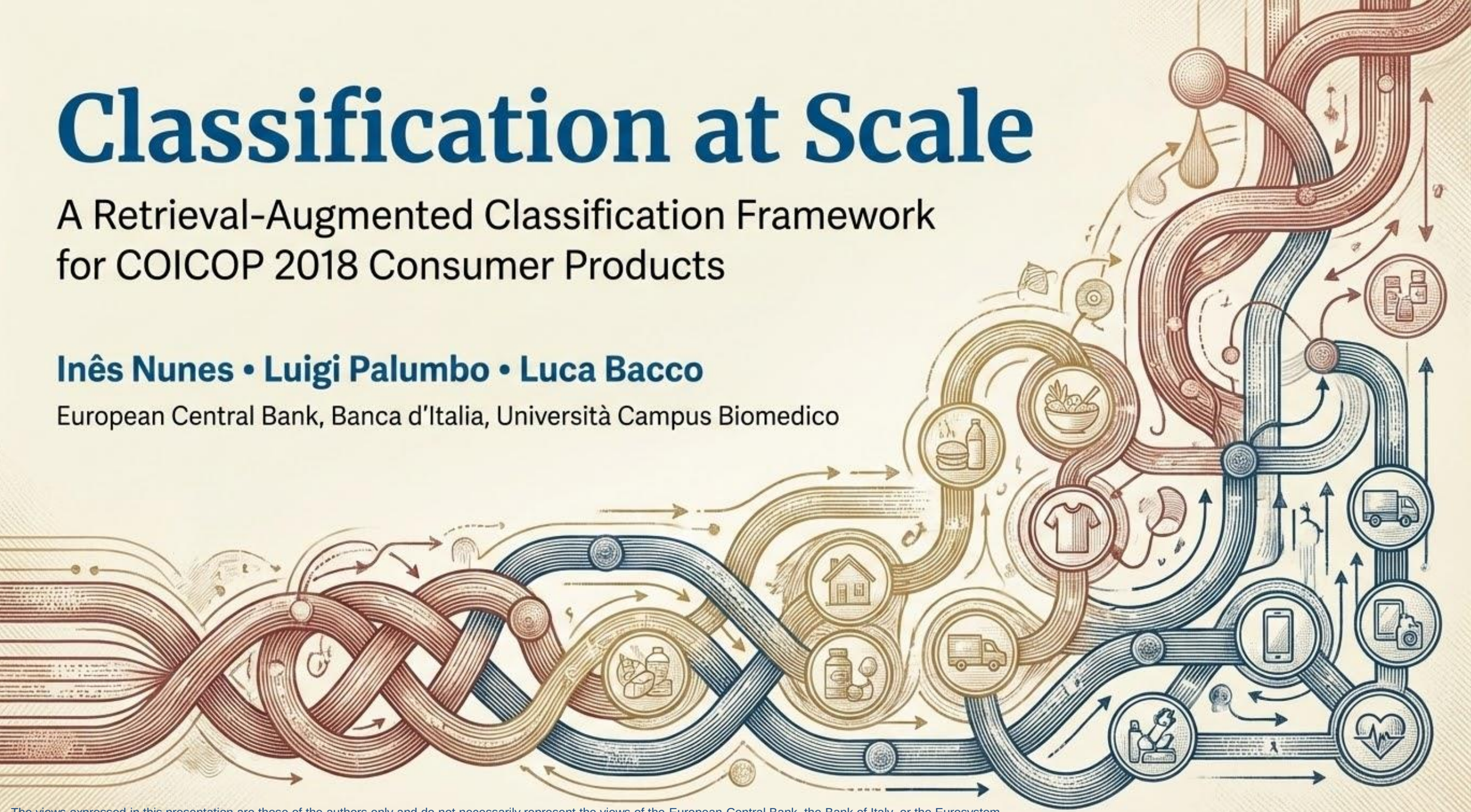
Classification at Scale

A Retrieval-Augmented Classification Framework for COICOP 2018 Consumer Products

Inês Nunes • Luigi Palumbo • Luca Bacco

European Central Bank, Banca d'Italia, Università Campus Biomedico

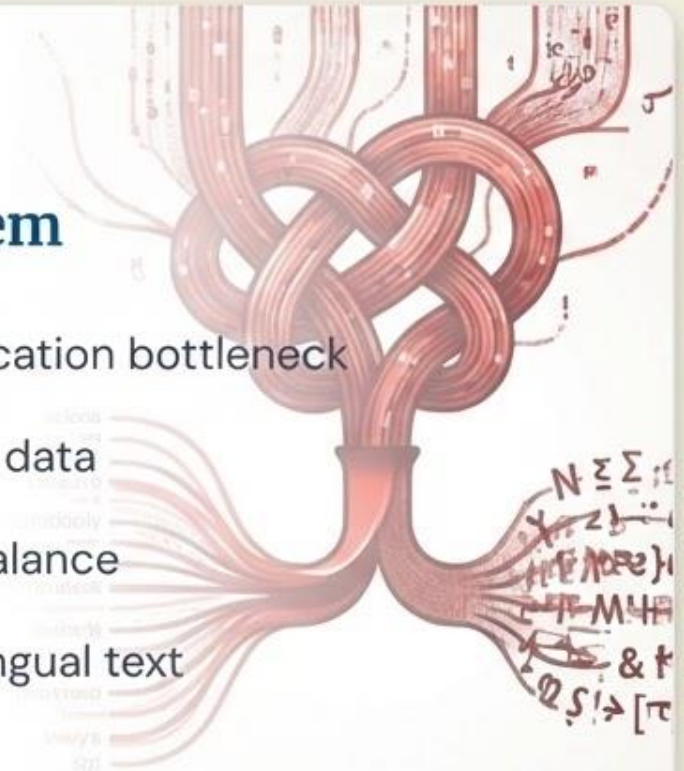


Context & Challenge

The Need

- Automated CPI compilation
- Inflation nowcasting
- Exploiting web-scraped data
- High-frequency microdata analysis

The Problem



- Manual classification bottleneck
- Scarce labeled data
- High class imbalance
- Noisy & multilingual text

The RAC Framework

Synthetic Knowledge Base

Derived from official COICOP definitions

Semantic Retrieval

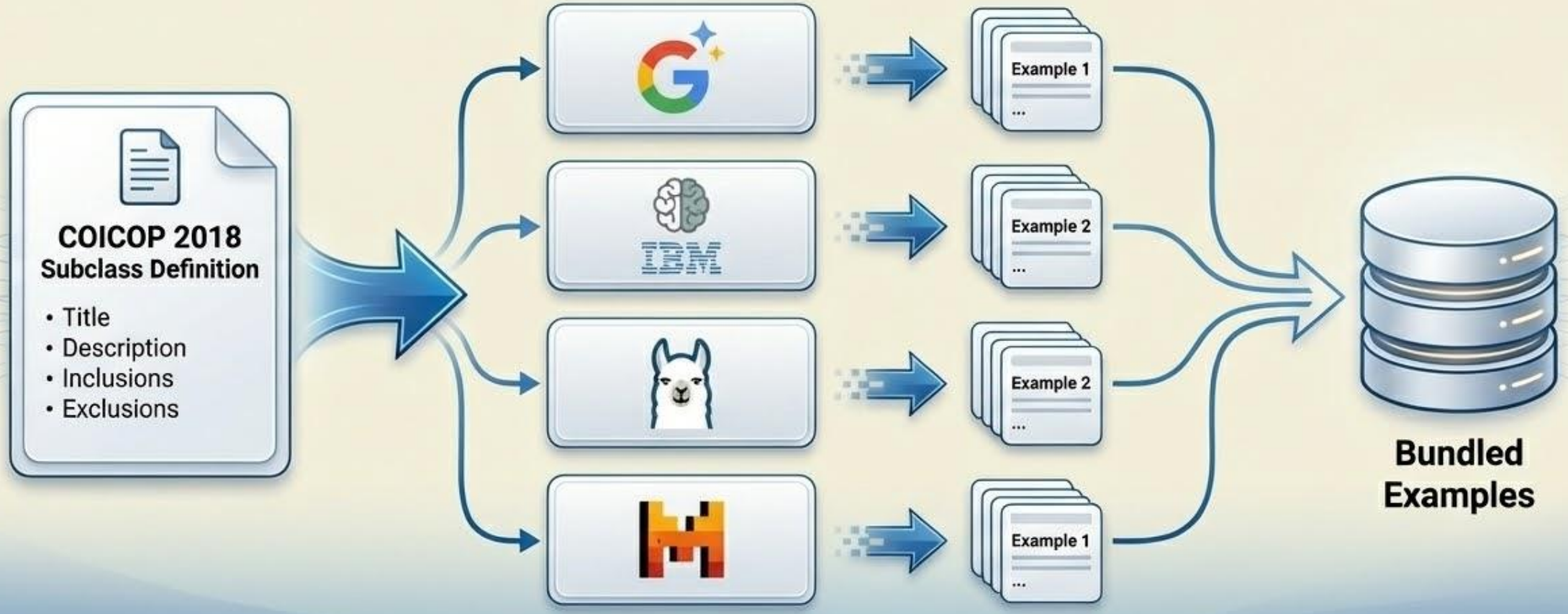
Identifying relevant exemplars

Generative Reasoning

Zero-shot classification

Synthetic Data Creation Process: From Definitions to Bundled Examples

- **Source:** COICOP 2018 Taxonomy
- **Scale:** 34,451 multilingual exemplars across 294 COICOP subclasses

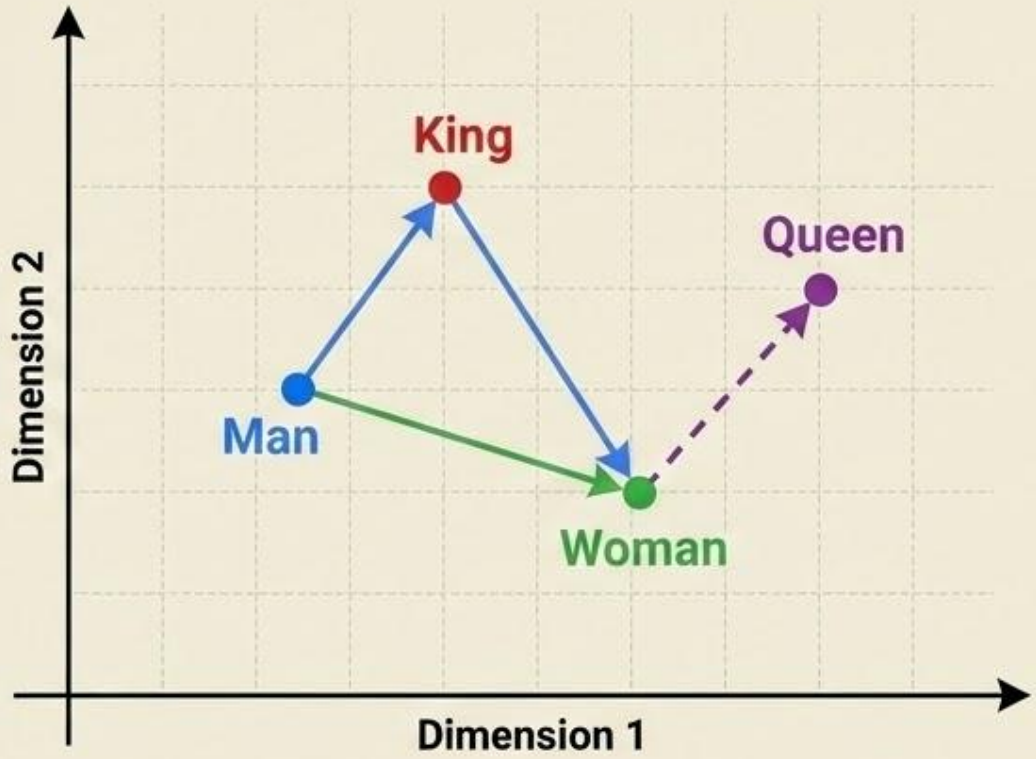


Semantic Similarity

Vector Space: Words mapped to numerical vectors

Principle: Geometric distance = Semantic meaning

Algebraic Operations: Relationship preservation



Vector(King) − Vector(Man) + Vector(Woman) ≈ Vector(Queen)

k-Nearest Neighbours Classification

A non-parametric classification method based on the proximity

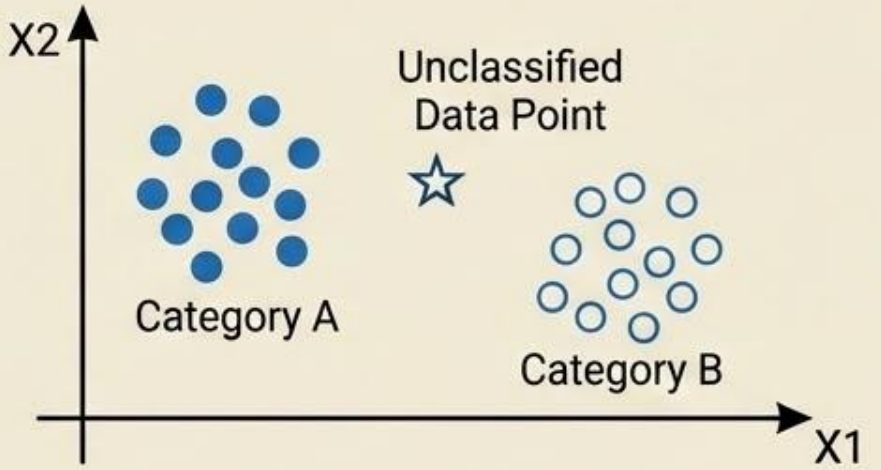
Baseline Method: Non-parametric approach, no explicit model training

Metric: Semantic Similarity (Cosine/Dot Product)

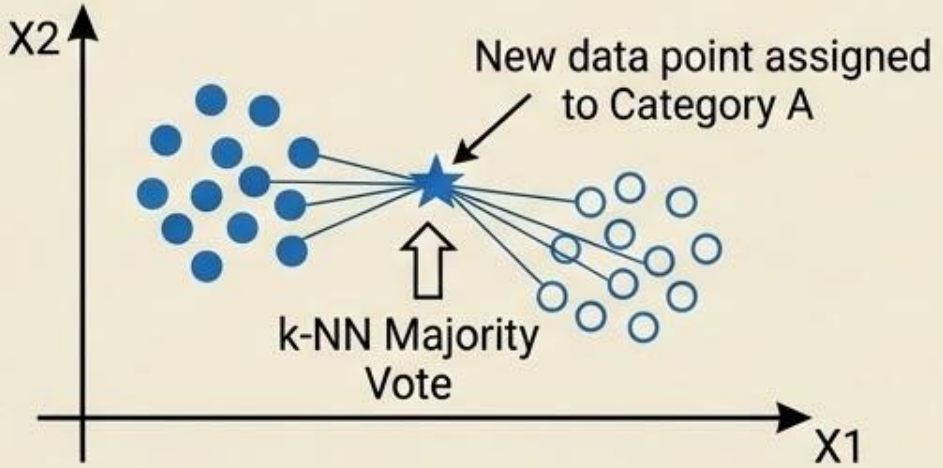
Mechanism: Retrieve top-k similar exemplars

Decision: Majority voting determines class

Before k-NN: Unclassified Data Point



After k-NN: Classified Data Point



RAC: Classification by Example

Item to be Classified
(e.g., Web-scraped Data)

1. Retrieve similar exemplars

Top-k similar synthetic examples found

LLM Prompt Construction

Insert examples into prompt as context

LLM Reasoning & Selection

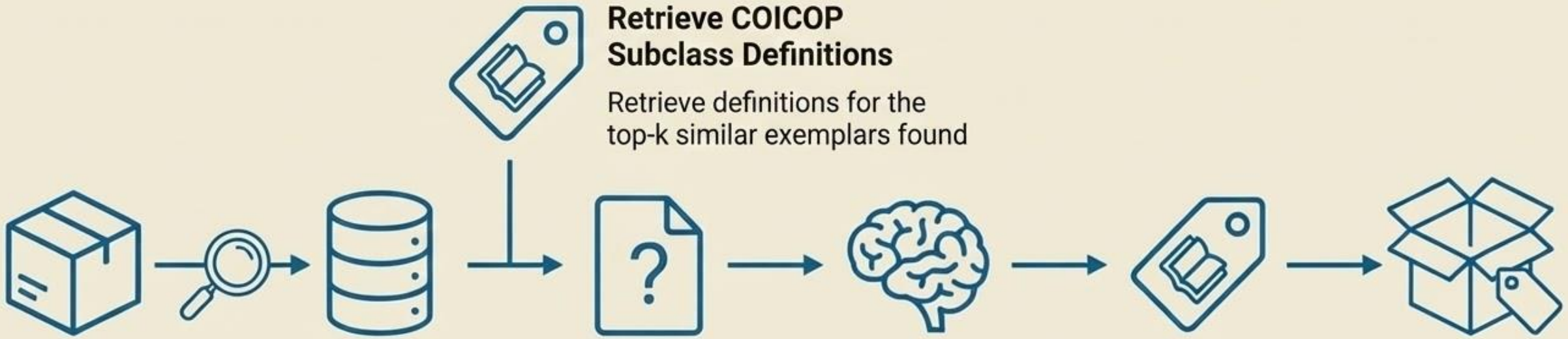
LLM selects most similar example or "none of the above"

COICOP 2018 Subclass Label

Retrieve & assign label of selected example

Classified Item

RAC: Classification by Definition



Item to be Classified
(e.g., Web-scraped Data)

1. Retrieve similar exemplars

Top-k similar synthetic examples found

Retrieve COICOP Subclass Definitions

Retrieve definitions for the top-k similar exemplars found

LLM Prompt Construction

Insert examples and their COICOP definitions into prompt as context

LLM Reasoning & Selection

LLM selects most appropriate COICOP definition or "none of the above"

COICOP 2018 Subclass Definition & Label

Assign definition & label of selected example

Classified Item

Evaluation Framework

Gold Standard Data

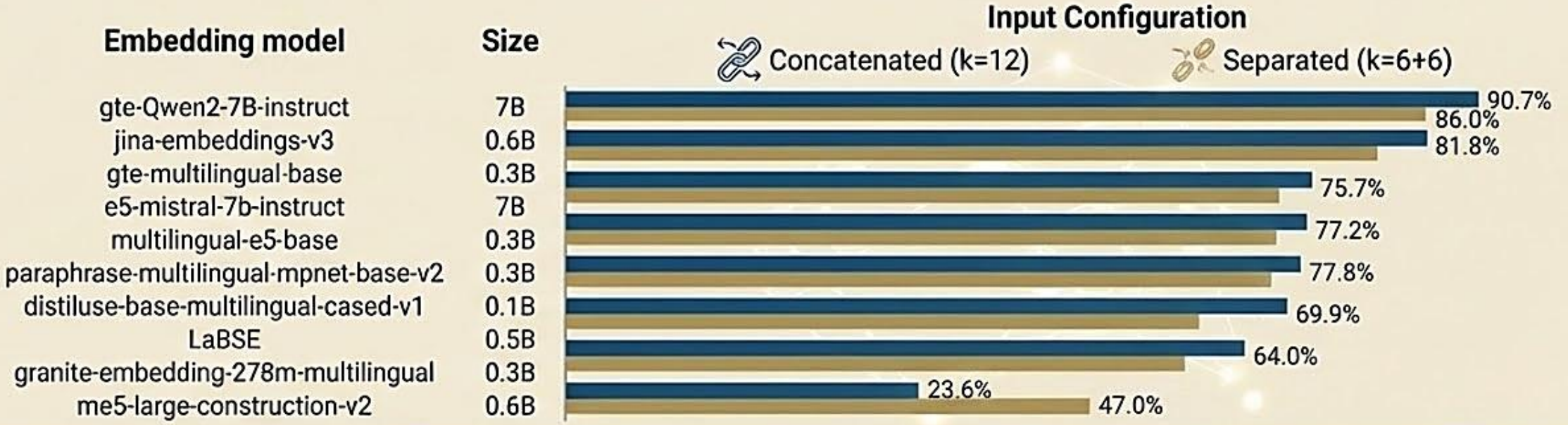
- 4,993 manually labeled products
- 61 COICOP subclasses
- Diverse markets: Italy, Malta, Germany
- High class imbalance preserved

Model Configuration

- **Embeddings:** GTE, E5, Jina, ...
- **Retrieval:** Concatenated vs Separated
- **LLMs:** Qwen 2.5, Llama 3, DeepSeek R1, ...
- **Metrics:** Overall Accuracy @ Level 4, 3, and 2; Accuracy @ Level 4, 3 and 2 on given labels.

Key Findings: Retrieval

Model Performance Comparison



Key Insights

Concatenated input consistently outperforms separated input.

Performance is strongly dependent on the embedding model, but not necessarily on model size.

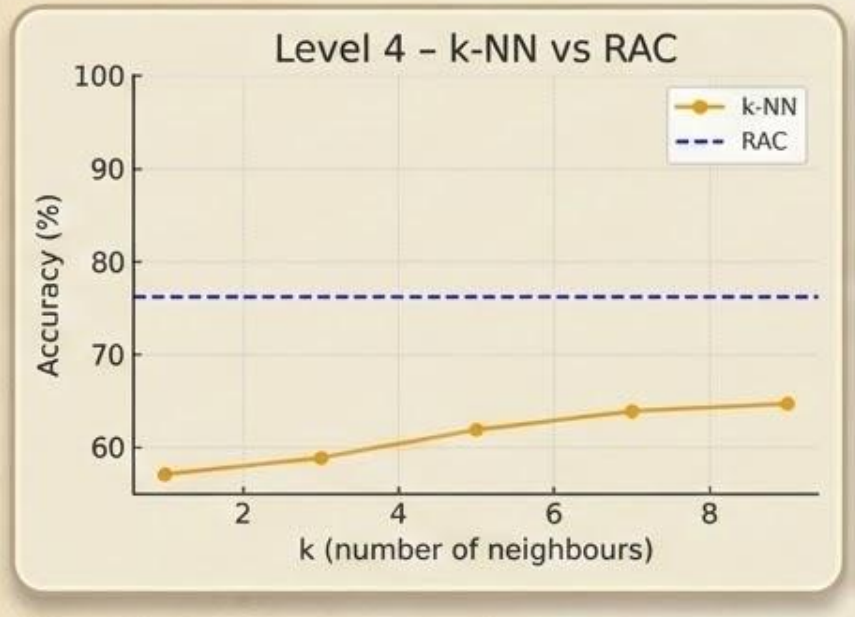
Key Findings: RAC by example

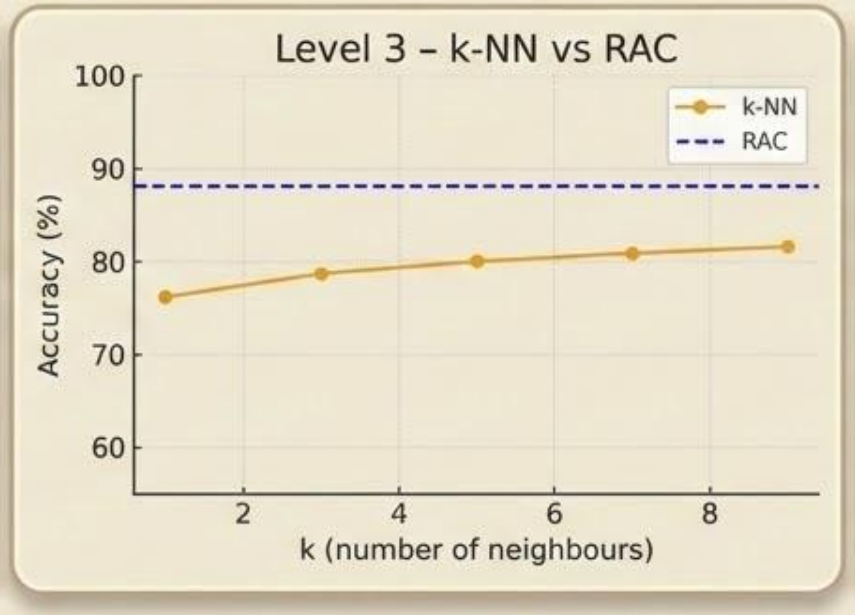
Language model	# classified	Overall			Correct	Classified		
		L4 (%)	L3 (%)	L2 (%)	opt-out (%)	L4 (%)	L3 (%)	L2 (%)
mistral-latest	4856	**63.8**	80.5	93.1	6.65	65.6	82.8	95.7
llama3-8b	4912	62.9	**81.4**	**94.5**	3.86	64.0	82.7	96.1
llama3.3-70b	4317	59.9	74.1	83.5	35.0	69.2	85.7	96.6
gemma3-4b	3940	49.6	64.1	74.7	29.2	62.9	81.3	94.7
qwen2.5-32b	3245	48.7	57.2	63.2	**65.5**	**74.9**	**88.0**	**97.2**
qwen2.5-3b-instruct	3830	46.1	60.1	72.0	40.6	60.1	78.3	93.8
llama3.2-1b	3916	41.6	56.5	72.4	22.5	53.0	72.0	92.3
granite3.1-moe-3b	3497	35.9	50.4	64.4	25.8	51.3	71.9	91.9
qwen2.5-1.5b-instruct	3082	34.4	45.0	57.0	47.2	55.7	73.0	92.4
deepseek-r1-70b	1279	19.0	22.6	24.9	85.8	74.1	88.2	97.3
deepseek-r1-8b	1299	14.7	19.8	24.3	84.1	56.5	76.1	93.5
deepseek-r1-1.5b	406	4.63	5.63	7.11	94.0	56.9	69.2	87.4
smollm2-135m	156	1.74	2.30	2.98	97.6	55.8	73.7	95.5

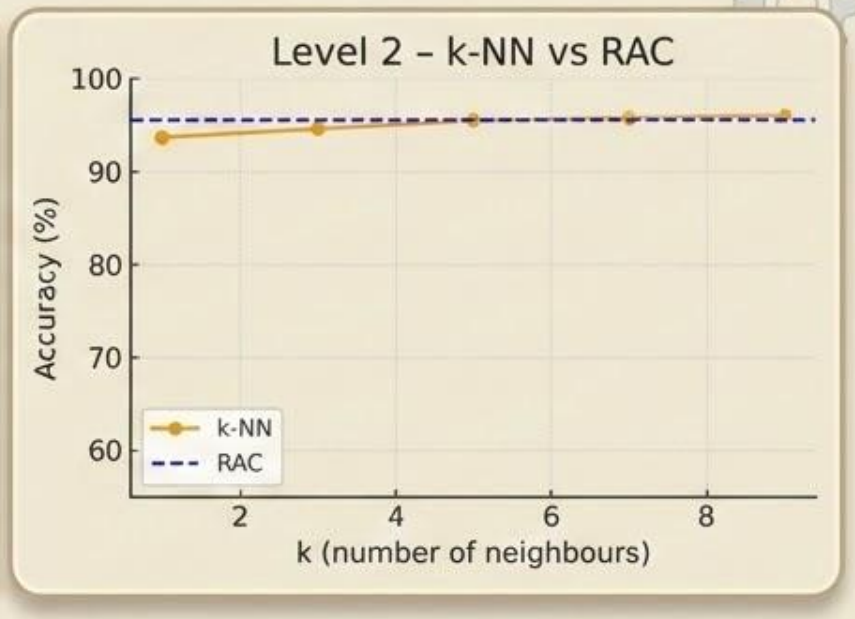
Key Findings: RAC by definition

Language model	# classified	Overall			Correct	Classified		
		L4 (%)	L3 (%)	L2 (%)	opt-out (%)	L4 (%)	L3 (%)	L2 (%)
qwen2.5-32b	4848	**76.2**	**88.1**	**95.5**	15.0	78.5	90.7	98.4
llama3.3-70b	4760	75.0	86.5	94.1	22.3	78.6	90.8	98.7
deepseek-r1-70b	4341	70.1	80.1	86.0	**40.1**	**80.6**	**92.2**	**98.9**
llama3-8b	4904	65.6	84.8	95.3	7.08	66.8	86.3	97.1
mistral-latest	4934	65.0	85.0	96.2	3.00	65.8	86.0	97.3
gemma3-4b	4818	61.8	82.4	93.8	11.2	64.1	85.4	97.2
qwen2.5-3b-instruct	4336	59.3	73.2	84.2	29.0	68.3	84.2	96.9
granite3.1-moe-3b	3919	43.7	60.7	72.3	33.0	55.7	77.3	92.1
deepseek-r1-8b	2890	29.6	44.7	54.6	53.2	51.1	77.2	94.3
qwen2.5-1.5b-instruct	2604	26.1	39.6	48.7	55.2	50.0	75.9	93.4
llama3.2-1b	3795	25.3	48.5	65.9	30.9	33.2	63.8	86.7
deepseek-r1-1.5b	1111	4.13	10.1	17.9	68.5	18.5	45.3	80.6
smollm2-135m	1250	3.36	10.4	16.9	69.1	13.4	41.6	67.7

RAC by definition vs kNN







Much better performance at the more detailed levels of aggregation

Impact & Conclusion

Scalable pipeline for official statistics

Real-time inflation nowcasting

Fully interpretable decision traces

Multilingual & taxonomy-agnostic

Thank you!

 Q&A

Luigi Palumbo – Luigi.Palumbo@bancaditalia.it