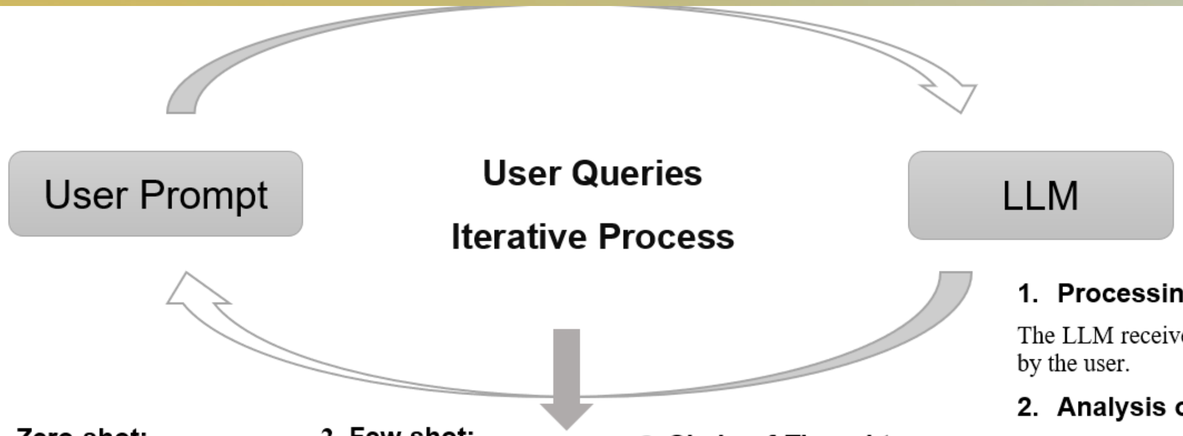
Workflow to collect labelled data for OJA



How to collect LLM labelled data ?

1. **Define the labeling task:**
   - Classify job ad for OCCUPATION at 4th level ISCO-08
   - ≈ 400 classes ISCO-08 4D
2. **Prepare prompt templates:**
   - Guide LLM to provide the desired labels: clear, concise, and provide sufficient context to make accurate judgments
3. **Dataset for annotation:**
   - OJA sample labelled by experts
4. **Interface with the LLM** (chatGPT-4):
   - Generate labels for the OJA sample
5. **Quality control:**
   - Implement quality control measures
   - Regularly checks on the annotations (use of separate prompts or Agents)

Prompting LLM



1. **Zero-shot:**

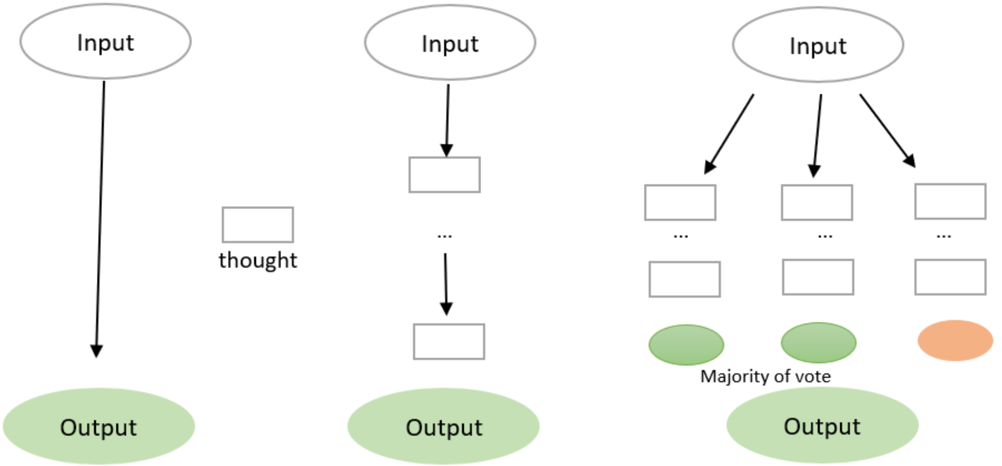
The user provides one (a series of) job description(s) and asks for the corresponding ISCO-08 codes at the 4-digit level.

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.

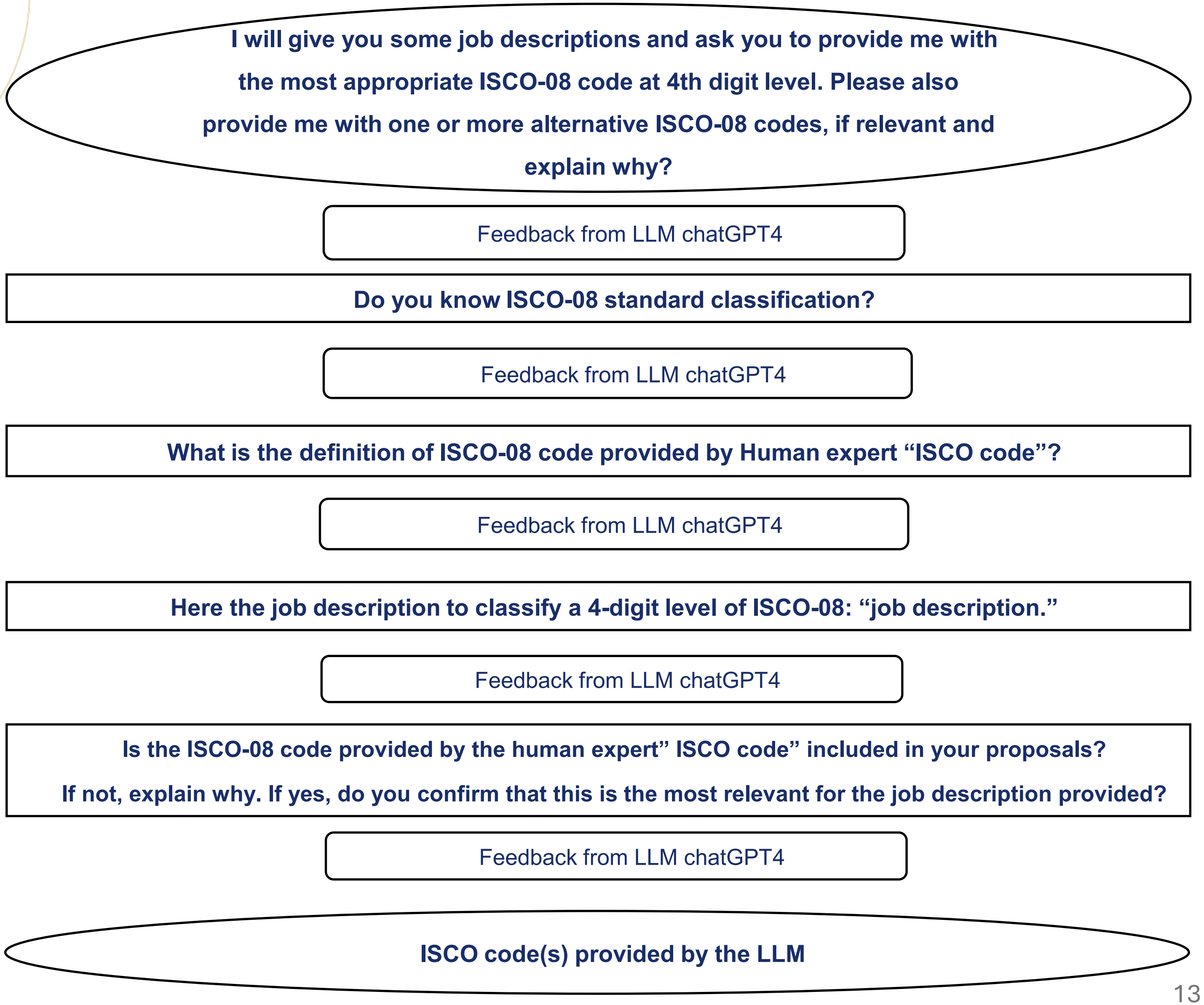
4. **Feedback to User:**

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.

EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Example of prompting



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

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

Agreement between Human expert & LLM (ChatGPT-4)

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 ISCO-08 4D 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 labelled sample (human expert): ‘*job description missing*’, ‘*wrong language*’, ‘*not a job ad*’

Agreement rate: human expert, OJA classifier and LLM

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 %

First analysis of results

***'Correct'* and *'Incorrect'*:**

- substantial difference suggests a disparity in the accuracy of labelling between the human expert and the LLM (ChatGPT-4).

- human expert's judgments may have been influenced by their awareness of the OJA classifier's results.

- 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.

***'Multiple ISCO-8 4D labels'* for the same job ad:**

Both the human expert and the LLM (ChatGPT-4):

- encountered cases where multiple ISCO4D labels were assigned.

- faced challenges in accurately classifying certain ads with multiple job categories.

Potential Bias in Human Expert Judgments:

- impacted the accuracy and consistency of the human expert-labelled data.

- LLM classification was not affected by this bias since the result of the OJA classifier was not provided in the prompting.

Conclusions

- Even if the LLM and human expert have low agreement rate, the LLM seem to be more accurate, after checking the explanations provided when proposing the ISCO label (for a sub-sample of the OJA dataset labeled)

Next steps

- Test the agreement response using more advanced prompting techniques (Tree-of-thought) to confirm the accuracy of LLMs in classifying occupation based on job description

Stay connected

Anca-Maria.KISS@ext.ec.europa.eu

ESTAT-WIH@ec.europa.eu

https://ec.europa.eu/eurostat/web/main/home

Thank you!