FOURTH WORKSHOP
ON METHODOLOGIES
FOR OFFICIAL STATISTICS

Language Models for Automated Coding in Official Statistics: Risks and Opportunities

Fourth Workshop on Methodologies for Official Statistics
Rome, Italy, 1-2 December 2025

Mauro Bruno, Marco Di Zio, Francesco Ortame (Istat)

December 1, 2025

Language Models for Automated Coding

1 Context

- Context
- Data
- Methodology
- Results

Textual Data in Official Statistics

1 Context

Modern statistical production increasingly relies on **textual data** that are **not generated under statistical control** and vary widely in structure and quality.

- **Administrative data:** balance sheets, patents, and other free-text descriptions often require assigning **classification codes** (e.g., ATECO/NACE, ISCO). Traditional rule-based systems are accurate but **costly to maintain** and difficult to scale across taxonomies.
- **User-provided descriptions:** services like CIRCE rely on manually updated thesauri and string-matching techniques to support coding activities. The transition to **ATECO 2025** would require substantial revisions.
- **Company websites:** web scraping combined with **Retrieval Augmented Generation (RAG)** workflows allow the estimate of enterprises' products.

Why Modern NLP Helps

1 Context

Language models and sentence embeddings allow NSOs to move from **token-based** matching to **meaning-based retrieval**.

- **Semantic understanding:** semantic search and sentence embeddings enable matching based on **meaning**, not token overlap, offering a more flexible alternative for automatic coding tasks.
- **Flexibility across taxonomies:** embedding-based systems can be adapted to new versions of classifications (e.g., **ATECO 2025**) with substantially lower maintenance than rule-based thesauri.
- **Probabilistic outputs:** NLP-based classifiers (e.g., LDA on embeddings) provide **uncertainty measures** that can be calibrated through statistical techniques such as conformal prediction.

Research Question

1 Context

In this work, we explore whether **embedding-based methods** can support the assignment of **ATECO codes** to user-provided descriptions of economic activities.

1. We use **sentence embeddings** to represent both activity descriptions and ATECO textual descriptors in a shared semantic space.
2. We evaluate two retrieval/classification strategies: **cosine similarity** (semantic search) and **linear discriminant analysis**.
3. We assess model outputs against **CIRCE-assigned labels**, interpreting accuracy as the degree of concordance with the existing rule-based system.
4. We quantify uncertainty through **conformal prediction** to derive statistically valid prediction sets for each query.

Language Models for Automated Coding

2 Data

- Context
- Data
- Methodology
- Results

Datasets

2 Data

We rely on three complementary datasets that together provide a comprehensive view of economic activity descriptions and their corresponding ATECO codes.

Dataset	Source	Size / Scope	Purpose in study
ATECO 2022	Istat, official classification	21 Sections, 87 Divisions, ∼1,200 Sub-categories	Hierarchical reference and base for data augmentation
CIRCE-labelled queries	Istat, classification tool	33,544 user-generated queries	Benchmark for evaluating semantic vs. rule-based classification
Synthetic dataset	GPT-5-mini (OpenAI, 2025)	12,400 generated examples	Data augmentation, balancing and knowledge base creation

These datasets include **33.544** real user queries and **12.400** synthetic examples covering all ATECO divisions.

ATECO 2022: Official Classification

2 Data

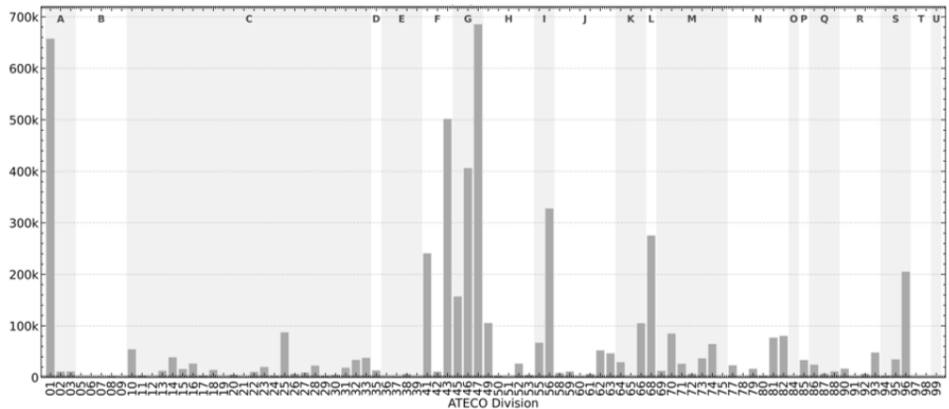
ATECO 2022 is the official classification of economic activities used in Istat.

- **Hierarchical structure:** 21 Sections, 87 Divisions, and more than 1.200 Sub-categories (6-digit level).
- **Rich textual descriptors:** titles, explanatory notes, inclusion and exclusion rules define the scope of each code.
- **Imbalanced distribution:** some divisions (e.g., retail trade) contain hundreds of thousands of enterprises, while others have only a few cases.
- **Role in this study:** provides the **taxonomy** and textual material used to build the knowledge base and to generate *synthetic descriptions*.

ATECO 2022: Official Classification

2 Data

Figure 2.1 - Distribution of active businesses in Italy by ATECO division (2-digit code) as of December 2024. Shaded background areas and capital letters indicate ATECO sections.



CIRCE-labelled Queries

2 Data

CIRCE is Istat's **rule-based** tool that helps enterprises identify the appropriate ATECO category from short text descriptions.

- Publicly available through Istat's classification portal since 2016.
- The dataset used here consists of **33.544** user-generated queries, each mapped by CIRCE to a **5-digit** ATECO code.
- Queries are short free-text descriptions of economic activities, reflecting **real interactions** with the classification tool.
- **Role in this study:** serves as a **benchmark** to evaluate the agreement between embedding-based methods and the existing rule-based system (concordance with CIRCE labels).

Synthetic Dataset Generated with GPT-5-mini

2 Data

To describe ATECO divisions in the form of user queries, we generate a synthetic datase tof activity descriptions.

- Starting point: official ATECO 2022 descriptors (titles, notes, inclusion/exclusion rules) at the 6-digit level.
- For each sub-category, GPT-5-mini is prompted to create **10 short queries** that mimic real user descriptions.
- This yields **12.400** synthetic examples, which are then aggregated at the **2-digit** division level.
- **Role in this study:** used for constructing representative embeddings for each **ATECO division** in the knowledge base.

Synthetic Dataset Generated with GPT-5-mini

2 Data

You are a generator of descriptions of economic activities carried out by businesses and professionals.

Here is what you need to do:

- You receive an economic classification as input: title + details (included and excluded activities, classification hierarchy. activities, classification hierarchy.
- You generate 10 examples of very short descriptions of activities carried out by enterprises, coherent with the classification.
- Focus on the specified activity. If present, be careful with the excluded activities.
- You do not generate text that is not one of the descriptions.

Important information: keep in mind that the descriptions that you generate serve as synthetic queries for a search engine, in which an enterprise/professional is asked to describe their activity with the aim of assigning a classification.

Here are some examples:

[8 examples of real queries here]

Format your output as a Python list.

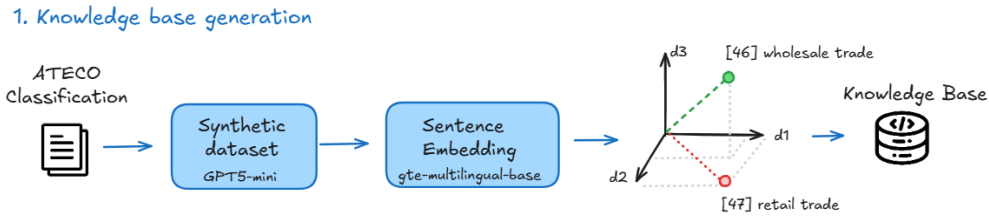
Language Models for Automated Coding

3 Methodology

- Context
- Data
- Methodology
- Results

1. Knowledge Base Generation

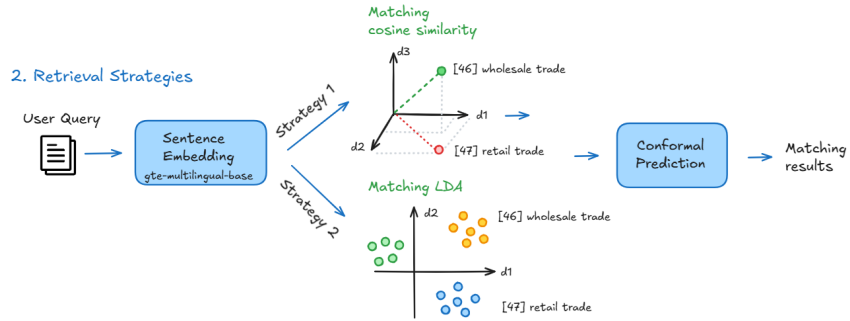
3 Methodology



The ATECO taxonomy is expanded using a synthetic dataset generated with GPT-5-mini. Each description is embedded with a *multilingual sentence encoder* to build a semantic **knowledge base** for all divisions.

2. Retrieval and Uncertainty Quantification

3 Methodology



User queries are embedded and matched to ATECO divisions using **cosine similarity** or **Linear Discriminant Analysis**. Conformal prediction calibrates a similarity threshold to produce prediction sets with valid statistical coverage.

Sentence Embeddings

3 Methodology

Sentence encoders map texts into a shared **high-dimensional semantic space**. This allows us to compare user queries and ATECO descriptors based on **semantic meaning**.

- We use the **gte-multilingual-base** sentence encoder to obtain 768-dimensional vectors for each description.
- Each ATECO division is represented by **multiple embedding vectors**, one for each synthetic example in the knowledge base.
- The resulting embedding space provides a compact semantic **fingerprint** of each economic activity.

This representation enables meaning-based retrieval rather than token matching.

Retrieval Strategies

We evaluate two distinct approaches for matching user queries to ATECO divisions, leveraging the semantic structure of the embedding space.

- **Cosine similarity (semantic search):** retrieves the divisions whose embedding vectors are most closely aligned with the query vector in the high-dimensional space.
- **Linear Discriminant Analysis (LDA):** fits a probabilistic classifier on embedding vectors to obtain **posterior probabilities** for each division.

Performance is assessed in terms of **Accuracy@k**, measuring agreement between retrieved divisions and CIRCE-assigned labels.

Conformal Prediction

3 Methodology

Conformal prediction provides **statistically valid prediction sets** based on cosine similarities or LDA probabilities, allowing explicit uncertainty quantification for retrieval-based classification.

- Addresses the lack of **uncertainty measures** in embedding-based retrieval, which is essential in official statistics.
- **Split Conformal Prediction (SCP):** computes calibrated thresholds on nonconformity scores to produce prediction sets with a target coverage (e.g. 90%).
- **Adaptive Prediction Sets (APS):** builds prediction sets by accumulating probability mass until the coverage level is reached, producing sets that adapt to query difficulty.

Language Models for Automated Coding

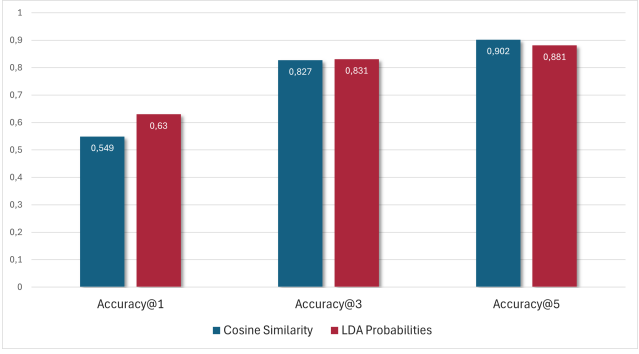
4 Results

- Context
- Data
- Methodology
- Results

Performance of Retrieval Strategies

4 Results

We evaluate the performance of the retrieval strategies: cosine similarity and LDA

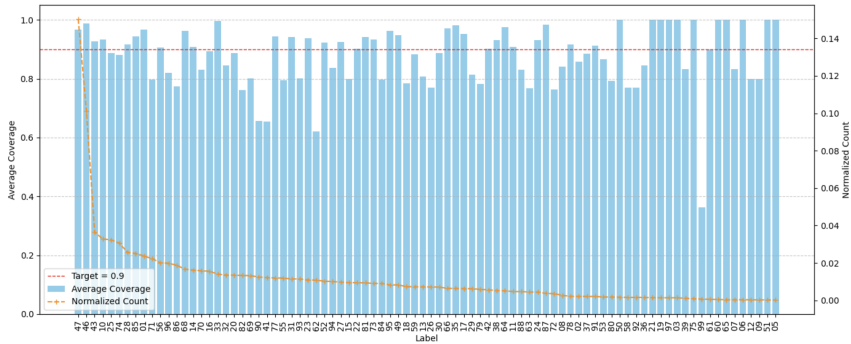


Both approaches achieve high agreement with CIRCE when we look at the top–5 retrieved divisions — around 90%.

Label-Stratified Coverage

4 Results

Label stratified coverage (LSC), with cosine similarity-based retrieval and standard split conformal prediction.



High coverage is achieved at the cost of relatively large prediction sets (average size ≈ 8).

Conclusions

4 Results

- We explored the use of language models for assigning ATECO codes to user-provided descriptions using synthetic data generation and sentence embeddings.
- Retrieval strategies based on **cosine similarity** and **LDA** show encouraging results, with agreement above 90% in top–5 compared to the rule-based CIRCE system.
- Conformal prediction provides **valid uncertainty quantification** for both retrieval approaches, ensuring calibrated coverage.

Limitations

- Achieving calibrated coverage leads to **large prediction sets** (average size ≈ 8), which reflects both model calibration and intrinsic query ambiguity.
- Further refinement is needed in retrieval strategies and discriminant models.