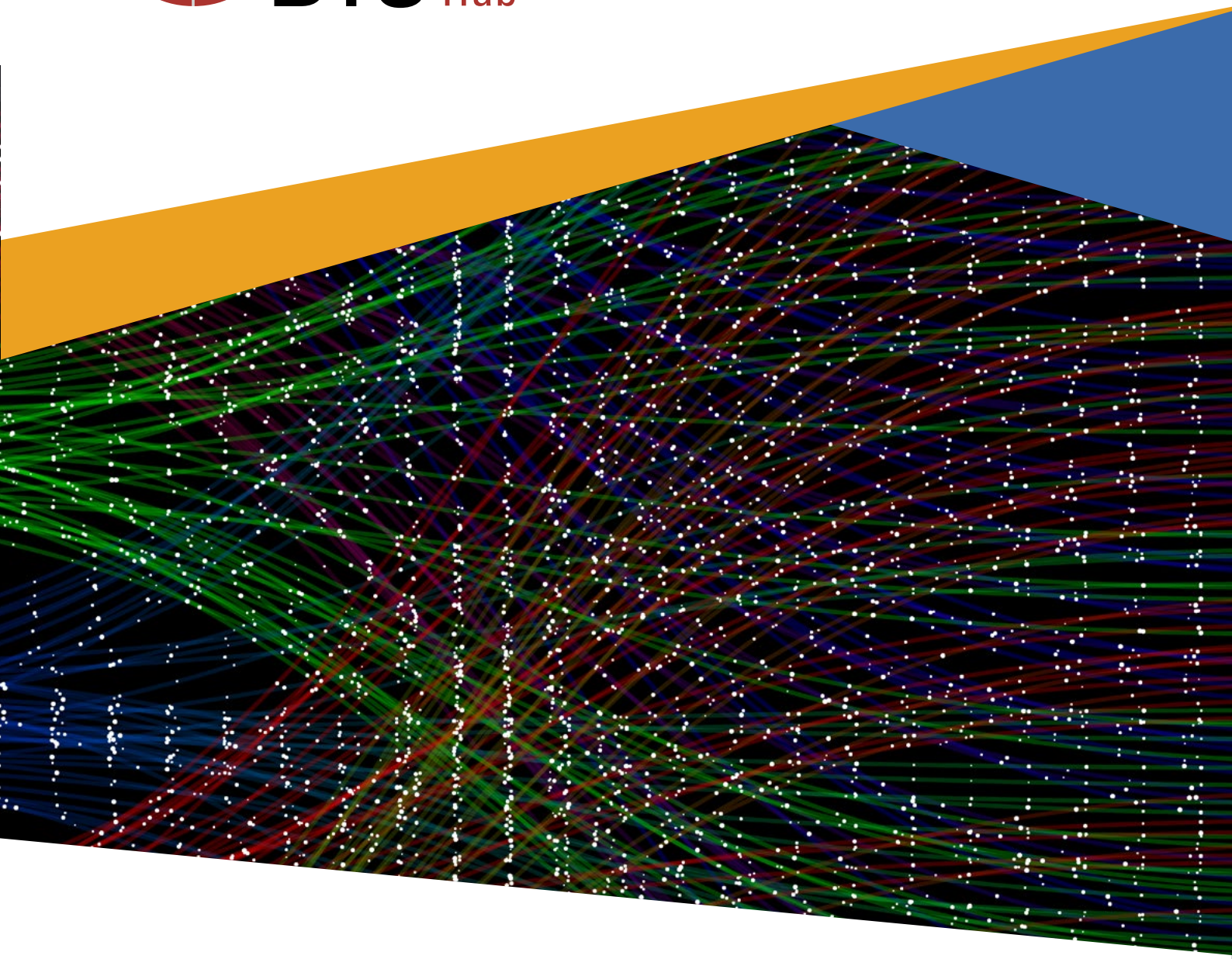
BIS Innovation Hub



Project Spectrum

Using generative AI to enchance inflation nowcasting

February 2026

DEUTSCHE BUNDESBANK EUROSYSTEM

EUROPEAN CENTRAL BANK

Follow us

Executive summary

The availability of web-scraped and scanner data sets provides central banks with unprecedented access to real-time data on individual product prices. However, to use these data for inflation nowcasting and forecasting, analysts need to classify products according to statistical conventions. In the absence of reliable, scalable classification methods, inflation analysts are flooded with data but lack actionable insight.

Product classification at the scale of web-scraped data represents a major challenge. Manually processing this amount of data is not feasible. Classification using large language models (LLMs) is promising, but with the LLM models currently available, the processing time and cost become prohibitively high. Project Spectrum used the European Central Bank's Daily Price Dataset (DPD), which contains billions of price-product daily observations for 34 million unique products. At the time of writing, classifying this data set using GPT-5 would take over six months of computing time at a cost exceeding EUR 0.5 million.

Project Spectrum – a collaboration between the Bank for International Settlements (BIS), the Deutsche Bundesbank and the European Central Bank – explored an alternative approach where artificial intelligence (AI) was used only to transform product descriptions into high-dimensional text embeddings. These were then classified into product categories using classic machine learning algorithms. Text embedding is a foundational AI technique used by many natural language processing applications. This method achieved accuracy levels comparable to LLM prompting, but at a fraction of the cost: the entire DPD was classified in just five days for approximately EUR 1,500.

Besides classifying all records in the current DPD, the project has developed a production pipeline solution that can classify new products as they are added to the DPD. In addition, to ensure continuous improvement, an iterative algorithm was implemented to gradually expand the reference data set. By selectively adding manually labelled data to the reference and validation sets, this algorithm systematically refines the classification logic, enhances overall predictive accuracy and adapts to a changing product range.

By turning raw, fragmented product descriptions into structured data, Project Spectrum equips analysts and policymakers with timely, detailed insights into price developments. Ultimately, the project contributes to an emerging new generation of AI-powered analysis, where data abundance can be translated more easily into actionable economic understanding.

This report is intended for monetary policy analysts who utilise high-frequency data for inflation nowcasting and data scientists within central banks looking for cost-effective alternatives to LLMs for large-scale classification. It also serves as a technical reference for statistical agencies seeking to automate the categorisation of scanner and web-scraped data into official indices. Finally, it provides a methodological framework for economic researchers studying price-setting behaviour at the individual product level.

Acronyms and abbreviations

AI	Artificial intelligence
CPI	Consumer price index
COICOP	Classification of Individual Consumption by Purpose
DPD	Daily Price Dataset
ECB	European Central Bank
ECOICOP	European Classification of Individual Consumption by Purpose
FFN	Feedforward neural network
GPT	Generative pre-trained transformer
KNN	K-nearest neighbours
LLM	Large language model
MIRACL	Multilingual information retrieval across a continuum of languages
MTEB	Massive text embedding benchmark
PRISMA	Price-setting Microdata Analysis Network

Table of contents

1. Using online price data for inflation nowcasting

Accurate and timely inflation nowcasts[1] and forecasts are central to effective monetary policy because inflation often responds to policy measures with a time lag.[2] By anticipating future price trends, policymakers can adjust interest rates and other policy tools earlier to maintain price stability and prevent costly economic fluctuations. Such foresight not only helps anchor inflation expectations and foster sustainable growth,[3] but also reinforces public confidence in the monetary authority's ability to respond promptly and effectively to emerging risks.

As a significant share of consumer spending is done online, data from e-commerce platforms are a rich source of real-time information for inflation nowcasting and price-setting analysis. In traditional statistical sampling and in-store data collection, field officers visit stores and document prices at periodic intervals. As online data capture real-time price variations, they have the potential to provide more timely and targeted insights than traditional macroeconomic indicators, which are often available with some delay and at regular publication frequencies.[4] The data also provide access to a richer set of information about products and price setting, including the frequency and size of individual price changes, price setting by large firms, shop-level data on the use of sales and discounts and the evolution of the range of offered products.[5]

However, while online price data are rich in detail, extracting actionable insights remains a challenge. The main issue is that web-scraped data are not standardised – they come in various formats, lack quality adjustments and contain key information in non-standardised textual form, making interpretation challenging.[6] This heterogeneity necessitates extensive harmonisation efforts to create a unified data set.

The principal challenge Project Spectrum addresses is the labelling of products in accordance with international classification standards, such as the Classification of Individual Consumption According to Purpose (COICOP). To harness information on individual prices for inflation analysis, it is crucial to map each product to a classification category. This enables the construction of consistent price series at a granular level. Ensuring accurate classification is also essential for maintaining comparability in inflation calculations across countries.

The advantages of improving the quality and accessibility of high-frequency price data are clear. At the same time, the challenges of creating a structured, unified data set from high-frequency data are manifold. In addition to the classification issue, the sheer scale of the collected data exceeds the capacity of traditional data-processing methods and requires extensive automation. While typical data sets used for economic policy contain hundreds or thousands of data points, web-scraped online price data contain millions, even billions of observations. Processing such massive amounts of data requires new skills and adjustments in the technical infrastructure for efficient storage and processing. Only when these data are structured and unified can central banks turn them into insights for decision-making, forecasting and economic analysis.

1. Nowcast refers to an estimate of the (quasi) real-time value of an indicator before the official data are published.
2. Friedman (1961) argues that monetary policy operates with "long and variable lags", making accurate inflation forecasts crucial for effective policy decisions. Sims (1992) uses vector autoregressions to analyse monetary policy transmission, showing that inflation reacts to policy changes with a delay.
3. For example, Woodford (2003) highlights the importance of forward-looking monetary policy in anchoring expectations and maintaining economic stability.
4. In the case of inflation, the first "flash" release on headline consumer price index (CPI) developments is available at the end of the month, and the details on CPI components are available a few weeks afterwards, depending on the Statistical Office. Therefore, web-scraped data provide information on the current price developments at least three weeks before the monthly official release.
5. For example, Alvarez-Blaser et al (2025) demonstrate the importance of large firms for the variability of aggregate inflation and the promise of structured big price data sets in refining inflation nowcasting.
6. Another challenge, not addressed by Project Spectrum, is the complexity and cost of collecting and maintaining web-scraped data.

To address these challenges, Project Spectrum used text embeddings and machine learning to automatically categorise web-scraped product descriptions, thereby enabling more timely and granular inflation analysis. The technological approach was validated using a large web-scraped price database: the DPD, which is collected by the European Central Bank in the context of the Price-setting Microdata Analysis Network (PRISMA) (see Osbat et al (2022)). The DPD is one of the most ambitious initiatives to collect price data by means of web scraping within the euro area. It collects approximately 4 million price-product observations per day. Due to high product churn, nearly a million new products that need to be classified may be added each month. Using this large data set allowed Project Spectrum to verify the efficiency of its approach with respect to execution time and cost and prove its suitability for processing large-scale, high-frequency data sets.

2. Related literature

Extraordinary economic episodes such as the Great Financial Crisis and the Covid-19 pandemic underscored the need for better and more timely data on key macroeconomic indicators to provide central banks with immediate insights about current economic conditions and detect sudden changes early on. This stresses the pivotal role of the data infrastructure for contemporary economic forecasting (Tissot and de Beer (2020)). Several research papers document the benefits of using online prices for economic policy (Cavallo (2013)); and the first publications building on the Billion Prices Project,[7] such as Cavallo and Rigobon (2016), provide empirical evidence that indices constructed from web-scraped prices effectively track the dynamics and the movements in official consumer price indexes (CPIs) in various countries and time horizons. This finding, confirming the utility of micro data, has been robustly supported by later works utilising micro prices from scanner data across various contexts, such as the German market (Beck et al (2024)), and broader international studies (Alvarez-Blaser et al (2025)).

High-frequency, web-scraped micro data on prices can improve forecast accuracy for headline inflation relative to well-established benchmark econometric models. This has been explored by Cavallo (2018b) and Harchaoui and Janssen (2018). More recently, focusing on the systematic analysis of using online prices for improving inflation forecasting tools, Aparicio and Bertolotto (2020) document for 10 countries that extending models with such data outperforms traditional benchmark models for predicting changes in the CPI.

Empirical evidence shows that online price data improve forecasting and nowcasting not only for headline inflation but also for its subcomponents. This is particularly important for CPI subcomponents that are notoriously difficult to predict due to their high volatility. Macias et al (2023) compare the performance of nowcasts for inflation of food and non-alcoholic beverages in Poland using online price data against standard univariate models and central bank judgmental forecasts. They show that incorporating web-scraped data can reduce forecast errors. A key contribution of their work is highlighting the critical role of accurate classification of online price observations in nowcasting. Proper classification ensures that observed price changes genuinely reflect price dynamics within the relevant goods categories. Similarly, Beer et al (2025) focus on web-scraped prices on food and non-alcoholic beverages in Austria, demonstrating that integrating web-scraped data into time-series-based short-term forecasts significantly improves the predictability of disaggregated inflation rates.[8]

The utility of web-scraped data extends beyond forecasting and nowcasting headline inflation or its subcomponents. Online prices offer an immense value for understanding firms' price-setting behaviour, including the frequency and size of price changes. Such data facilitate the analysis of the distribution of price changes, allowing for a better understanding of inflation and helping to pin down the micro-foundations in macroeconomic models (Gautier et al (2023); Gautier et al (2024); ECB (2025); Dedola et al (2025)).

7. The Billion Prices Project was a research initiative founded in 2008 by Alberto Cavallo and Roberto Rigobon that collected and analysed high-frequency online price data from retailers around the world to measure inflation in real time. See Cavallo and Rigobon (2016).
8. The ECB Price-Setting Microdata Analysis Network (PRISMA) is actively deepening the understanding of price-setting behaviour and inflation dynamics in the European Union. In addition to collecting various micro data – including online prices, scanner data and underlying official micro prices – PRISMA is conducting substantial ongoing research, the details of which can be found at https://www.ecb.europa.eu/pub/research-networks/html/researcher_prisma.en.html.

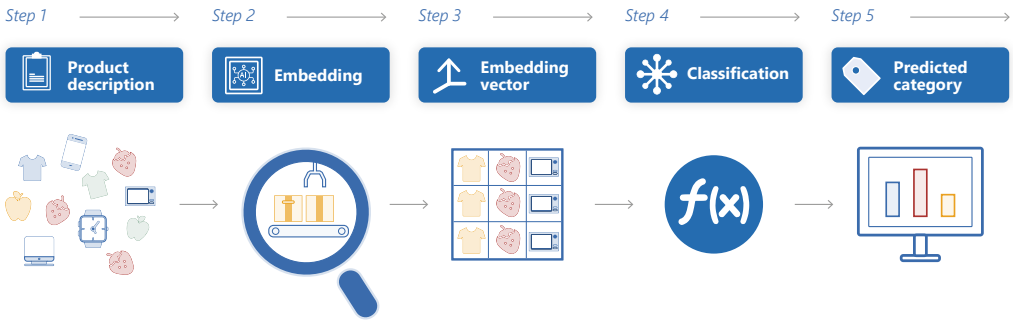
3. Spectrum overview

Project Spectrum explores the potential of artificial intelligence (AI) to categorise product descriptions to improve inflation nowcasting automatically. It uses text embeddings and large language models (LLMs) with multilingual capabilities to transform raw, high-volume, unstructured product data into standardised, accessible information. By automating product classification, Project Spectrum facilitates the integration of high-frequency, online price data into the process of inflation nowcasting and forecasting, enabling, for example, more accurate short-term inflation rate forecasts.

A straightforward generative AI approach to product classification is to provide an LLM with a product description and the definition of product categories and to prompt it to predict the appropriate category. This approach, hereafter referred to as "direct LLM prompting", is effective for a small number of products but has scalability challenges. Since the prompt needs to include the complete handbook of category descriptions, assuming a few hundred product categories, the prompt can reach around 56,000 tokens.[9] Using currently available LLM models, using this approach at scale – with millions of goods and billions of observations – becomes problematic due to the time and cost of processing prompts of that size. For example, at the time of writing, classifying 34 million records using GPT-5 could take over 200 days and cost over EUR 650,000.[10,11]

Project Spectrum explores an alternative approach, where AI is used only to transform product information into embeddings, which are then classified into product categories by a classic machine learning algorithm (Graph 1). The project aims to show that this approach, hereafter referred to as the "embedding-based classifier", can achieve similar accuracy to direct LLM prompting while significantly reducing classification time and cost. In addition to being both time- and cost-efficient, text embedding eliminates the randomness associated with prompting. The dynamic nature of LLM models helps direct LLM prompting to increase their performance and efficiency in the long run, but it can reduce replicability and reliability in classification.

From unstructured data to structured statistics — Graph 1



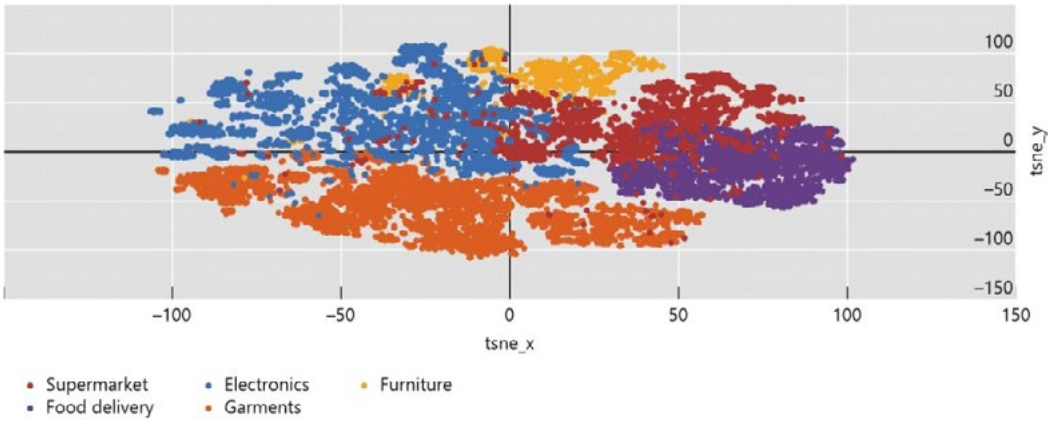
This graph illustrates the process of turning unstructured product data into established categories, enabling the subsequent calculation of category-specific indices.

9. A token is the fundamental unit of text (often a word, part of a word, or a piece of punctuation) that the underlying AI model uses for analysis and generation. An LLM prompt is first translated into a sequence of tokens for processing. The size of the prompt (measured in number of tokens) is the key parameter determining the processing time and cost.
10. This extrapolation is based on empirical tests performed on a smaller number of records by Project Spectrum. Similarly extrapolating from Project Spectrum, one can assume that a human annotator can label two products per minute. During a dedicated 40-hour work week, this person would label 4,800 products. It is clear that this approach becomes infeasible when applied to a data set of 34 million.
11. Most LLM models have some notion of product categorisation from their pre-training. As a result, a rudimentary classification can also be performed by a much shorter prompt that includes the new product description but not the definition of product categories. However, this method has no control over the rules of categorisation, and hence it does not provide the reliability and repeatability that is necessary for inflation analysis.

Text embedding is an AI technique that transforms text into a high-dimensional space based on its semantic content. The result is an n-dimensional vector, where n ranges from hundreds to thousands, depending on the embedding model. Embeddings locate semantically similar texts close to each other in the vector space. This is illustrated in Graph 2, where around 30,000 product descriptions are mapped to a 3,072-dimensional space and then projected to two dimensions using the t-SNE algorithm for illustration. Graph 2 shows that similar products are clustered together, suggesting that embeddings are a promising approach for classification. Once product descriptions are mapped onto this multidimensional space, they can be processed by traditional classification algorithms.

Product clustering in an embedding space — Graph 2

t-SNE of K-means clusters



This graph illustrates the distribution of approximately 30,000 products (each shown as a dot). Product descriptions were first transformed into 3,072-dimensional embeddings, which were subsequently grouped into five distinct clusters using the K-means algorithm. Next, for visualisation, these high-dimensional vectors and clusters were projected to two dimensions using the t-SNE algorithm. Colours represent the algorithmic clusters, while textual labels of each cluster (eg food delivery) were manually assigned based on a visual inspection and selecting the thematic pattern within each group.

Sources: PRISMA DPD and authors' calculations.

Table 1 summarises the main advantages and shortcomings of some product classification approaches. Manual data labelling cannot scale to handle massive web-scraped data. At the same time, traditional keyword-based algorithms (eg regular expressions, lexical or semantic dictionaries) struggle with unstructured, heterogeneous and multilingual product data sets and require constant dictionary updates. Direct LLM prompting can achieve high accuracy for multilingual data, but it is not easy to scale due to the time and cost of processing. Project Spectrum is testing the hypothesis that combining embedding models with classification algorithms is a practical, accurate approach that scales to handle massive web-scraped data sets.

Comparison of classification approaches — Table 1

Approach	Advantage	Drawback
Manual	Accuracy	Scalability
Keyword based	Simplicity	Accuracy, language dependent
Embedding-based classifier	Accuracy, speed, cost, multilingual	Training/reference data required
Direct large-language-model prompting	Accuracy, simplicity, multilingual	Processing time and cost

Source: Authors' elaboration.

To test this hypothesis, Project Spectrum implemented an embedding-based classification method and evaluated it by classifying products in the ECB's DPD according to the European Classification of Individual Consumption by Purpose (ECOICOP) 2018.[12] This is the classification system established by Eurostat (the European Statistical Office) for constructing CPIs. The ECOICOP classification structures consumption items into hierarchical levels: divisions, groups, classes and subclasses, each offering progressively detailed categorisation. Project Spectrum aims to classify products at the subclass level, that is, the five-digit level. As an example, Table 2 shows that a loaf of bread would be classified as 01.1.1.3.

Example of European Classification of Individual Consumption by Purpose (ECOICOP) at different levels — Table 2

Digit	Level description	ECOICOP code	Description
2	Division	01	Food and non-alcoholic beverages
3	Group	01.1	Food
4	Class	01.1.1	Cereals and cereal products
5	Subclass	01.1.1.3	Bread and bakery products

Source: Authors' elaboration.

12. The classification used is the ECOICOP 2018 v1 (European Classification of Individual Consumption According to Purpose, 2018 version 1). ECOICOP is the European adaptation of the United Nation's COICOP (Classification of Individual Consumption According to Purpose). While based on the latest COICOP 2018 structure, this version (ECOICOP 2018 v1) does not achieve complete consistency with the final international COICOP 2018 version. For more details on the methodology and on ECOICOP Classification, see Eurostat (2024), and for details on the classification, see United Nations Statistics Division (2018), https://unstats.un.org/unsd/classifications/unsdclassifications/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf. The description of each ECOICOP five-digit category in different languages and tabulated for analytic purposes can be found at https://showvoc.op.europa.eu/#/datasets/ESTAT_European_Classification_of_Individual_Consumption_according_to_Purpose_%28ECOICOP%29/downloads. ShowVoc is a web-based, multilingual platform for publishing, browsing and consuming data sets that comply with Semantic Web standards. The Publications Office of the European Union uses an instance of ShowVoc to disseminate its vocabularies, statistical classifications (such as NACE codes used by Eurostat) and code lists.

4. Data

4.1. Daily Price Data set (DPD)

The DPD is a comprehensive data set that combines high-frequency price data with extensive metadata – such as product names, descriptions and shop-level details – enabling granular analysis of price developments. It focuses on euro area retailers with large market shares, in relevant cities, that sell both offline and online.[13] The foundation of the DPD data collection initiative was the establishment of PRISMA by the European System of Central Banks to explore the use of various micro data sources on prices, including online price data (Osbat et al (2022)).

The collection of online prices spans multiple countries and languages. Each retailer is assigned a unique shop identifier that remains unchanged over time, facilitating consistent tracking of shops and their observed products. Each observation includes the following:

- **Date:** the date and time of price collection.
- **Identifiers:** an anonymised product ID and anonymised shop ID.
- **Name:** a concise textual explanation of the product (eg "Pizza Margherita").
- **Description:** a more detailed text containing additional characteristics (eg flavour, size and composition).
- **Shop category:** the category label used by the online retailer (eg "Frozen Foods"), indicating how the product is organised on the website.
- **Sector:** the retailer sector being scraped (eg supermarket, electronics, fashion, furniture or food delivery platforms).

The unstructured and "noisy" nature of the raw data driven by language differences, retailer-specific naming conventions, frequent missing attributes, and temporal changes in product identifiers creates substantial hurdles for data treatment. Table 3 shows a few examples of individual observations (with anonymised identifiers).

13. DPD data collection complies with strict ethical standards. The targeted institutions are consulted before conducting web scraping, and the ECB is committed to ensure data confidentiality. Web scraping follows protocols that minimise the impact on website traffic. The anonymised data provided to European System of Central Banks researchers ensures the anonymity of the retailer and protects sensitive information.

Examples of observations in the DPD — Table 3

Name	Shop_ category	Sector	Shop_id	Product_id	Description
Soda 33cl	Boissons	Food_delivery	Id_A	Id_1	Gasseuse
Olive oil 500ml	Home | Dispensa	Supermarket	Id_B	Id_2	Frantoio Toscano
Lavadora carga frontal	Gran electrodomestico | Lavadoras y secadoras | Lavadoras carga frontal	Electronics	Id_C	Id_3	Capacidad de 9kg y rendimiento para familias... motor inverte...
Jeans	Herren | Bekleidung | Hosen | Cargohosen	Fashion	Id_D	Id_4	Straight fit kernige Kontrastnate 5 pocket. Loose fit. Used look.
Chaise teinte rouge	Meubles | Tables & bureaux	Furniture	Id_E	Id_5	Solide structure en bois massif...

Source: Authors' elaboration.

The ECB provided Project Spectrum with a data set containing all uniquely identifiable products in the DPD as of December 2025, around 34 million.[14] The data set includes 111 unique shop identifiers (around 60 distinct retailers, some of which operate in multiple countries). The data set consists of data fields relevant to classification, but it excludes price information. The length and completeness of the textual fields vary significantly. The name and description fields average 60 and 581 characters, respectively. In some cases, descriptions reach over 100,000 characters. However, not all products have complete textual information – approximately 400,000 lack both a name and a description.

4.2. Ground truth via manual labelling of reference and test data sets

Project Spectrum required manually labelled data for two purposes. First, a **reference data set** is used to train the classification algorithms. Embeddings transform product descriptions into numerical vectors, and classification algorithms then map descriptions to product categories. These algorithms are trained on reference data – example product descriptions that have been manually assigned to a given category. The project's reference data set includes 30,000 products. This is not a random sample but overweights smaller categories (which also require a minimum number of example records).

The second manually labelled data set is the **test data set**, used to assess how well the classification performs versus human labelling. The test data set – also 30,000 product descriptions – is selected randomly from the available DPD data. This allows for comparing the identified classifications and establishing the goodness of fit. Both data sets were cleaned from duplicates.

14. Duplicate entries often result from either scraping errors, the same product being sold by different retailers or retailers changing the product identifier over time.

The **reference** and **test data sets** used in the project to train and assess the classification models, respectively, were manually labelled. Manual labellers had clear instructions indicating which classification handbook to use and some initial training through feedback on their assigned labels in repeated rounds of iteration. For the reference set, each batch was manually checked for errors; if any were found, the correct labels were communicated back to the labellers, who then re-labelled the same batch or proceeded with the next one. This iterative process helped improve consistency and accuracy. For the test set, no such iteration took place – the trained labellers received the batch once and labelled it directly.

Some inaccuracy is expected in manual labelling due to ambiguities and human error. Table 4 illustrates this by comparing labels assigned by two different annotators for three examples of web-scraped product descriptions. The first product is a baby towel. One of the labellers classified it as a baby product, which may seem correct at first sight, but the proper label is 05.2.0.3, bathroom linen. The second example is uncooked pasta available via a food delivery service, which was labelled as a pasta product by one labeller and as takeaway food by another. These two examples illustrate the intrinsic ambiguity of categorisation. The third example illustrates human error: a cleaning product was mistakenly labelled as a hygiene article by one of the two labellers.

Examples of manual labelling disagreement — Table 4

Name	Shop_category	Sector	Labeller 1	Labeller 2
Bybadetuch mit Kapuze 	Hirsch | Braun 80x80 cm & Kinder	Furniture	12.3.2.2 Article for babies 	05.2.0.3 Table linen and bathroom linen 
Terre d'Italia Trofie della Liguria 	Pasta, Riso e farine	Food_delivery	01.1.1.6 Pasta products and couscous	11.1.1.2 Fast food and take away food services 
Lejia perfumada con detergente frescor marinol 		Supermarket	05.6.1.1 Cleaning and maintenance products 	12.1.3.2 Articles for personal hygiene and wellness, esoteric products and beauty products 

Source: PRISMA DPD and authors' elaboration.

To assess the quality of manual labelling used in the Spectrum Project, a random subsample of the labelled data (250 records) was validated by additional expert annotators. Based on this subsample, the error rate of human labellers is estimated at 8–10%, which is considered tolerable. The degree of agreement between annotators, as measured by Fleiss' Kappa index, is moderate at the ECOICOP five-digit level, indicating the difficulty and ambiguity of properly labelling products at such a fine level.

4.3. Sample representativeness

Whereas traditional household consumption surveys are built on strict probabilistic sampling to represent the entire economy, web-scraped data simply reflect the digital inventory and commercial priorities of specific retailers. Consequently, the composition of web-scaped data sets does not naturally scale to represent the entire economy. First, broad categorical gaps exist because entire sectors – such as energy, housing rents, telecommunications and motor cars – are either not sold online or are technically difficult to track. Second, even within covered sectors, the data often suffer from subclass imbalances where the volume of scraped items does not match actual consumer behaviour. For instance, a retailer's online catalogue may offer hundreds of individual listings for books, while providing very few for fresh fish, creating a mismatch between the data set's composition and the actual weights of the CPI basket.

Despite these sampling challenges, the DPD at the time of writing covers approximately 50% of euro area CPI expenditures.[15] While coverage varies by data set, based on website selection and national consumption habits, coverage figures confirm that web-scraped data can capture a substantial portion of the inflationary landscape. The DPD covers key economic sectors, including food, clothing, personal care, electronics and furniture. The ad hoc nature of web-scraped data collection can introduce specific selection biases. Because web scraping is limited to retailers with an online presence, certain market segments – such as large distributors in the food sector – are heavily represented, while smaller "corner shops" are excluded. The specific choices made regarding which data to web scrape can further compound these biases. The selection biases that may occur when using web scraping have been extensively analysed in Cavallo (2015) and Cavallo and Rigobon (2016). The data collection approach of the DPD aims to minimise selection bias by targeting representative retailers that operate through both online and brick-and-mortar channels, while also ensuring coverage across a variety of sectors.

The Spectrum data set's coverage of ECOICOP categories is illustrated in Graph 3. Since the test data set is a random sample, the graph is illustrative of the entire Spectrum data set's composition. The number of categories covered is 154 out of 253, totalling 50.16% of CPI. A notable challenge remains at the granular level, where several categories are under-represented with fewer than five data points each. These imbalances are partly mitigated by the use of a curated reference data set and by the iterative refinement process that further expands the reference and validation data sets after deployment.

15. To compute the data set's coverage in terms of the CPI basket, the project used Eurostat weights for euro area–19. This coverage may vary for individual economies due to differences in their national consumer baskets. This coverage estimate aligns with the scope reported in comparable studies, such as Cavallo and Rigobon's (2016), in which, using web-scraped data for 25 countries, they cover at least 70% of the weights of their corresponding CPI; and Gautier et al (2023), who, using micro-CPI data for 11 euro area countries, on average, cover 60%. This notwithstanding, there is substantial heterogeneity in terms of CPI coverage across countries. Coverage depends primarily on the selection of websites for scraping or, in the case of micro-price data provided by National Statistical Offices, the specific prices chosen for collection and disclosure.

Product coverage of Project Spectrum's test data set — Graph 3

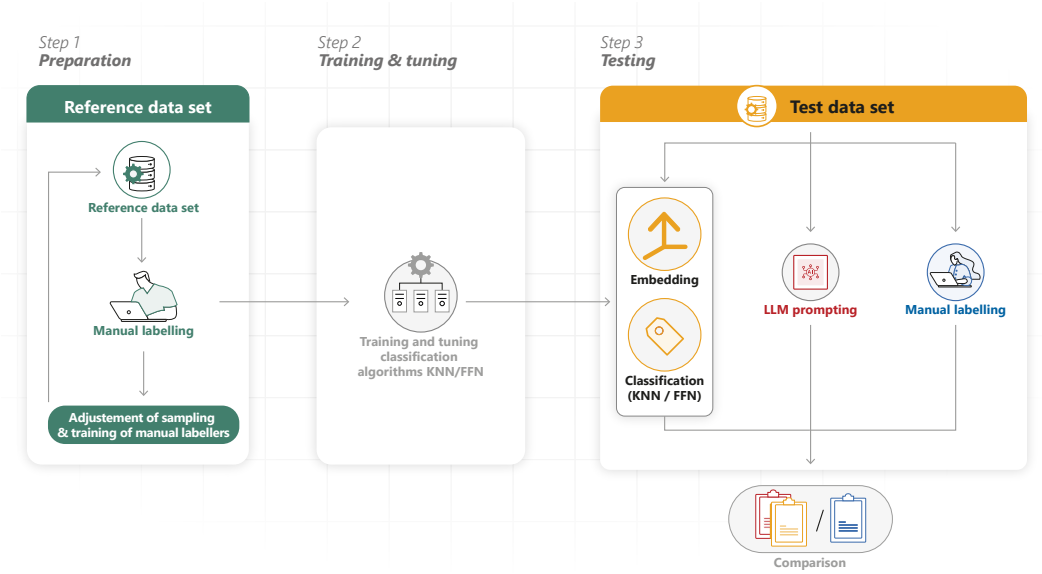
This treemap displays the euro area–19 CPI basket by ECOICOP divisions (parent) and subclasses (child). Box sizes are proportional to their expenditure weight, while colours reflect the number of records within the test data set. Dark blue represents subclasses with five or more records, light blue indicates less than five records, and white denotes zero records. Division shading reflects the cumulative coverage of its underlying subclasses.

Sources: Eurostat and Project Spectrum's manually labelled test data set.

5. Implementation

Graph 4 illustrates the main steps carried out in Project Spectrum. Step 1 involved preparing a curated reference data set of 30,000 records and simultaneously training the manual labellers. In step 2, the reference data set was used to configure, train and fine-tune the classification algorithms. Step 3 consisted of classifying a test data set of 30,000 records using two alternative methods (the embedding-based classifier and direct LLM prompting) and then comparing the predicted classifications with manual labels.

Project Spectrum at a glance — Graph 4



Note: KNN = k-nearest neighbours; FFN = feedforward neural network; LLM = large language model.

Source: Authors' elaboration.

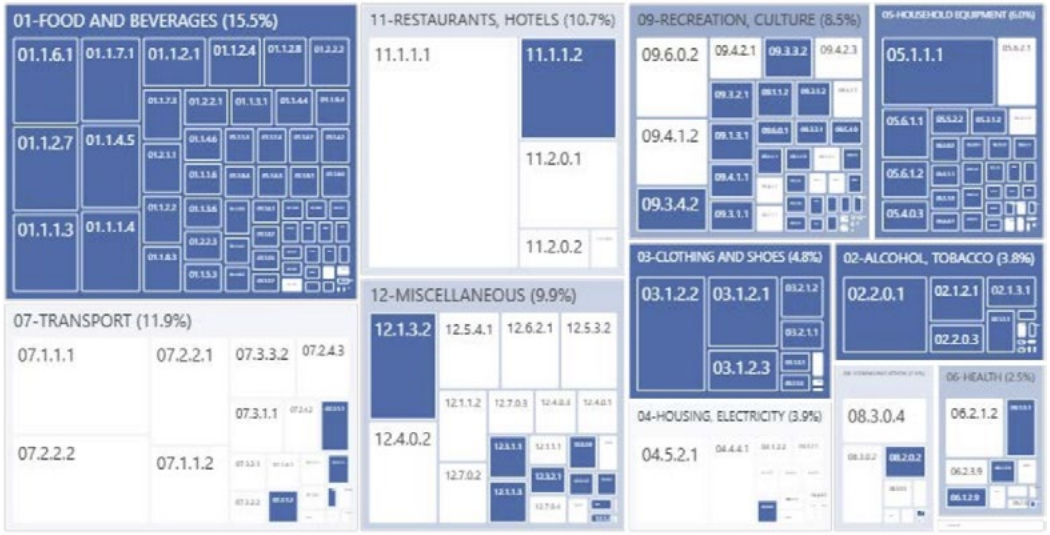
The embedding-based classifier was implemented in two variants: one using the k-nearest neighbours (KNN) algorithm, and one using a feedforward neural network (FFN). Both algorithms are widely used for multiclass prediction problems. The two variants used the same embedding vectors as input and were trained on the same curated reference data set.

5.1. Curated reference data set

The Project Spectrum data set presents some classification challenges because of its severe class imbalance – characterised by a high number of records within some categories and sparsity within others. To overcome this problem, the reference data set was curated by reducing the number of products in over-represented categories (eg food deliveries) and increasing the relative frequency of records in rare categories. Technically, this was achieved by starting with a larger random sample and removing records from categories with high representation.

Graph 5 shows the coverage of the resulting curated data set. Though many categories still have low or zero coverage, the number with adequate coverage has increased compared with the random sample. Notably, the reference data set shows a good coverage in Food and beverages (division 01). Results presented hereafter are limited to the coloured categories, excluding those for which reference data were not available to the project.

Product coverage of the curated reference data set — Graph 5



This treemap displays the euro area–19 CPI basket by ECOICOP divisions (parent) and subclasses (child). Box sizes are proportional to their expenditure weight, while colours reflect the number of records within the curated reference data set. Dark blue represents subclasses with five or more records, light blue indicates less than five records, and white denotes zero records. Division shading reflects the cumulative coverage of its underlying subclasses.

Sources: Eurostat and Spectrum Project's manually labelled reference data set.

5.2. Embedding model

The project experimented with several embedding models, including Text-Embedding-3-Large and Text-Embedding-3-Small, both part of OpenAI's third-generation embedding series. These embeddings are optimised for various natural language processing tasks, particularly those requiring semantic understanding (OpenAI (2024a)).

Text-Embedding-3-Large supports up to 3,072 dimensions, demonstrating superior performance on benchmarks like MIRACL (multilingual information retrieval across a continuum of languages) and MTEB (massive text embedding benchmarks). MIRACL evaluates multilingual information retrieval, while MTEB benchmarks text embeddings across tasks such as clustering, classification and retrieval. In contrast, Text-Embedding-3-Small has 1,536 dimensions and offers a balance between performance and efficiency. It is ideal for use cases where computational resources are limited but efficient text embeddings are still required (OpenAI (2024a)). Comparative tests (not included in this report) showed that classification accuracy is sensitive to the choice of model. All results presented in this report were obtained using Text-Embedding-3-Large, as it performed better than smaller models.

To prepare the data, the text fields of each record – *name, description, shop category, and sector* – were concatenated into a single, unified text field. This combined text was then converted into numerical vectors using the embedding model. In a future production deployment, these preparatory steps will have to be performed for each new product that needs to be classified.

5.3. Classification using the k-nearest neighbour algorithm

K-nearest neighbours (KNN) is a multiclass classifier algorithm that predicts a data point's subclass based on its proximity to neighbours in a multidimensional feature space. In Project Spectrum, this feature space corresponds to the embedding space; consequently, KNN operates by calculating the similarity between high-dimensional text embeddings.

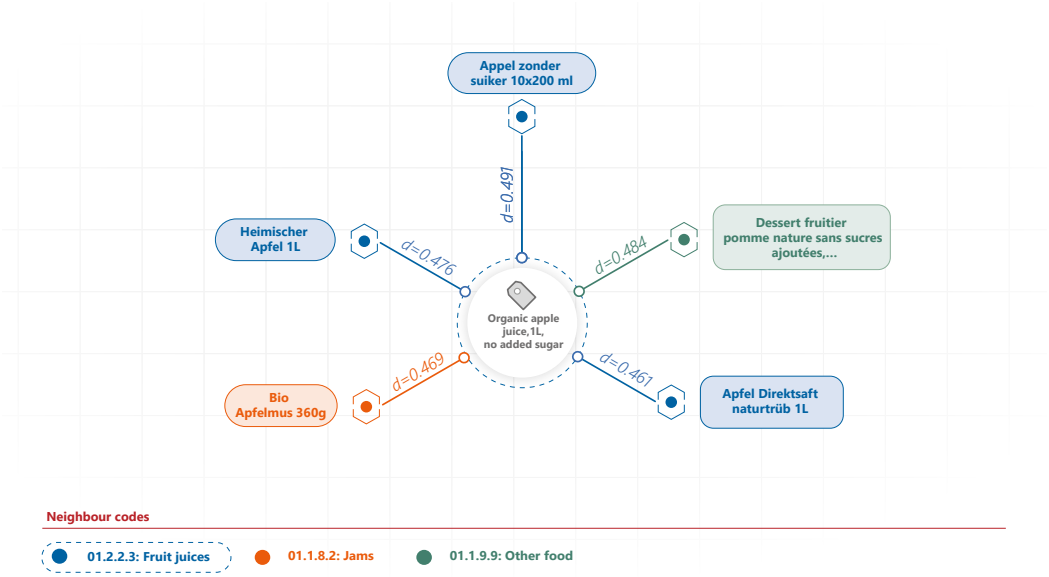
For each target product, the system generates an embedding vector and identifies the k reference products with the highest vector similarity. Vector similarity ranges from 1 (complete similarity) to –1 (opposite directions). The predicted class is determined by a majority vote among these k neighbours, ensuring that it reflects the most frequent category in the local neighbourhood.

An empirical evaluation of various distance metrics and hyperparameter configurations was performed to identify the setting that provides the highest classification accuracy. For the parameter k, odd values between 1 and 15 were tested. While k = 11 achieved the highest weighted-average F1 score (by a margin less than 1 percentage point), k = 5 was selected for its superior performance when accounting for CPI category weights. For vector similarity, Euclidean, Manhattan and Cosine distances were evaluated. Euclidean and Manhattan distances focus on coordinate-based gaps, while Cosine distance focuses on the orientation of the feature vectors. In the tests, Euclidean and Cosine distances yielded similar classification results, with Manhattan distance performing worse. Cosine distance was ultimately selected as it is a standard approach for the KNN algorithm. To resolve ties in the majority voting scheme, random selection and distance-based weighting were tested, and the former was selected for the final implementation.

Graph 6 illustrates the KNN inference procedure for a new product with the name and description "Organic apple juice, 1L bottle, no added sugar". In the graph, the circle in the middle represents the new product to be classified, while the five circles around it correspond to the five nearest reference data points in the embedding space. As indicated by the colour coding, three of these neighbours belong to the ECOICOP subclass 01.2.2.3, one to subclass 01.1.8.2 and one to subclass 01.1.9.9. Based on majority voting, the new product will be classified as 01.2.2.3 – fruit and vegetable juices – which is the correct label.

Classifying products based on k-nearest neighbours (KNN) Graph 6

Under which categories are the closest five products?



This graph shows how a new product is categorised based on its similarity to existing items. First, all products are converted into numerical vectors using the text-embedding-3-large model. The central circle represents a new, unclassified product. The algorithm identifies the k = 5 closest items from the curated reference set, where "closeness" is measured by Cosine distance d. Finally, the new product is labelled with the most frequent category (the majority label) among these five neighbours; any ties are resolved by random selection. In this example the predicted category is 01.2.3 "Fruit juices".

Source: Authors' elaboration.

5.4. Classification using a feedforward neural network

As a second option for the embedding-based classifier, an alternative to KNN, the project designed a custom feedforward neural network (FFN). The aim was to develop a supervised method that could learn features from labelled embedded data, process them and make label predictions.

First, the *curated reference data set* was split into two sets: one for training (85%) and the other for fine tuning (15%). The splits were stratified by ECOICOP codes to maintain balanced category distributions and to prevent any data leakage or memorisation.

The FFN architecture consisted of an input layer that received the embedding vectors, followed by two hidden layers that progressively transformed the input into more meaningful representations for classification. The final output layer produced the predicted class.[16]

16. *The input layer matched the dimension of the embedding space (3,072); the two hidden layers had 128 and 64 neurons, respectively; and the output layer had a number of neurons matching the number of ECOICOP subclasses. The model was trained using the Adam optimisation algorithm with cross-entropy loss. Regularisation techniques such as dropout, batch normalisation and weight decay were used to improve generalisation. In addition to 50 epochs, early stopping was applied to retain the model configuration that achieved the best validation performance.*

Using two hidden layers, the models were able to capture complex patterns in the data encoded in the embedding vectors and learn their relationships. The model was trained for 50 rounds (epochs), each followed by evaluation on unseen data, during which the weights were adjusted to ensure proper information capture and prevent overfitting. Once training was complete, the best model was saved, and the final performance was measured on the test data set.

5.5. Direct large language model prompting

For direct LLM prompting, the project used OpenAI's GPT-5 model with an instruction prompt that explicitly referenced the ECOICOP 2018 v1 handbook. The model was instructed to assign the most appropriate ECOICOP five-digit code to each product, based on the official ECOICOP definitions and descriptions. As an input, the model was asked to consider the following columns: name, description, shop category, sector and item type. The process was iteratively applied across all products, yielding an ECOICOP prediction for each.

One challenge with direct LLM prompting was classifying food delivery items. These products often overlap semantically with items belonging to division 01–Food and non-alcoholic beverages, while being formally categorised under "11.1.1.2–Take-away food services." Without additional guidance, the model would tend to classify, for example, a can of soda, under "01.2.2.2–Soft drinks", which aligns semantically with the product description, but does not align with the intended label of the data set. This is where the columns "item type" and "sector" help navigate to the right division: the sector value "supermarket" indicates that a product belongs to division 01, that is, a food item or beverage that can be purchased in the supermarket for consumption at home. To address this issue, the prompt was refined to explicitly instruct the model to assign products to class 11.1.1.2 if "sector" corresponds to food delivery.

6. Evaluation

6.1. Classification accuracy

To evaluate performance, the predicted ECOICOP categories were compared with the manual labels, which serve as the ground-truth benchmark. The project defined classification accuracy as the ratio of predictions that match the manual label at the subclass level (ie the five-digit level of ECOICOP). Graph 7 illustrates the classification accuracy of the embedding-based classifiers (KNN and FFN) for each ECOICOP category.[17]

Some categories, such as food delivery (11.1.1.2) or household furniture (05.1.1.1), achieve a very high classification accuracy, over 90%. Many clothing and home appliance categories reach an accuracy above 70%. Some other categories, notably health and housing, see a lower accuracy. Note that the graph excludes categories not covered in the reference data set, such as education or energy, in order to focus on the targeted categories.

Accuracy of the embedding-based classifier at the European Classification of Individual Consumption by Purpose (ECOICOP) five-digit level — Graph 7

17. *Accuracy on a sub-class level is calculated by dividing the number of correct predictions on that sub-class by the total number of products manually labelled in the same sub-class. Using the terminology of classification algorithms, this corresponds to in-class recall, often written as TP / (TP + FN) where TP stands for "true positive" and FN for "false negative" predictions.*

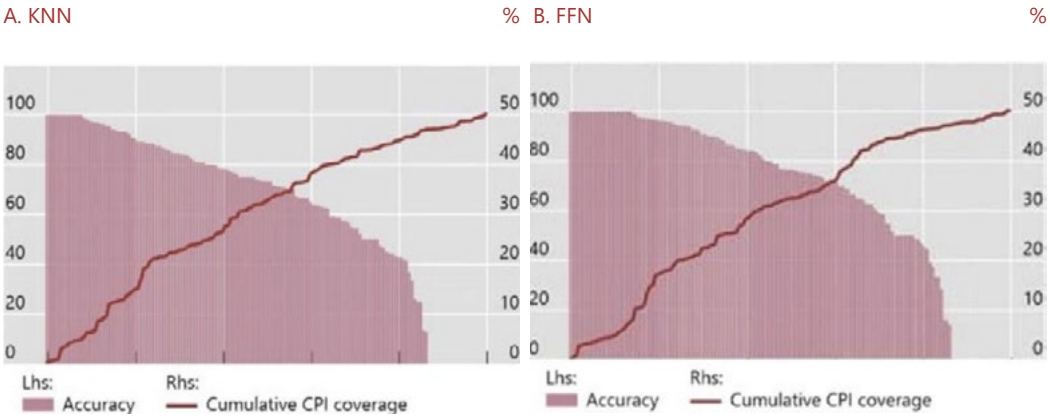
B. FNN

Treemaps of in-class prediction accuracy for KNN (panel A) and FFN (panel B) combined with text-embedding-3-large on the test data set of 30,000 records. Box sizes are proportional to expenditure weights in the euro area–19 CPI basket across ECOICOP divisions (parent) and subclasses (child). Shading represents accuracy levels ranging from dark green (high accuracy) to white (low accuracy). Categories lacking reference data are excluded.

Sources: Eurostat and Spectrum data set.

In Graph 8, each bar corresponds to one ECOICOP five-digit category (ie subclass). The categories are ordered by classification accuracy, from most to least accurate. The line tracks the cumulative CPI coverage of the categories, concluding at approximately 50%, which aligns with the data set's total coverage. The leftmost few categories have a very high classification accuracy, but they collectively cover a small portion of CPI. Moving progressively towards the right, the cumulative CPI coverage increases but the classification accuracy decreases with each new category. The graph shows a substantial degree of heterogeneity in terms of accuracy across categories. For example, categories with at least 70% accuracy collectively cover 36% of CPI (KNN).

Model performance across ECOICOP categories Graph 8



In-class prediction accuracy for KNN (panel A) and FFN (panel B) combined with text-embedding-3-large on the test data set of 30,000 records. Each bar represents a specific five-digit ECOICOP category, ranked by its prediction accuracy. The overlaid line tracks the cumulative weight of these categories within the euro area-19 CPI basket. This allows for a comparison between model precision and the actual economic significance of each product group.

Sources: Eurostat and Project Spectrum data set.

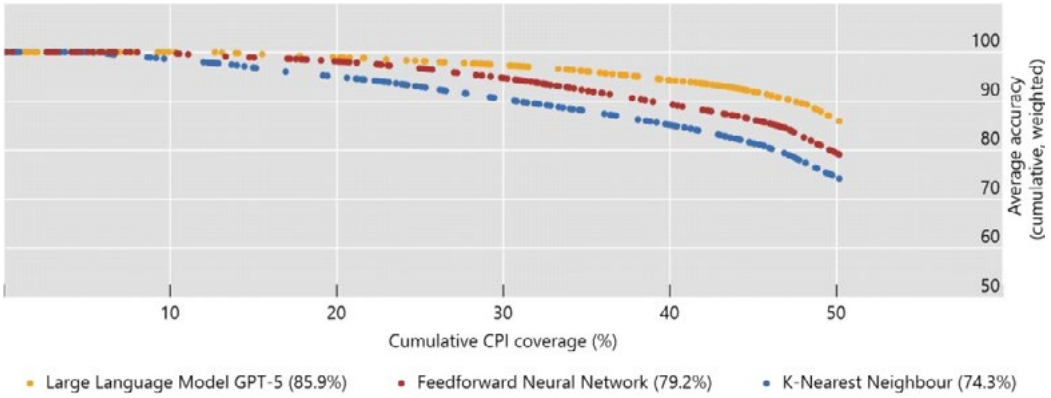
If one were to set a minimum threshold for accuracy, then CPI coverage would naturally be limited. Each possible level of such a threshold represents a different balance between accuracy and CPI coverage for the given classification approach.

This is illustrated in Graph 9, which shows the cumulative weighted average classification accuracy against cumulative CPI coverage for three classification approaches. On the left, high accuracy is achieved for a small number of categories, resulting in low overall CPI coverage. Moving to the right, CPI coverage increases but weighted average classification accuracy decreases. The rightmost end of the curves corresponds to the entire Project Spectrum data set, covering approximately 50% of CPI. At this point, direct LLM prompting achieves 86% accuracy, while the embedding-based classifiers reach 80% and 75% for FFN and KNN, respectively.

The three curves have similar shape, confirming the natural trade-off between accuracy and CPI coverage. The graph also shows that the embedding-based classifier achieves reasonable accuracy, albeit somewhat lower than full LLM prompting, in line with the project's initial hypothesis.

Applying FFN after text embeddings achieves higher classification accuracy than applying KNN on the same embeddings. This is not surprising, as neural networks often perform better on complex multiclass classification tasks than simpler, distance-based algorithms. In contrast, KNN is a more intuitive approach and offers greater traceability than FFN, which operates as a black box. In the context of inflation analysis, this trade-off might be considered when selecting the most suitable algorithm.

Classification accuracy – CPI coverage trade-off Graph 9



Average weighted accuracy as a function of cumulative CPI coverage for the euro area–19. The curves compare direct prompting (GPT-5) against FFN and KNN classifiers using text-embedding-3-large. Results are evaluated on the test data set of 30,000 records, with final aggregate accuracy scores noted in the legend.

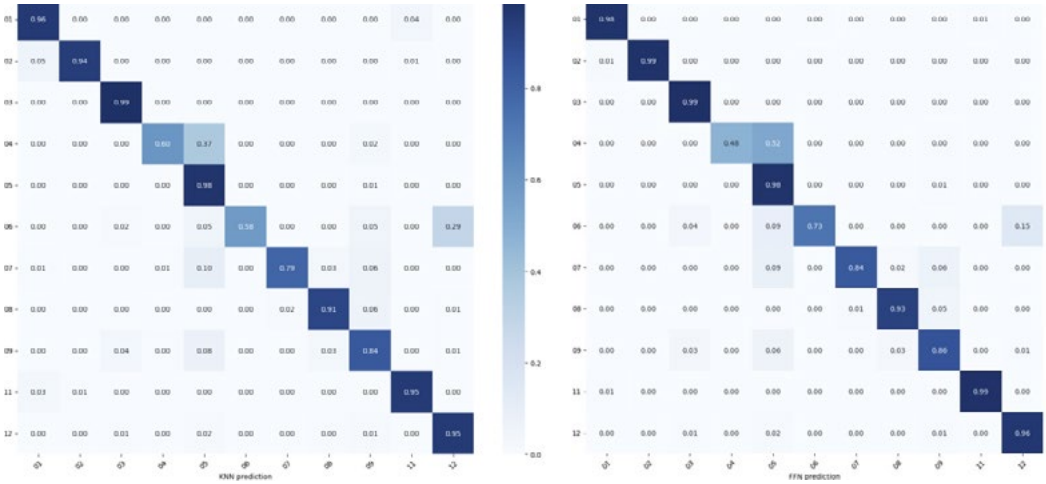
Sources: Eurostat and Project Spectrum data set.

Project Spectrum aims to correctly classify products at the five-digit (ie subclass) level. In this analysis, a prediction is classified as erroneous if it fails to match any of the five ECOICOP digits. However, for inflation nowcasting, classification errors at the subclass (five-digit) level usually have a lower impact than those at higher hierarchy levels. This will depend on divergent patterns in pricing across certain categories; for example, no great divergences are expected between women's and men's clothing, while there might be large differences in price developments for olive oil or butter, as they can be subject to different shocks.

The importance of accurate predictions at granular ECOICOP levels will depend on the underlying economic question. For example, when analysing VAT reductions on targeted goods, accuracy at a very granular level is crucial to evaluate the VAT pass-through to final prices (see Forteza (2025)). This granularity is also relevant when analysing, for example, the transmission of specific shocks or the degree of volatility. In contrast, studies show that for inflation analysis, classifying products at the division level, and ignoring more granular levels, can be sufficient, as explored by Beer et al (2025) using web-scraped prices for food forecasting.

To illustrate the embedding-based classifier's accuracy at the division level, Graph 10 compares classifier predictions with manual labels. The graph shows that most predictions stay within the correct division, with only sporadic errors occurring across divisions. This is a promising result, which further supports the applicability of the embedding-based classifier for inflation analysis.

Evaluation of classification errors on a division level — Graph 10



Normalised confusion matrices comparing predictions by text-embedding-3-large combined with KNN (panel A) and FFN (panel B) to manual labels in the test data set of 30,000 records. Each row corresponds to a true division according to manual labels, and each column corresponds to a division predicted by the embedding-based classifier. Rows are normalised by the total number of manual labels in that division. Numbers and shading correspond to the ratio of products with the given manual label and predicted division. Diagonal cells represent predictions that are correct on the division level, while off-diagonal cells correspond to division-level classification errors.

Sources: Eurostat, PRISMA DPD and authors' calculations.

6.2. Feasibility and cost comparison of classification methods

Table 5 compares processing cost and execution time between direct LLM prompting and an embedding-based classifier combined with KNN or FFN. The results confirm that the embedding-based classifier requires a fraction of the cost and time of LLM prompting. Text embedding is a light AI function, and the classification algorithms can be performed very efficiently using, for example, vector similarity operations. In contrast, for direct LLM prompting, every product classification involves an LLM inference step, which is much more resource intensive.

Average processing latency and operational cost per product record — Table 5

Model approach	Mean time per 1,000 products (seconds)	Mean cost per 1,000 products (EUR)
Direct LLM (GPT-5)	~500.0	22,2
Embedding + KNN	< 8.1	< 0. 030
Embedding + FFN	< 6.8	< 0. 031

Average processing time and cost per product record. The analysis compares direct LLM prompting (GPT-5) with classification via text-embedding-3-large combined with KNN or FFN.

Sources: PRISMA DPD and authors' elaboration.

Besides cost and execution time, another advantage of the embedding-based classifier is its modularity. Embedded product descriptions can be saved and reused, for example, if new classification categories are introduced or if the classification algorithm is improved. This results in a flexible system that can adapt to the latest technology. As embedding models evolve, they can be updated in the embedding-based classifier, and if better classifiers are created, those can be deployed while keeping the embedding model unchanged. Users can combine the embedding model and classifier that best suits their specific needs.

In contrast, direct LLM prompting is a one-step process that must be repeated for each new classification. The prompt workflow has only a few intermediate steps; hence, reusability is limited if, for example, a different LLM needs to be tested. Furthermore, direct LLM prompting is a black-box approach, whereas certain classification algorithms, notably KNN, offer higher levels of explainability.

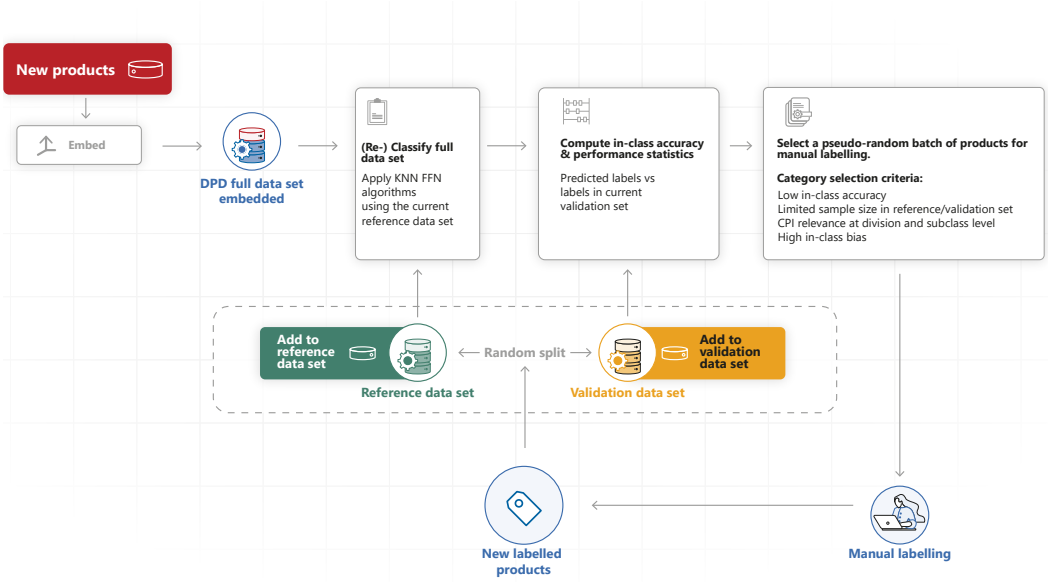
7. Initial deployment and a continuous refinement process

The results presented above demonstrate the concept's applicability to the set of product categories covered by the manually labelled reference data set, which at the time of writing covered around 50% of the euro area CPI expenditure basket. Besides classifying the entire existing DPD, the project has developed the solution as a production pipeline that can classify new products as they appear on the market. This is important due to the high turnover of products in web-scraped data; each month new products enter the data set and need to be classified. The Spectrum pipeline performs this process at a high speed – it takes around three hours of computing time for one million new products.

Classification accuracy and coverage can be further improved by enlarging the reference data set. With Project Spectrum's initial reference data set of 30,000 products, the accuracy varies across ECOICOP categories due to sample sparsity and heterogeneity in the underlying product classes. This is an unavoidable consequence of starting from a random sample: by construction, sampling variability leaves some categories under-represented and others over-represented, which in turn depresses classification accuracy for the under-represented product classes.

The Spectrum pipeline therefore supports an iterative human-in-the-loop process to gradually extend the reference and validation data sets, selectively focusing on under-represented product categories. This is implemented as a semi-automated process with a human labeller in the loop (see Graph 11). It allows for the gradual increase in overall accuracy while narrowing the performance gap between categories. Adding more reference data also helps address potential new product categories appearing in an ever-evolving product landscape.

Iterative process to expand the reference data set — Graph 11



Source: Authors' elaboration.

The process to enlarge the reference data selects new records for manual labelling in batches, for example 1,000 products at a time. These records are selected by stratified sampling, prioritising subclasses based on a combination of criteria. The initial set of prioritisation criteria include manual testing results, classification accuracy, number of manually labelled records, CPI weight and the statistical properties of each class in the embedding space (specifically in-class bias).[18]

The sample selection is performed based on the records' location in the embedding space, oversampling the vicinity of prioritised subclasses. After a new batch of products is manually labelled, these products are added either to the reference or to the validation data sets, and all DPD products are reclassified.[19] Finally, in-class accuracies and other relevant statistics are re-computed, as these will serve as an input to selecting the next batch of products to label manually. This process is designed as a self-reinforcing path to progressively improve classification accuracy across all categories and make the embedding-based classifier an increasingly accurate tool for structuring product data supporting inflation analysis.

Project Spectrum has implemented this methodology, initiated the iterative refinement process and completed the first few iteration cycles. This involved manually labelling another 6,000 products, which were equally distributed between the reference and validation data sets. The resulting data – the product embeddings and preliminary classifications for all 34 million products in the DPD, as well as boosted reference and validation data sets – have been made available to project partners.

Further improving classification accuracy and keeping the classification algorithm up to date with an evolving product portfolio will require continuous iteration and occasional manual labelling of new data batches. Project Spectrum has proved that such a continuous operation is realistic from a cost and time perspective, providing analysts and policymakers with increasingly accurate and detailed insights into price developments.

18. It is expected that these prioritisation criteria might need to be fine-tuned as the iteration is progressing.
19. Classifying the entire data set is computationally efficient, totalling approximately EUR 150 in processing costs. Linear scaling is possible by increasing concurrency. The embeddings do not need to be re-computed.

8. Conclusions and next steps

Project Spectrum has explored the potential of AI in transforming unstructured data into actionable economic intelligence. As digitalisation and e-commerce expand, leveraging online price data is becoming increasingly vital for effective monetary policymaking. The project has proved that combining text embeddings with machine-learning-based classification provides an efficient approach to structuring these high-frequency, large-scale online price data sets.

The embedding-based classifier achieves comparable accuracy with direct LLM prompting while demanding significantly less cost and processing time. Given the vast volume and diversity of product-level data, automating and refining product classification at scale represent major steps forward that could ultimately improve inflation nowcasting when using web-scraped data. In dynamic economic environments, where product-level price shifts can signal broader inflationary trends before they appear in aggregated indices, access to granular, near-instant insights can give policymakers a forward-looking advantage.

This research contributes to the growing body of literature on automated classification systems applied to economic data. It offers practical implications for statistical agencies and economic researchers working with web-scraped product information. In addition, knowledge about automatically structuring data at scale also supports broader applications of AI in financial stability assessment, risk modelling and data-driven policy strategies.

The experimental results presented in this report focus on data collected in the euro area; however, the approach is globally applicable thanks to the multilingual capabilities of AI models. For broad use in economic analysis, analysts will need to expand the reference data set – for which this report has presented a viable pathway. Already scalable up to production, the embedding-based classifier promises to turn data abundance into actionable economic understanding.

Project Spectrum also opens the door to further work in this area. A potential next step is to test the solution with different data sets and languages. Though Project Spectrum has focused on data sets and retailers within the euro area, the multilingual capabilities of embedding models make the approach globally applicable.

A further validation step involves constructing CPI indices at the ECOICOP subclass level to benchmark them against the "gold standard" of official inflation time series. Executing this historical back-test requires historical price data, which remains outside the scope of the current Project Spectrum phase but is identified as a priority in the follow-up phase.

9. Project participants and Acknowledgements

BIS Innovation Hub Eurosystem centre

Enikő Gábor-Tóth, Adviser
Elvira Prades Illanes, Adviser
Franca Seneviratne, Adviser
Oleksandra Toporko, Research Analyst
Timothy Aerts, Eurosystem Centre, Deputy Head
Andras Valko, Eurosystem Centre, Deputy Head
Raphael Auer, Eurosystem Centre Head

European Central Bank

Inês de Almeida Lopes Nunes, Data Science Expert
Jack Horgan, Trainee
Chiara Osbat, Adviser, Monetary Policy Research
Luigi Palumbo, Data Scientist

Deutsche Bundesbank

Manuel Fangmann, Economist
Julia Lassmann, Team Lead, Strategy & Innovation
Martin Pester, Team Lead, Strategy & Innovation
Johannes Rogowsky, Economist

Acknowledgements

The authors are grateful to Morten Bech, Miguel Díaz, Karmela Holtgreve, Fernando Pérez-Cruz and Víctor Rayado Pérez for reviewing this report and providing extremely valuable feedback.

10. Bibliography

Aparicio, D, and MI Bertolotto (2020): "Forecasting inflation with online prices", *International Journal of Forectasting*, vol 36, no 2, pp 232–47.

Alvarez-Blaser S, R Auer, S Lein and A Levchenko (2025): "The granular origins of inflation", *BIS Working Papers,* no. 1240, January.

Beck, G W, K Carstensen, J-O Menz, R Schnorrenberger and E Wieland (2024): "Nowcasting consumer price inflation using high-frequency scanner data: evidence from Germany", *European Central Bank Working Papers,* no 2930.

Beer, C, R Ferstl and B Graf (2025): "Improving disaggregated short-term food inflation forecasts with webscraped data", *Oesterreichische Nationalbank Working Papers,* no 262.

Cavallo, A (2013): "Online and official price indexes: measuring Argentina's inflation", *Journal of Monetary Economics,* vol 60, no 2, pp 152–65.

——— (2015): "Scraped data and sticky prices", *NBER Working Papers*, no 21490.

——— (2018a): "More Amazon effects: online competition and pricing behaviors", *NBER Working Papers*, no 25138.

——— (2018b): "Scraped data and sticky prices", *The Review of Economics and Statistics*, vol 100, no 1, pp 105–19.

Cavallo, A, and R Rigobon (2016): "The Billion Prices Project: using online prices for measurement and research", *Journal of Economic Perspectives,* vol 30, no 2, pp 151–78.

Dedola, L, L Henkel, C Höynk, C Osbat and S. Santoro (2025): "What does new micro price evidence tell us about inflation dynamics and monetary policy transmission?" *ECB Economic Bulletin*, Issue 3.

ECB (2025): "A strategic view on the economic and inflation environment in the euro area", *ECB Occasional Paper Series*, Issue 371.

Eurostat (2024): *Harmonised index of consumer prices (HICP): Methodological Manual – 2024 edition*, Publications Office of the European Union.

Forteza, N, E Prades and M Roca (2025): "Analysing the VAT cut pass-through in Spain using web-scraped supermarket data and machine learning", *Journal of the Spanish Economic Association,* vol 16, no 1, pp 137–89.

Friedman, M (1961): "The lag in effect of monetary policy", *Journal of Political Economy*, vol 69, no 5, pp 447–66.

Gautier, E, C Conflitti, RP Faber, B Fabo, L Fadejeva, V Jouvanceau, J-O Menz, T Messner, P Petroulas, P Roldan-Blanco,F Rumler, S Santoro, E Wieland and H Zimmer (2024): "New facts on consumer price rigidity in the euro area", *American Economic Journal: Macroeconomics*, vol 16, no 4, pp 386–431.

Gautier, E, P Karadi et al (2023): "Price adjustment in the euro area in the low-inflation period: evidence from consumer and producer micro price data", *European Central Bank Occasional Papers,* no 319.

Harchaoui, TM, and RV Janssen (2018): "How can big data enhance the timeliness of official statistics? The case of the US consumer price index", *International Journal of Forecasting,* vol 23, no 2, pp 225–34.

Long, A, W Yin, T Ajanthan, V Nguyen, P Purkait, R Garg, A Blair, C Shen and A van den Hengel (2022): "Retrieval augmented classification for long-tail visual recognition, *arXiv*.

Macias, P, D Stelmasiak and K Szafranek (2023): "Nowcasting food inflation with a massive amount of online prices", *International Journal of Forecasting*, vol 39, no 2, pp 809–26.

OpenAI (2024a): *New embedding models and API updates*, https://openai.com/index/new-embedding-models-and-api-updates/.

Osbat, C et al (2022): *What micro price data teach us about the inflation process: web-scraping in PRISMA, SUERF Policy Brief,* no 470, November.

Sims, C (1992): "Interpreting the macroeconomic time series facts: the effects of monetary policy", *European Economic Review*, vol 36, no 5, pp 975–1000.

Tissot, B, and B de Beer (2020): "Implications of Covid-19 for official statistics: a central banking perspective", *IFC Working Papers*, no 20, November.

Woodford, M (2003): *Interest and prices: foundations of a theory of monetary policy. Princeton University Press.*

Bank for International Settlements (BIS)

ISBN 978-92-9259-929-4 (Online)