AI-Based Scanner Data Classification: Kazakhstan Case



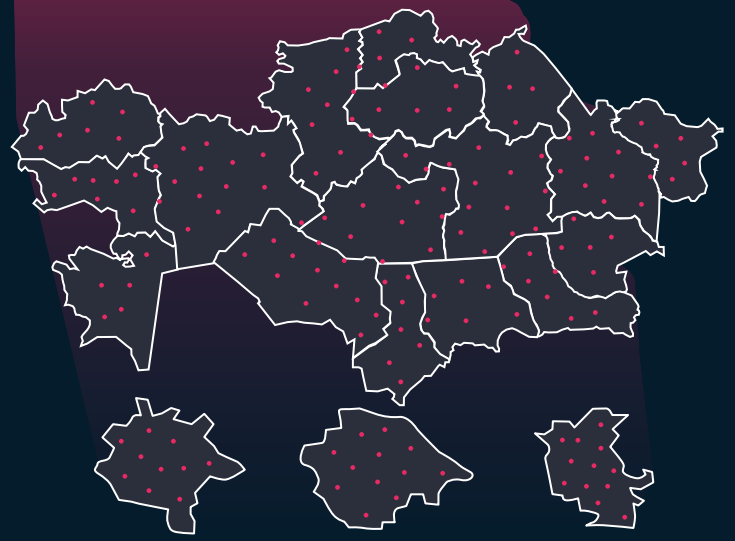
2026

Using fiscal receipt (scanner) data to monitor inflation

Traditional method (AS IS)

120 thousand quotes

12 thousand basic objects
(shops, markets)

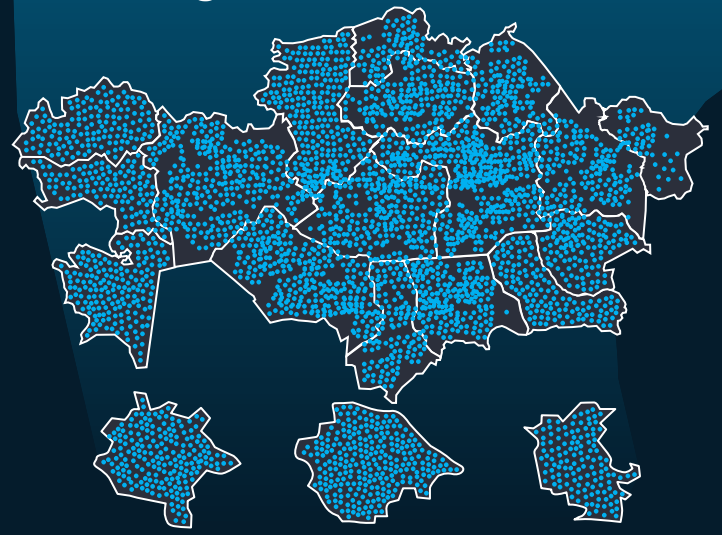


Prices are fixed by price registers
(over 300 employees)

Use of fiscal receipt data (TO BE)

36,9 million quotes

124 thousand retail outlets covered by the categorization



Data from fiscal receipts using AI

The classification challenge



- **6.8 billion annual records** of fiscal receipt data were received from the **Ministry of Finance** in 2024
- The datasets specifically included key variables such as **product names**, **transaction quantities**, and **unit prices**
- Product names **vary significantly** across different retail entities, meaning traditional tools cannot classify them effectively
- A single product can have **more than 100 name variations** within scanner data
- An **automated classification** tool was necessary to process this volume of data for use in Consumer Price Index measurements

Implementation of AI tools for classifying scanner data



Project Goal

Development and implementation of an automated system for classifying product names using modern language models and machine learning methods to use scanner data in CPI calculations



Tasks

- Categorization **product names** within scanner data for further use in **inflation calculation**



Expected Results

- A **fully automated product classification system**
- **Expansion of the product list** used to estimate CPI



Key Effects

- **Accurate inflation monitoring** across a wider range of goods
- Reduced the costs associated with traditional manual field surveys



Data Sources

- Fiscal data from **Fiscal Data Operators**
- Data from the **Statistical Business Register**

Implementation roadmap and resources

1 — Data Cleaning & Algorithm Development (4 months)

- Development of data cleaning algorithms
- Processing and normalization of fiscal receipt data
- Preparation of training dataset

2 — Model Training (5 months)

- GPT-based labeling of 100,000 product names
- Embedding generation using BGE-M3
- Training of classification layer
- Optimization of model performance

3 — Deployment & Integration (2 months)

- Integration into inflation calculation pipeline
- Automation of categorization process
- Scaling to full fiscal dataset

4 — Testing & Validation (1 month)

- Accuracy assessment
- Error analysis
- Validation against traditional inflation methodology

Key Technical Resources

- **GPT** – used for labeling 100,000 product names
- **BGE-M3 model** – embedding generation
- Secure Big Data processing environment

Project Team

- Project Manager
- Data Scientists (ML model development)
- Data Engineer (data preparation & validation)
- IT Engineers (deployment & infrastructure)
- Domain Experts (inflation methodology)

290 product categories are classified with high accuracy

Product Name	Category	Evaluation	Comment
zelenaya marka traditsionnaya ura 0.7 vodka	Vodka	✅	Correct classification, brand name did not affect the result
maslo kisloslivochnoe prostokvashino 82%	Unsalted butter	✅	Accurately identified by product composition and brand
conditioner	Conditioner	✅	Short but unambiguous name
banana kg	Bananas	✅	Correctly recognized based on word form
drink Coca-Cola 2L carbonated	Carbonated beverages	✅	Identified correctly based on brand and volume
?ant a? ??m ?lsh	Failed to classify	❌	Unrecognizable text - likely OCR error
sigma 18-50 mm f2.8 dc dn for Sony E	Camera	❌	Misclassification - this is a lens, not a camera
biomed silver medium k	Toothpaste	⚠️	Partially correct - brand matches, but incomplete description
assortiment	Unrecognized assortment	❌	Insufficient context, lacks product indicators
tennisika	Category not found	❌	Recognition error, likely a distorted word

Typical Successful Examples

- **Brand + quantity:** «Ariel automatic 3 kg», «Nescafe Gold 100 g»
- **Keywords and composition:** «Mayonnaise 67%», «UHT Milk»
- **Clear category reference:** «Gasoline A-92», «Motor oil»
- **Products with measure or volume:** «Potatoes by weight», «Conditioner 200 ml»

Typical Errors (Failed Cases)

- **Unreadable or corrupted strings:** symbols, noise, OCR artifacts `(?ant a, b/lim/^a^i)`
- **Service descriptions instead of products:** «bath visit», «medical services»
- **Ambiguous short terms:** «oil», «laminate», «assortment»
- **Brand confusion:** lens → camera, toothpaste → face cream
- **Mixed alphabet and language issues:** Cyrillic + Latin `(keratin aktobe el, sigma 18-50 mm)`

Integration into Official Statistics

Inflation from 2026. 537 goods and services

TRADITIONAL METHOD

Price collection at retail outlets

379 goods and services
120 thousand price quotations
12 thousand retail outlets

ALTERNATIVE SOURCES

Website scraping

83 goods and services (in 2025 – 43) **63** - non-food products **20** – services
Monthly Data from more than 42 websites
22 thousand prices

Scanner data

75 goods **60** - food products **15** - non-food products.
Monthly inflow **+120 million records**

ADVANTAGES

- **Coverage**
  High level of sales data granularity
- **Speed**
  Prompt access to price information
- **Efficiency**
  Reduced data collection costs
- **Representativeness**
  Improved accuracy of inflation measurement

DEVELOPMENT PROSPECTS

Expansion of Scanner data usage through:

- ✓ Cleaning and structuring of labeled goods
- ✓ Increasing the level of fiscalization across regions

THANKS FOR YOUR ATTENTION

