How much does product classification cost in basis points?

Multilingual classification of Thai product text for consumer price measurement

Thanatchon Chongmankhong*

October 2026

Abstract

Web and scanner prices enter a consumer price index only after each product has been assigned to an item of the official basket. We ask what that classification step costs in measured inflation. Using public Thai data only, we score multilingual text embeddings, a fine-tuned encoder, LLM prompting and cascades on 3,910 held-out official price-collection specifications, weight every score by CPI expenditure, and pass each method's confusion matrix through a simulator of the Thai CPI. Transfer from non-Thai products alone classifies 60.3% of the evaluated weight correctly, and all official labels raise this to 68.8%. A frontier LLM reaches 97.9%, and a cascade scores 97.8% at 27% of its cost on the held-out list. With the LLM's confusion matrix, headline year-on-year inflation would be misstated by 1.9 basis points (RMSE); with cheaper classifiers, by 1.9 to 10.4. Accuracy does not rank methods by this error. Code and derived data are open.

JEL codes: C55, C81, E31. **Keywords:** consumer price index, product classification, multilingual text embeddings, large language models, web prices, measurement error, Thailand.

1 Introduction

Statistical offices and central banks increasingly measure prices from retailers' websites and scanner data (Cavallo and Rigobon, 2016; Eurostat, 2017, 2020). These sources carry millions of products, but each product must first be assigned to an item of the official basket. Done by hand, that step does not scale; done by a model, it introduces an error. The literature on classification for price statistics reports this error as accuracy (Martindale et al., 2020; Harms and Spinder, 2019; BIS Innovation Hub, 2026). A central bank does not target accuracy. It needs to know how far measured inflation would move.

This paper puts that question to one language and one index. We classify Thai product text into the in-scope items of the Thai CPI with four families of method: multilingual text embeddings with simple classifiers, a fine-tuned multilingual encoder, few-shot prompting of two commercial LLMs, and cascades that send a cheap classifier's unsure products to an LLM. We score them on 3,910 held-out official price-collection specifications, published by the Ministry of Commerce with their official item codes, and weight every score by CPI expenditure. We then pass each method's confusion matrix through a simulator of the Thai CPI, built from public weights and item indices, which returns the error the classifier would add to measured headline, food and core inflation in basis points. Three questions organise the paper.

RQ1: how far does transfer from other languages go, and what are official labels worth? Thai labelled product data are scarce, while Open Food Facts holds millions of products in other languages. An e5-large classifier trained only on non-Thai products classifies 60.3% of the evaluated CPI weight correctly. Up to 1,000 official labels add little, and all 24,456 of them raise it to 68.8%.

RQ2: what does accuracy cost? On official Thai text the LLMs are far ahead: Claude Opus 5.5 reaches 97.9% and Claude Haiku 5.5 95.4%, against 69.5% for the best cheap method with the full input (name,

*The views expressed are the author's own and do not necessarily reflect those of the Bank of Thailand. This work uses only publicly available data. Code and derived data: `https://github.com/thanatchon36/cpilens`.

brand and quantity). A cascade from the e5-large classifier to Claude Opus 5.5 scores 97.8% at 27% of its cost on the held-out list, a difference from Claude Opus 5.5 that the reference set cannot tell from zero. Most of the remaining LLM error comes from two places where our description of the items differed from the office's practice, not from misreading Thai.

RQ3: what does classification error cost in basis points? With Claude Opus 5.5's confusion matrix, headline year-on-year inflation would be misstated by 1.9 basis points (RMSE; at most 3.7 at the upper end of a 95% interval that allows errors to cluster) and food inflation by 3.2. The cheap classifiers trained on all labels misstate headline inflation by between 1.9 and 10.4 basis points. Accuracy does not rank methods by this error: two classifiers with about the same weighted accuracy have point estimates that differ several-fold in basis points, though their intervals overlap, because what matters is whether the confused items' prices moved apart.

The paper makes three contributions. First, it turns a confusion matrix into basis points of measured inflation, with a decomposition that attributes the error to individual confusions and an uncertainty band that reflects how few observations some rows rest on. The simulator needs only public weights and item indices, so any office can apply it to its own classifier. Second, it evaluates classification for a CPI in a non-European language, on the office's own held-out specifications and with expenditure weights on both accuracy and coverage. Third, the whole pipeline is public and reproducible: it uses only open data, every number in the paper is generated by code, and the LLM runs cost USD 14.72 in total.

The closest work is the BIS Innovation Hub's Project Spectrum (BIS Innovation Hub, 2026), which classifies euro-area web prices and names testing other languages and back-testing indices as next steps (BIS Innovation Hub, 2026, p. 30). Our taxonomy, language, product universe and labels differ from Spectrum's, so our numbers are not directly comparable with theirs. Section 2 reviews the literature, Section 3 describes the data and Section 4 the methods. Section 5 answers the three questions, Section 6 draws lessons for central banks in Asia, Section 7 lists the limitations and Section 8 concludes.

2 Related work

Online prices. Prices collected from retailers' websites have become a measurement source in their own right (Cavallo and Rigobon, 2016). For large multi-channel retailers, online and offline prices are largely the same (Cavallo, 2017), and online prices help forecast and nowcast inflation (Aparicio and Bertolotto, 2020; Macias et al., 2023). Every such index needs each product assigned to a category of the official basket first; this paper is about that step.

Classification in official statistics. Eurostat's guides on scanner data and web scraping for the HICP treat classification as one of the main processing steps (Eurostat, 2017, 2020), and euro-area statistics have moved towards these sources (Eiglsperger, 2019). National statistical offices have classified products with machine learning: the ONS with clustering and decision trees (Metcalfe et al., 2016) and with semi-supervised label propagation into the items of the UK CPIH basket (Martindale et al., 2020), Statistics Netherlands with supervised classifiers, of which logistic regression performed best (Harms and Spinder, 2019), and the US Census Bureau with text classification of product descriptions for the Economic Census (Roberson, 2021). The Australian Bureau of Statistics reported that it was investigating machine learning to automate the classification of web-scraped products (Australian Bureau of Statistics, 2020), and a UNECE project surveyed such uses across offices (United Nations Economic Commission for Europe, 2021). These studies report accuracy; they do not translate it into the error of the resulting index.

Project Spectrum. The BIS Innovation Hub's Project Spectrum (BIS Innovation Hub, 2026) classifies the ECB's Daily Price Dataset into ECOICOP 2018 subclasses. It finds that text embeddings with classic classifiers reach accuracy comparable to LLM prompting at a small fraction of the cost and time (BIS Innovation Hub, 2026, p. 3), and reports accuracy weighted by CPI expenditure at the subclass level (BIS

Innovation Hub, 2026, pp. 22–24). It argues that, for nowcasting, errors at the subclass level usually matter less than errors at higher levels of the hierarchy, depending on how differently the confused categories' prices move (BIS Innovation Hub, 2026, p. 25), and it names two next steps: testing the approach on other data sets and languages, and building CPI indices from the classified data to back-test them against official inflation (BIS Innovation Hub, 2026, p. 30). This paper takes up the first for one non-European language and complements the second: it does not back-test a web-price index, but measures the error a given confusion matrix adds to measured inflation, in basis points. Our setting differs from Spectrum's in taxonomy, language, product universe and labels, so the numbers are not directly comparable.

Models. We use multilingual sentence encoders (Reimers and Gurevych, 2019; Wang et al., 2024; Chen et al., 2024) and a fine-tuned multilingual encoder (Conneau et al., 2020). LLMs have been shown to match or beat crowd workers on text annotation (Gilardi et al., 2023), and cascades that send only hard items to an expensive model can cut cost at little loss of accuracy (Chen et al., 2023). Calibrated confidence (Guo et al., 2017) makes selective classification possible, trading coverage for accuracy (Geifman and El-Yaniv, 2017); we weight both axes of that trade-off by CPI expenditure.

3 Data

All data are public, and each source has one job. The monthly indices and weights of the Trade Policy and Strategy Office define the target: the items a product must be assigned to and the weight each carries in headline, food and core inflation (Section 3.1). The same weights show which part of the index a product classifier can reach at all (Section 3.2). Product text comes from two places: Open Food Facts and Open Beauty Facts, crowd-sourced product names in many languages with weak category labels; and the Ministry of Commerce's price-collection specifications, Thai text that the office itself labels with item codes (Section 3.3). Specifications held out from training form the reference set on which every method is scored (Section 3.4). Table 1 in Appendix A lists the sources and their licences; the repository records the retrieval date and checksum of every file.

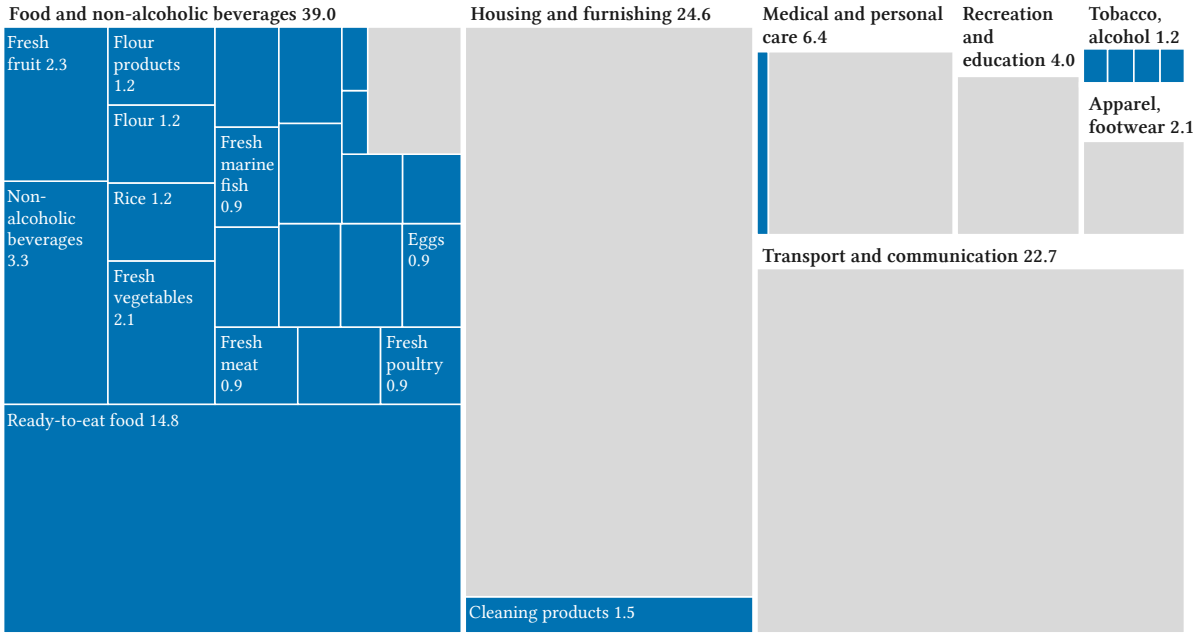
3.1 The Thai CPI

The Thai consumer price index is compiled monthly by the Trade Policy and Strategy Office (TPSO) of the Ministry of Commerce. The current base year is 2566 in the Buddhist era (2023), and the series on this base has been in use since January 2568 (2025) (Trade Policy and Strategy Office, 2025). Like most national CPIs it is a Lowe index (International Monetary Fund et al., 2020): expenditure weights from a base period, price-updated to a price reference month, here December 2567 (2024). TPSO's open API publishes monthly indices for the headline, divisions, groups and items (Trade Policy and Strategy Office, 2026), but no weights. The monthly release publishes price-updated weight shares for the official aggregates only (headline, the divisions, the food groups and the analytical series such as core CPI). We recover those shares from the releases and check that they add up across the hierarchy and reproduce the official headline from the published sub-indices.

Below the published aggregates, item weights are not public. We partition the item codes into *cells*: each official aggregate owns the items whose deepest published ancestor it is, and the cell weight is the aggregate's weight minus that of the published aggregates nested below it. Within a cell we estimate item weights in two ways and report both: an equal split (primary) and non-negative least squares (NNLS) on the Lowe identity, which chooses the split that best reproduces the cell's index from its items' indices (sensitivity). Neither is an official weight, and the paper never calls it one.

3.2 Scope: the items a product classifier can reach

A classifier of packaged-product text can only measure items sold as packaged products. We define 28 in-scope CPI items: food at home, ready-to-eat food, non-alcoholic beverages, alcoholic beverages,



Source: TPSO consumer price indices and weights (base 2566); author's calculations.

Note: Area: price-updated CPI weight (% of headline). Blue: the 28 items in the classifier's scope; grey: the rest of each division. The 16 smallest are unlabelled; the appendix lists every item. Item weights inside an official weight cell are an equal split (estimated).

Figure 1: The items a product classifier can reach are a large share of the Thai CPI basket, and most of that share is food. Area: price-updated CPI weight (% of headline), by division; blue, the in-scope items; grey, the rest of each division. Item weights inside an official weight cell are an equal split (estimated). Source: TPSO; author's calculations.

tobacco, personal care and cleaning products. Each is mapped to its COICOP 2018 class (United Nations Statistics Division, 2018). Everything else (rents, energy, transport, services, clothing, durable goods) is out of scope (OOS). In-scope items carry 40.5% of the headline weight under the equal split and 43.5% under NNLS, most of it food (Figure 1).

3.3 Product text

Open Food Facts. Open Food Facts (OFF) is a crowd-sourced, openly licensed database of packaged products (Open Food Facts contributors, 2026). We use its Parquet export and that of its sister project Open Beauty Facts. Our extract holds 76,269 products from 18 markets: Thailand, other Asian markets and a sample of European products. Of the 10,526 products tagged as sold in Thailand, 7,303 have a name in Thai script. Model input is the product's name, brand and quantity, joined as `name | brand | quantity`.

Weak labels. OFF products carry crowd-sourced category tags. A hand-curated crosswalk maps tags to CPI items, following TPSO's own item headings rather than intuition about COICOP (for example, snacks belong to flour products in the Thai CPI). This gives weak labels: 52,255 non-Thai products and 3,945 Thai products. Weak labels are used for training and development, never for evaluation.

Official specifications. The Ministry of Commerce (MOC) publishes the product specifications TPSO uses for price collection: short Thai descriptions with brand and pack size, each under an official item code (Ministry of Commerce, Thailand, 2026). They cover the whole basket, so specifications of items

outside our scope carry the label OOS. Each specification sits under a sub-item (a seven-digit code; jasmine rice under rice, say). We hold out whole sub-items: within every item, a fifth of its sub-items (at least one) form the test split, another fifth validation and the rest training, so that a test specification's sub-item is never seen in training; an item with a single sub-item is split by specification instead. We drop the 35 test specifications whose text duplicates a training one. This leaves 24,456 specifications for training and validation and 7,331 held out, 4,921 of them out of scope. The 2 in-scope items with no specifications in the published file, ready-to-eat food and sugar and confectionery, cannot be evaluated; the 26 evaluated items carry 24.9% of the headline weight under the equal split (40.5% for the full scope).

3.4 The reference set

Every method is evaluated on the reference set, drawn from the held-out specifications: labels assigned by the office that defines the classification, on text it wrote for its own price collection. The set holds every held-out in-scope specification and, because the out-of-scope class would otherwise be two thirds of it, a fixed random 1,500 of the held-out out-of-scope specifications, carried with the matching design weight so that every estimate still refers to the whole held-out list: 3,910 rows in all, over 26 items and OOS. Cascade thresholds and classifier settings are chosen on the training and validation splits only. The reference set was not sealed, however: the development grid logged its scores for every run, so they were visible while the methods reported here were chosen.

The original design called for a sample of Thai Open Food Facts products labelled by hand, blind, in a purpose-built tool, with a second pass to measure the human label ceiling. No such labels were collected; the tool and the held-out product sample remain in the repository for a later version, and Section 7 states what the official text cannot tell us.

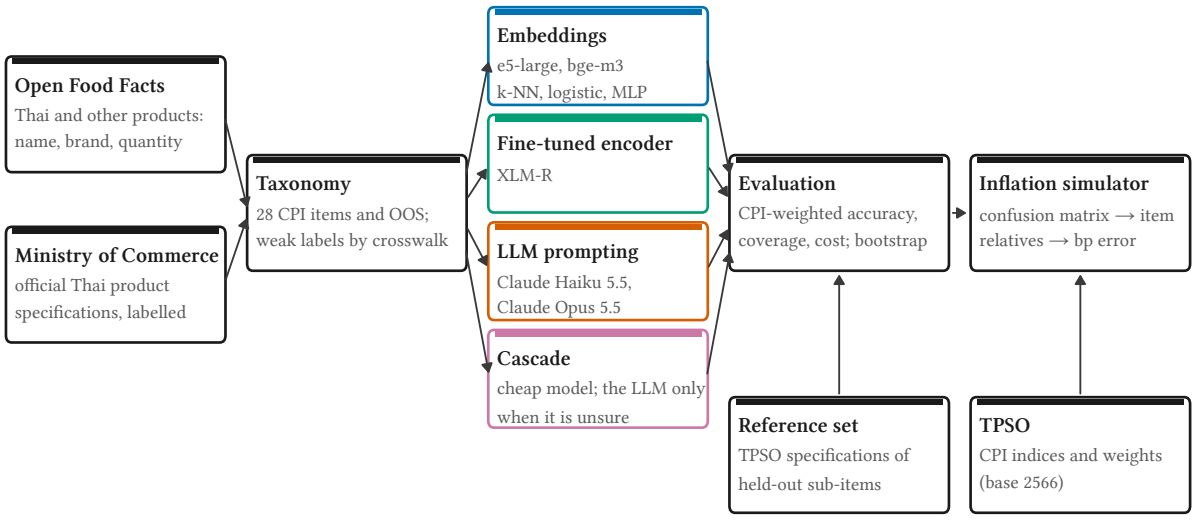
4 Methods

Figure 2 summarises the pipeline. Every method maps a product text to one of the 28 in-scope items or OOS, with a calibrated confidence; some may abstain. Section 4.1 describes the four families of classifier: light classifiers on frozen multilingual embeddings, a fine-tuned multilingual encoder, few-shot prompting of two commercial LLMs, and cascades that pass a cheap classifier's least confident products to an LLM. Section 4.2 defines how they are scored: accuracy weighted by CPI expenditure, the accuracy–coverage and cost–accuracy frontiers, and bootstrap intervals for paired comparisons. Section 4.3 builds the simulator that turns a confusion matrix into basis points of measured inflation.

4.1 Classifiers

Frozen multilingual embeddings. Following Spectrum's design (BIS Innovation Hub, 2026, pp. 3, 19–20), the first family embeds each text once and trains a light classifier on the vectors. We use two open multilingual embedding models run locally, e5-large (`intfloat/multilingual-e5-large`; Wang et al., 2024) and bge-m3 (`BAAI/bge-m3`; Chen et al., 2024). On top of each we fit three classifiers: cosine k-nearest neighbours with similarity-weighted votes, $k \in \{1, 3, 5, 10, 20, 50\}$ chosen on validation macro recall; multinomial logistic regression; and a one-hidden-layer perceptron (512 units, dropout 0.2). The last two share one trainer: Adam with learning rate 10^{-3} and weight decay 10^{-4}, class weights proportional to the inverse square root of class frequency, and early stopping on validation loss (patience 3).

Fine-tuned encoder. XLM-R base (`FacebookAI/xlm-roberta-base`; Conneau et al., 2020) with a classification head is fine-tuned end to end on the same training pools with the same class weights and early stopping (4 epochs at most, batch 64, AdamW at 3×10^{-5} with linear warm-up and decay, inputs truncated to 64 tokens).



Source: author's design.

Note: Every result in the paper is computed from public data by the code in the repository.

Figure 2: The pipeline runs from public product text to basis points of measured inflation. Data sources (left) feed a taxonomy and weak labels, four families of classifier, an evaluation on held-out official specifications and the inflation simulator. Source: author's design.

LLM prompting. We prompt two models of different cost, Claude Opus 5.5 and Claude Haiku 5.5, through the Anthropic API, one product per request. A fixed system prompt gives the instructions, the item list grouped by CPI group with short descriptions grounded in TPSO's item headings, and 2 official specifications per item drawn from the training split (Appendix C). The answer is constrained by a JSON schema: the model names the group, then an item inside it, and a verbalised confidence, the probability in per cent that the item is right. Refusals count as abstentions. Every response is cached and every billed call is logged with its tokens and cost; total spend was capped at USD 60 and came to USD 14.72.

Cascade. A cheap classifier answers when its calibrated confidence is at least a threshold t; otherwise the product is routed to an LLM. We pair the cheap classifiers with each LLM. For each pair, t is the threshold with the highest weighted accuracy on the Thai weak-label validation split, with ties broken by lower cost, chosen without reference-set scores; the reference set then scores that one threshold. The routed share, and with it the cascade's cost, is a share of the held-out list: each specification counts with its design weight $1/\pi_i$, the inverse of its probability of entering the reference set.

Training data and the learning curve (RQ1). Each embedding classifier is trained on a sequence of pools: non-Thai weak labels only (zero-shot cross-lingual transfer to Thai); plus Thai weak labels; plus $n \in \{100, 250, 500, 1{,}000\}$ or all official Thai labels; and official labels alone. The budget n counts every official label a run touches, validation included, so a small-budget run cannot tune on thousands of extra official labels. The n labels are the first n specifications of the official training and validation pool in a fixed pseudo-random order, not stratified by item, so smaller budgets are subsets of larger ones; four fifths of them train the classifier and one fifth is its validation set.

Calibration. Each run's logits are divided by one scalar temperature fitted on its own validation rows by negative log-likelihood (Guo et al., 2017); for k-NN the logits are log vote shares. The calibrated top-class probability is the confidence used by the coverage frontier and the cascade.

4.2 Evaluation

Expenditure-weighted accuracy. What matters for a price index is not how many products are classified correctly but how much CPI weight is measured correctly. Let w_k be the CPI weight of item k and r_k the recall of item k on the reference set, estimated with the design weights $1/\pi_i$ of Section 3.4 (Horvitz and Thompson, 1952). The main metric is

$$\mathrm{acc}_w = \frac{\sum_k w_k r_k}{\sum_k w_k}, \qquad r_k = \frac{\sum_{i:\, y_i=k} \mathbb{1}[\hat{y}_i = k]/\pi_i}{\sum_{i:\, y_i=k} 1/\pi_i}, \tag{1}$$

the CPI-weighted mean recall over the evaluated items, a weighted version of Spectrum's in-class recall (BIS Innovation Hub, 2026, p. 22). Every held-out in-scope specification is in the reference set, so $\pi_i = 1$ there and the design weights matter only for the whole-list accuracy and the routed share. We also report accuracy over the whole held-out list, which counts OOS specifications, and macro-F1. Abstentions are errors at full coverage. acc_w is a recall: it does not penalise out-of-scope products sent into an item, which the simulator of Section 4.3 does count.

Coverage, cost and uncertainty. Sweeping a confidence threshold traces a weighted accuracy–coverage frontier (Geifman and El-Yaniv, 2017): the CPI-weighted share of products answered against acc_w on those. Cost per million products is the measured mean API cost per call plus local compute (measured seconds per item at an assumed USD 1.00 per GPU-hour). All intervals come from 1,000 bootstrap resamples of the reference set, stratified by item (Efron and Tibshirani, 1993). Because a method tends to get a sub-item all right or all wrong, each resample draws whole sub-items within an item: the 2,410 in-scope specifications fall in 54 sub-items, and 13 items have a single sub-item in the reference set; their specifications are resampled one by one, so their share of the uncertainty is understated. Method comparisons are paired: every method is scored on the same resamples, and the list of comparisons is fixed in code.

4.3 From confusion matrices to basis points (RQ3)

Setting. For an aggregate A with member items i, the Lowe index satisfies

$$\frac{I_{A,t}}{I_{A,L}} = \sum_{i \in A} \frac{w_i}{W_A} R_{i,t}, \qquad R_{i,t} = \frac{I_{i,t}}{I_{i,L}}, \tag{2}$$

where L is the price reference month, w_i are price-updated weights and $W_A = \sum_{i \in A} w_i$. We take the official item relatives $R_{i,t}$ as the true price movement of every product in item i. This isolates the error due to classification: a real web-price index would add sampling, matching and quality-adjustment error on top.

Misclassification. Let n_k be the universe share of products whose true item is k and P a method's confusion matrix, P_{kj} the probability that a product of item k is classified as j. Item j is then measured from a mixture of the products sent to it:

$$\tilde{R}_{j,t} = \sum_k \phi_{kj} R_{k,t}, \qquad \phi_{kj} = \frac{n_k P_{kj}}{\sum_{k'} n_{k'} P_{k'j}}. \tag{3}$$

Products sent to OOS or abstained on are dropped, and so are products sent to the two in-scope items that have no specifications (Section 7), since those items are held at their official values; none of the main methods sends more than a handful of reference specifications there. If nothing flows into an item, its weight is redistributed within its CPI group. True OOS products that leak into an item move with the official indices of their own items: the design-weighted mix of the out-of-scope specifications

that the method sent to that item, or of all out-of-scope specifications if it sent none there. Out-of-scope items are measured perfectly. The measured index is

$$\hat{I}_{A,t} = I_{A,t} + I_{A,L} \sum_{j \in A} \frac{w_j}{W_A} \big(\tilde{R}_{j,t} - R_{j,t}\big), \tag{4}$$

and the error is reported in basis points of month-on-month and year-on-year inflation for the headline index, the food index and a core proxy: TPSO's core CPI excludes raw food and energy, and in the simulator every in-scope item outside the four raw-food groups counts as core, so the series is a proxy for it. A perfect classifier gives $\tilde{R} = R$ and reproduces every official series exactly, which is a unit test of the code.

Product universe. The shares n_k are not observed: the reference set is the office's list, not a stream of products. We state them as an assumption and report two universes. In the primary one, products arrive in proportion to CPI expenditure weights, with the weight outside the evaluated items arriving as out-of-scope products; in the alternative they arrive in proportion to the office's specification list, that is, the design-weighted shares of the reference set. In-scope items without specifications are held at their official values, like out-of-scope items.

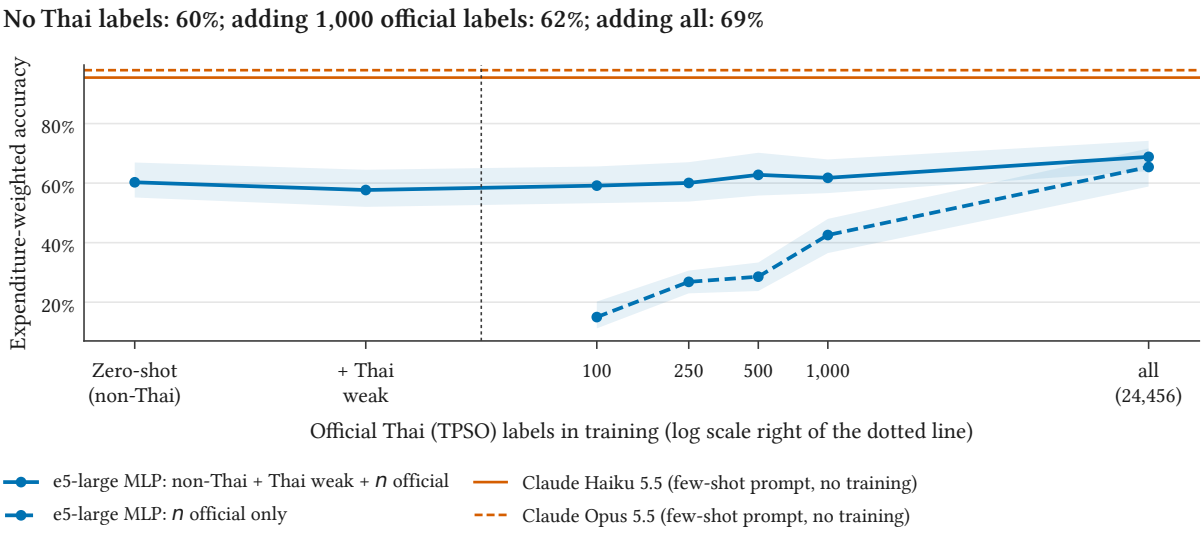
Decomposition. Since $\sum_k \phi_{kj} = 1$, $\tilde{R}_{j,t} - R_{j,t} = \sum_k \phi_{kj}(R_{k,t} - R_{j,t})$. The contribution of each confusion $k \to j$ is therefore the item weight times the misallocated mass times the gap between the two items' price movements. A confusion matters only if it is frequent, falls on a heavily weighted item and joins items whose prices move differently.

Uncertainty. The confusion counts come from the reference set. We draw 1,000 confusion matrices from a Dirichlet posterior, row by row. Sparse rows need a prior, and a flat one is not neutral: it sends mass to arbitrary items and creates error even for a perfect classifier. We therefore use an empirical-Bayes prior worth $\alpha = 1$ product per row, whose mean puts the method's pooled accuracy on the diagonal and spreads the rest over the columns where that method's errors actually go. Under this prior a perfect classifier has zero error in every draw. A variant adds product-level sampling noise: each product's log price drifts from its item's as a random walk with an assumed annual standard deviation of 5% or 10%, and item j averages the products sent to it, the size of the held-out list standing in for the number of products. A perfect classifier with the same noise gives the error of the noise alone, the baseline for that variant.

The Dirichlet draws treat specifications as independent, but a method's errors cluster: the reference set is made of whole held-out sub-items, and a method tends to get a sub-item all right or all wrong. We therefore also resample the 112 sub-items of the reference set (54 in scope, the rest out of scope) with replacement, 1,000 times, without stratifying by item, so an item with a single sub-item can be absent from a resample; each resample is smoothed with the same prior and simulated. Neither interval covers the assumptions of the simulator; Section 5.3 varies those separately. For each aggregate we report the root mean squared error of the inflation rate over the simulated months. The point estimate uses the posterior-mean confusion matrix and the interval is the 95% range of the RMSE across draws or resamples. The point uses the posterior-mean matrix, which averages away the scatter of the draws, so where a method's errors rest on few observations the point can sit near the lower end of the interval; there the upper end is the better guide.

5 Results

This section takes the three questions in turn, and each subsection opens with its answer. In brief: transfer from non-Thai products alone gets an e5-large MLP to 60.3% of the evaluated weight, and all official labels raise it to 68.8% (RQ1, Section 5.1). The LLMs are far ahead: Claude Opus 5.5 reaches



Source: TPSO consumer price indices and weights (base 2566); Open Food Facts (ODbL); TPSO price-collection specifications (MOC open data; official labels, held-out split); author's calculations.

Note: Budgets count every official label a run touches (training and validation). Bands: 95% stratified bootstrap.

Figure 3: Transfer from non-Thai products gives 60.3% weighted accuracy with no Thai label; adding 1,000 official labels gives 61.8% and adding all of them 68.8%. e5-large MLP by training pool: zero-shot (non-Thai weak labels), plus Thai weak labels, plus n official labels (solid), and n official labels alone (dashed). Horizontal lines: the two LLMs, prompted with 2 official examples per item. Bands: 95% stratified bootstrap. Source: Open Food Facts; MOC; TPSO; author's calculations.

97.9%, and a cascade from the e5-large MLP to Claude Opus 5.5 scores 97.8% at 27% of its cost on the held-out list (RQ2, Section 5.2). With Claude Opus 5.5's confusion matrix, headline year-on-year inflation would be misstated by 1.9 basis points, against 1.9 to 10.4 for cheaper classifiers trained on all labels, and accuracy alone does not rank the methods by this error (RQ3, Section 5.3).

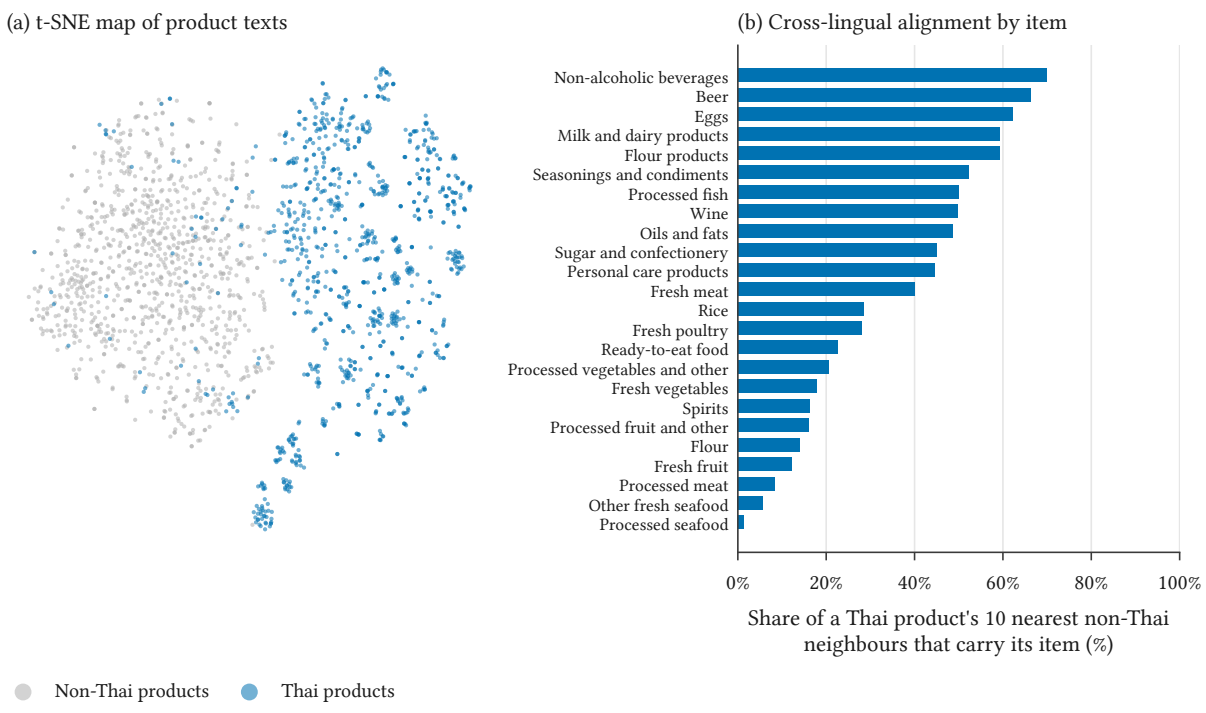
All scores are on the reference set (Section 3.4): 3,910 held-out official specifications, 2,410 of them in scope. Weighted accuracy is acc_w of equation (1) under the equal split; brackets give 95% intervals from 1,000 bootstrap resamples that draw whole sub-items within each item (Section 4.2). The NNLS split is reported where it changes a conclusion.

5.1 RQ1: cross-lingual transfer and the value of official labels

Answer. Transfer from non-Thai products does most of the work. An e5-large MLP trained only on non-Thai products classifies 60.3% [55.2, 66.9] of the evaluated CPI weight correctly. Thai weak labels do not help, budgets of up to 1,000 official labels add little, and all 24,456 official labels raise it to 68.8% [64.0, 74.2].

Figure 3 traces the learning curve. With Thai weak labels added the score is 57.7%; with 1,000 official labels on top of the weak ones it is 61.8%. Only the full official set moves it by more than the sampling error: the paired gain of all official labels over zero-shot transfer is 8.5 percentage points [2.3, 14.3]. That gain depends on how the weight is spread inside the official cells: under the NNLS split it is 1.5 points [−4.2, 6.2], and the interval includes zero. Official labels on their own need the whole set: with 1,000 of them the classifier reaches only 42.6%, and with all of them 65.4%. For the other encoders transfer does less: zero-shot, the bge-m3 MLP reaches 50.7% and the fine-tuned XLM-R 41.2%. Official labels add much more for them, and with every label the three end within a point of one another (68.8%, 69.2% and 69.5%).

Plain accuracy hides all of this. The held-out list is mostly out-of-scope specifications, and a classifier that answers oos to almost everything scores well on it: the e5-large MLP trained on a hundred official labels alone is right on 75.5% of the list but on only 15.0% of the evaluated CPI weight. Conversely,



Source: Open Food Facts (ODbL); TPSO price-collection specifications (MOC open data); author's calculations.

Note: Labels: official (TPSO) or weak (Open Food Facts crosswalk); reference set excluded. Items with no non-Thai products are omitted. The map is for display only.

Figure 4: Thai products sit next to non-Thai products of the same item for some items and not for others. (a) t-SNE map (van der Maaten and Hinton, 2008) of a balanced sample of product texts (display only). (b) Cross-lingual alignment: the share of a Thai product's 10 nearest non-Thai neighbours, in the full e5-large space, that carry its item. The reference set is excluded. Source: Open Food Facts; MOC; author's calculations.

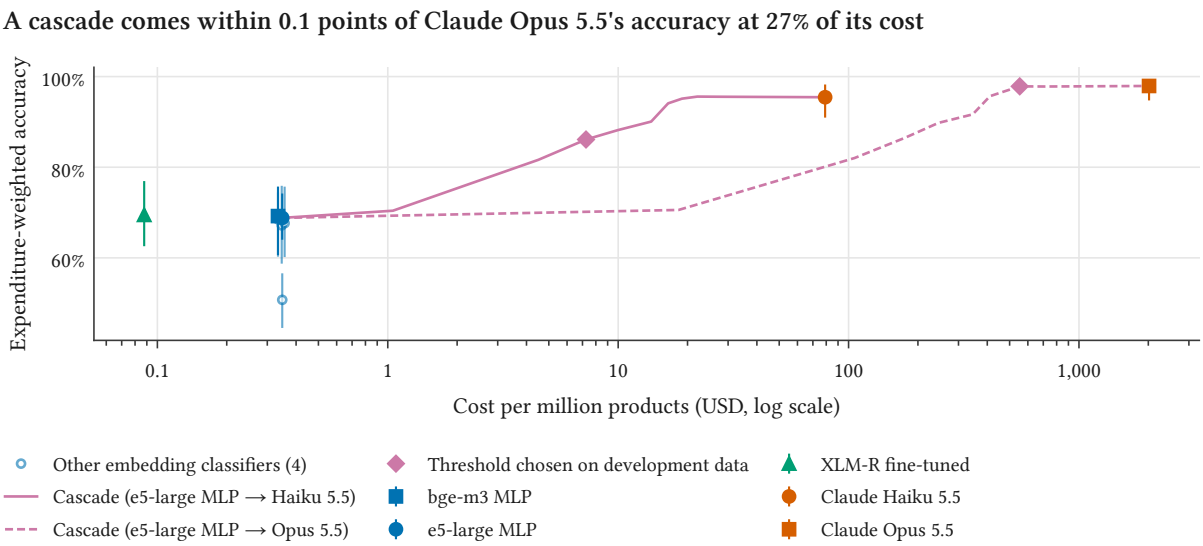
the zero-shot classifier is right on only 16.0% of the list because it was never shown an out-of-scope Thai text, yet it measures most of the weight correctly. A metric weighted by expenditure is not a refinement here; it reverses the ranking.

The input fields matter less than the labels. Removing the brand or the quantity from the Open Food Facts training texts (the official specifications keep theirs) changes weighted accuracy by amounts we cannot tell from zero: 2.9 points [−1.2, 6.6] for the brand and −0.2 [−2.6, 3.3] for the quantity, and under the NNLS split 1.2 [−1.7, 4.2] and 2.1 [−0.2, 5.1]. Each variant is a single training run, so the intervals leave out training-seed variation. Figure 4 shows why transfer works for some items and not for others: in the embedding space, Thai beverages, eggs and dairy products sit among non-Thai products of the same item, while Thai processed seafood, other fresh seafood and processed meat do not.

5.2 RQ2: accuracy, coverage and cost

Answer. On official Thai text the LLMs are far ahead (the specifications are public and may have been in the LLMs' pre-training data; Section 7). Claude Opus 5.5 classifies 97.9% [94.7, 99.3] of the evaluated weight correctly and Claude Haiku 5.5 95.4% [90.9, 98.3], against 69.5% for the best cheap method with the full input (the fine-tuned XLM-R; one text-ablation run, the bge-m3 MLP without the brand, scores 69.9%, a single training run). A cascade that sends the e5-large MLP's unsure products to Claude Opus 5.5 scores 97.8% at 27% of its cost on the held-out list.

The gap between the LLMs and every cheap method is large and certain: Claude Haiku 5.5 is ahead of the e5-large MLP by 26.6 points [19.4, 32.1], of the bge-m3 MLP by 26.2 and of the fine-tuned XLM-R by 26.0. Claude Opus 5.5 is ahead of Claude Haiku 5.5 by 2.5 points [0.3, 5.3] and costs 25 times as much:



Source: Open Food Facts (ODbL); TPSO price-collection specifications (MOC open data; official labels, held-out split); Anthropic API price list; author's calculations.

Note: Full coverage throughout. API cost measured per call; local compute at an assumed USD 1.00 per GPU-hour. Lines: the cascade over its thresholds; at the chosen threshold the Claude Opus 5.5 cascade routes 27% of the held-out list (design-weighted) to the LLM; the cost share depends on that product mix.

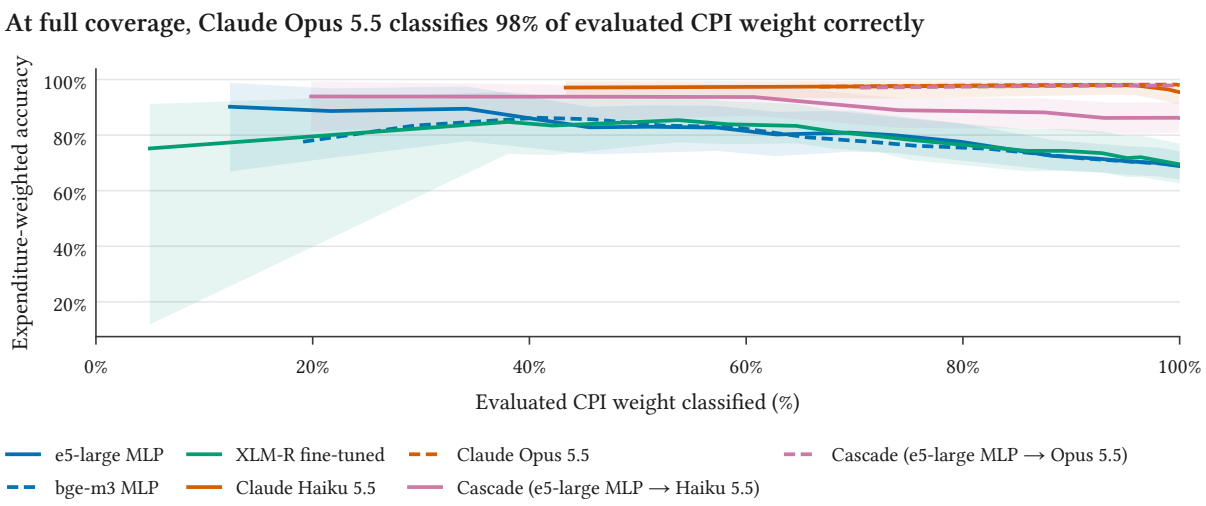
Figure 5: A cascade from the e5-large MLP to Claude Opus 5.5 comes within sampling error of Claude Opus 5.5 at 27% of its cost on the held-out list; no cheap method comes close to either LLM. Weighted accuracy at full coverage against cost per million products (log scale). Lines: the two e5-large cascades over their candidate thresholds; diamonds: the threshold chosen on development data. API cost measured per call; local compute at an assumed USD 1.00 per GPU-hour. Source: Open Food Facts; MOC; TPSO; Anthropic API price list; author's calculations.

USD 2,014.73 per million products against USD 79.12, while the embedding classifiers cost well under a dollar (USD 0.35 for the e5-large MLP; Figure 5). Under the NNLS split the scores of the main methods fall, because NNLS moves weight onto the item with the largest LLM error (personal items, below), but the LLMs stay ahead (Claude Opus 5.5 91.8%, Claude Haiku 5.5 90.6%, e5-large MLP 59.9%).

Cascades work when the threshold is set where the cheap model is reliable. Every threshold was chosen on the development data, the Thai weak-label validation split, which is Open Food Facts text. For the e5-large MLP and Claude Opus 5.5 it chose a confidence threshold of 0.95. Counted with its design weights, 27% of the held-out list went to the LLM (72% of in-scope specifications and 6% of out-of-scope ones, so the cost share rises with the in-scope share of the products). The saving comes mostly from out-of-scope specifications, which the cheap model keeps; on a stream of packaged products the cost share would be closer to the in-scope routed share. The cascade scored 97.8%, a paired difference from Claude Opus 5.5 of −0.1 points [−2.6, 0.4]. The development set, with its weak labels, rates the LLMs themselves far below their reference-set scores (Claude Opus 5.5 76.6%, Claude Haiku 5.5 71.2%), so it cannot see what routing to an LLM is worth: for Claude Opus 5.5 the highest candidate thresholds tied and the cost tie-break chose the lowest of them, and for Claude Haiku 5.5 it preferred thresholds that route few products. The other cascades were given lower thresholds, routed fewer products and differed from the LLM they call by between −6.4 and −9.3 points. Confidence calibrated on one kind of text was too high on the other: a cheap classifier is confidently wrong on sub-items it has never seen.

Abstaining does not close the gap. Figure 6 trades coverage for accuracy: the cheap methods gain accuracy as they leave their least confident products unclassified, but even on the small share of weight they are surest about they stay below the LLMs' accuracy at full coverage.

Where the errors are. The cheap methods miss whole items. The fine-tuned XLM-R misses every held-out specification of fresh meat, freshwater fish, processed fish and tobacco (each a small item,



Source: TPSO consumer price indices and weights (base 2566); Open Food Facts (ODbL); TPSO price-collection specifications (MOC open data; official labels, held-out split); author's calculations.

Note: Accuracy is on the classified products; those below a confidence threshold are left unclassified. Bands (first method of each family): 95% stratified bootstrap (1,000 resamples). Weights: equal split within residual cells.

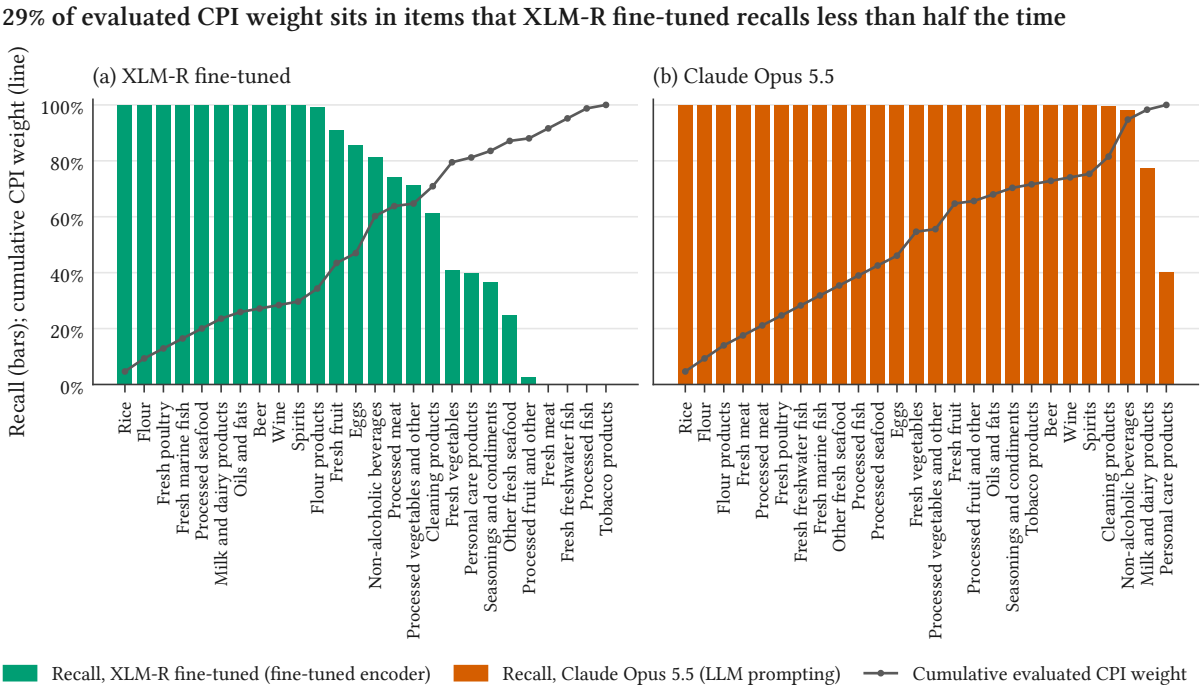
Figure 6: At full coverage Claude Opus 5.5 classifies 97.9% of the evaluated CPI weight correctly; the cheap methods stay below the LLMs at any coverage. Weighted accuracy on the classified products against the evaluated CPI weight classified, as the confidence threshold rises. Bands (first method of each family): 95% stratified bootstrap. Source: Open Food Facts; MOC; TPSO; author's calculations.

with 1 to 17 held-out specifications, so these are thin evidence one by one), sends most personal-care specifications to oos, and confuses seasonings with beverages and cleaning with personal-care products (Figures 7 and 8); its 897 in-scope errors fall on 24 sub-items. The LLMs' errors are concentrated. Claude Opus 5.5's 424 in-scope errors fall on 6 sub-items, and 399 of them (94%) on three; Claude Haiku 5.5's largest three are the same. The TPSO item whose Thai name means "personal items" covers wristwatches, school bags and hair dryers as well as soap and toothpaste. Our label description, written from the item headings, named only personal-care goods, and both LLMs sent the held-out wristwatches, school bags and hair dryers to oos. The next largest error is margarine, which TPSO counts under milk and dairy products and our description listed under oils and fats; the LLMs followed the description. These are not failures to read Thai. They are gaps in our documentation of the official item contents, and better documentation is the likely remedy; we did not test it. We did not change the descriptions after seeing these errors, since that would tune on the reference set. Because NNLS gives the personal-items item a much larger share of its cell's weight than the equal split does, it is this one gap that lowers the LLM scores under NNLS.

5.3 RQ3: classification error in basis points

Answer. With Claude Opus 5.5's confusion matrix, measured headline inflation would differ from official inflation by 1.9 basis points (RMSE of the year-on-year rate), food inflation by 3.2 and the core proxy by 2.6. The 95% interval for the sampling of specifications reaches 2.6 basis points for the headline; resampling whole sub-items, within which errors cluster, widens it to 3.7, and to 7.9 under the NNLS split. The cheap methods trained on all labels misstate headline inflation by between 1.9 and 10.4 basis points. The LLMs' lead is clear in accuracy; in basis points it is large on food against two of the three cheap methods, but once errors are allowed to cluster the intervals overlap. Accuracy does not rank methods by this error: two cheap methods with about the same weighted accuracy have point estimates that differ several-fold in basis points, though the intervals overlap.

The simulator covers 10 year-on-year rates (December 2025 to September 2026) and 21 month-on-month rates (January 2025 to September 2026). Figure 9 gives the year-on-year error of every main



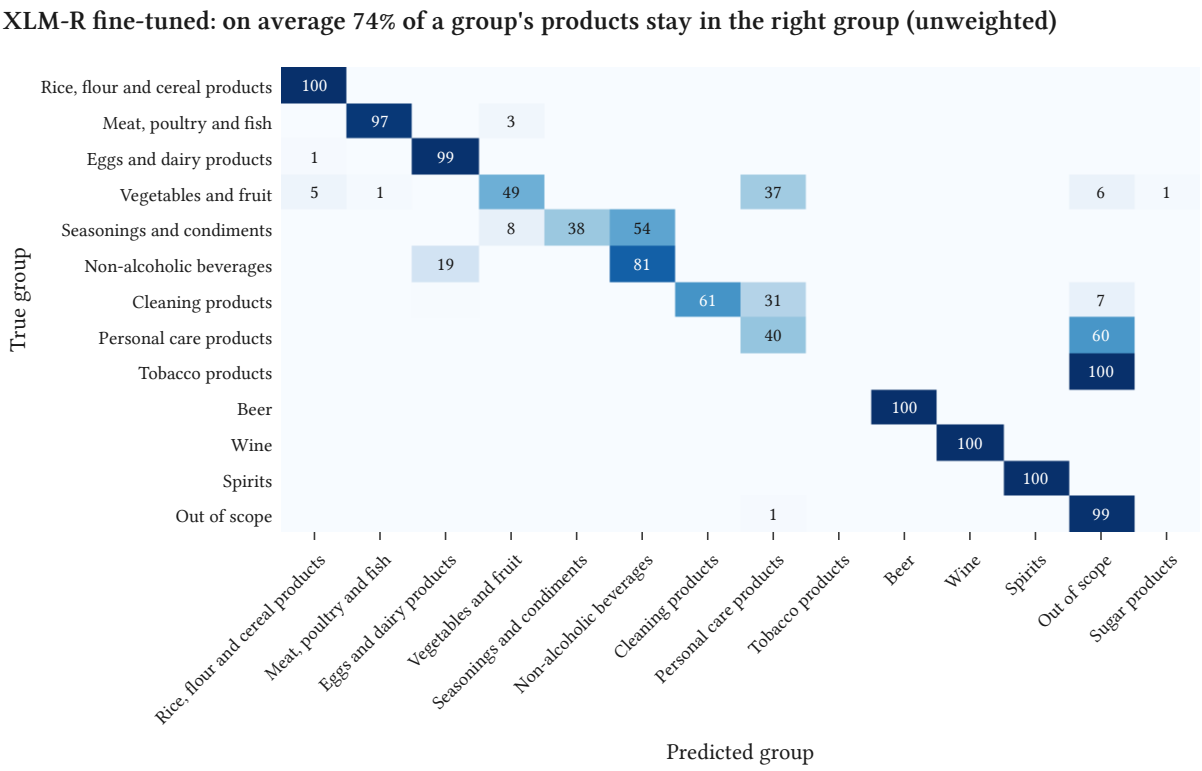
Source: TPSO consumer price indices and weights (base 2566); Open Food Facts (ODbL); TPSO price-collection specifications (MOC open data; official labels, held-out split); author's calculations.

Note: Bars: share of each item's held-out official specifications classified correctly; grey line: cumulative evaluated CPI weight in bar order. Items without specifications are omitted. Bar colours are the method families used in every figure.

Figure 7: The fine-tuned XLM-R misses whole items; Claude Opus 5.5 misses large parts of only two. Recall of each item on the reference set, ranked, with the cumulative evaluated CPI weight in bar order (line). Items without specifications are omitted. Source: MOC; TPSO; author's calculations.

method. Both LLMs and the e5-large–Claude Opus 5.5 cascade err by about two to five basis points on every aggregate; on month-on-month inflation Claude Opus 5.5's headline error is 0.5 basis points. The errors lean one way: Claude Opus 5.5's mean headline error is 1.5 basis points, so measured inflation would on average be slightly too high. For the e5-large MLP and the fine-tuned XLM-R, food is where errors bite: 24.2 basis points for the e5-large MLP [17.0, 31.6] and 22.3 for the fine-tuned XLM-R, whose errors are one-sided (a mean of −19.5 basis points): it understates food inflation in every month (Figure 10). The core proxy (Section 4.3) reverses the ranking. It excludes raw food, where most of the cheap methods' costly confusions lie, and keeps oils and fats and personal care, where the LLMs' confusions lie: the fine-tuned XLM-R's core error is 0.8 basis points [0.7, 1.9] against 2.6 for Claude Opus 5.5 under the equal split; under the NNLS split the two are close (3.3 against 3.8). Zero-shot transfer, which RQ1 found to get most of the weight right, misstates headline inflation by 14.7 basis points and food inflation by 38.7: getting most of the weight right does not bound the error from the rest, nor from the out-of-scope products pulled into the basket, which acc_w does not count.

The bge-m3 MLP shows why accuracy is not enough. Its weighted accuracy is about the same as the e5-large MLP's (69.2% against 68.8%), but its point estimate of the headline error is 1.9 basis points against 10.4. What matters is not how much weight is misclassified but whether the confused items' prices diverged. The interval for the bge-m3 MLP is wide (1.2–6.1) and the point is at its lower end, so we would not choose a classifier on this point alone. The decomposition of Section 4.3 shows where the basis points come from (Figure 11): for the fine-tuned XLM-R, fresh vegetables classified as fresh fruit contribute more than any other confusion, and confusions in opposite directions partly cancel. The decomposition is of the index level; a year-on-year error is the difference of two level errors, so its sign can differ from that of the largest contribution. For Claude Opus 5.5 the two largest contributions are out-of-scope products classified as personal care, and margarine counted under oils and fats. The



Source: TPSO consumer price indices and weights (base 2566); Open Food Facts (ODbL); TPSO price-collection specifications (MOC open data; official labels, held-out split); author's calculations.

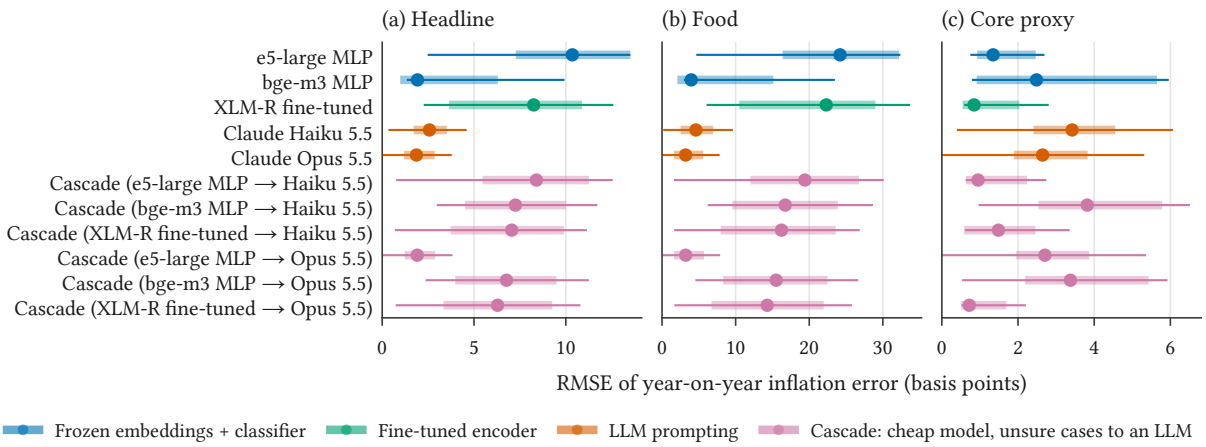
Note: Cells: % of the true group's held-out official specifications; rows sum to 100.

Figure 8: The best cheap method with the full input keeps rice and alcoholic beverages in the right group but sends tobacco and most personal-care products out of scope, sends many vegetables and fruit to personal care, and mixes seasonings with beverages and cleaning with personal-care products. Row-normalised confusion between CPI groups for the cheap method with the full input and the highest weighted accuracy (fine-tuned XLM-R); cells are % of the true group's held-out specifications. Source: MOC; TPSO; author's calculations.

wristwatches, school bags and hair dryers sent to OOS contribute nothing: a product sent out of scope leaves measurement without entering another item, so in the simulator it lowers accuracy but adds no bias.

Robustness. Four assumptions move these numbers. *The weights inside cells:* under the NNLS split Claude Opus 5.5's headline error is 2.7 basis points [1.8, 3.6], the e5-large MLP's 14.3 and the bge-m3 MLP's 2.8. *Clustered errors:* resampling sub-items gives Claude Opus 5.5 a headline interval of 0.0–3.7 and the bge-m3 MLP one of 1.4–9.9. *The product universe:* if products arrive in proportion to the office's specification list rather than to CPI weights, Claude Opus 5.5's headline error is 4.4 basis points and the e5-large MLP's 7.4. *Product-level noise:* if each product's price drifts from its item's with an annual standard deviation of 10%, Claude Opus 5.5's headline error rises to 4.7 basis points [1.7, 11.5]. Most of that is not classification: a perfect classifier with the same noise has an error of 4.3, the cost of measuring each item from a finite sample of products; its level depends on the assumed number of products, so only comparisons between classifiers under the same noise are meaningful. At 5% the two are 2.8 and 2.2. Without noise a perfect classifier has an error of 0.0, as it must; that is a unit test of the simulator, not a finding.

Basis points need accuracy alongside. A product sent out of scope leaves measurement, and without product-level noise an item measured from a few correctly classified products is measured exactly.



Source: TPSO consumer price indices and weights (base 2566); Open Food Facts (ODbL); TPSO price-collection specifications (MOC open data; official labels, held-out split); author's calculations.

Note: Dots: posterior-mean confusion matrix from the held-out official specifications. Thick lines: 95% interval over 1,000 Dirichlet draws of the matrix (specifications independent); thin lines: 95% interval over resampled official sub-items (errors clustered). Products in proportion to CPI weights. Out-of-scope items and items without specifications held at official values. Panels have their own scales.

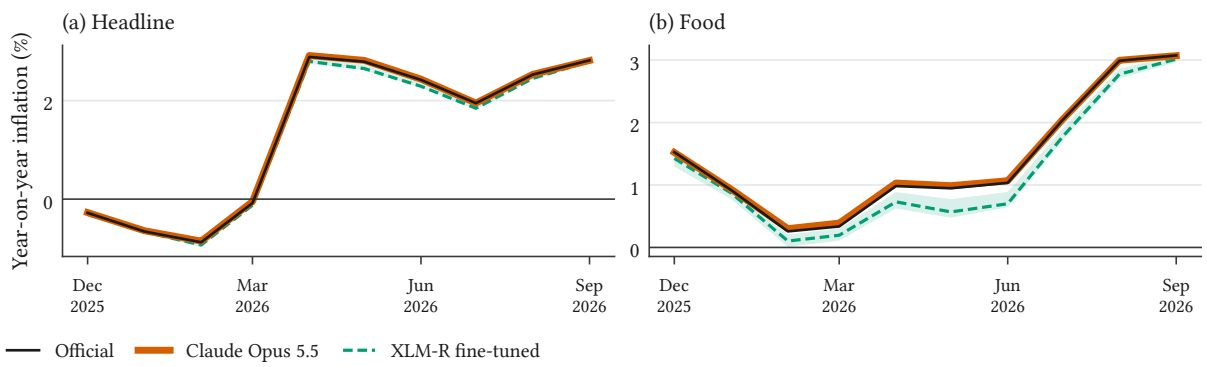
Figure 9: Claude Opus 5.5 misstates headline year-on-year inflation by at most 3.7 basis points (RMSE, upper end of the wider 95% interval); for every cheap method trained on all labels the upper end is higher. RMSE of the year-on-year inflation error over 10 months, headline, food and the core proxy. Dots: posterior-mean confusion matrix; thick lines: 95% interval over 1,000 Dirichlet draws (specifications independent); thin lines: over 1,000 resamples of the 112 official sub-items (errors clustered). Products in proportion to CPI weights; out-of-scope items and items without specifications held at official values. Source: MOC; TPSO; author's calculations.

A classifier that sends nearly everything out of scope therefore looks good in basis points: the e5-large linear classifier trained on a hundred official labels alone gets only 7.3% of the weight right, yet its headline error is 0.8 basis points. Product-level noise exposes it: at 10% its error is 30.8 basis points, against 4.7 for Claude Opus 5.5 and 4.3 for a perfect classifier. Neither metric ranks methods alone; a method should be chosen on both.

6 Discussion: what this means for central banks in Asia

Central banks and statistical offices in Asia that build price measures from web and scanner data work in languages for which little labelled product text exists. The Thai results suggest five points for that work. Each holds for one index, one language and the office's own text, and should be checked on other data before it is relied on.

Basis points are the unit that matters, alongside accuracy. Weighted accuracy reversed the ranking that plain accuracy gave; even weighted accuracy did not rank methods by their effect on measured inflation; and basis points alone flatter a classifier that measures little of the weight. Two classifiers that get 68.8% and 69.2% of the weight right have point estimates that differ several-fold in basis points, though their intervals overlap. The simulator of Section 4.3 needs only a confusion matrix on held-out labels, the published weights and the published item indices. An office that already publishes an index can therefore state the classification error of a new source in the unit its users care about before the source enters production. The decomposition says which confusions to fix first.



Source: TPSO consumer price indices and weights (base 2566); Open Food Facts (ODbL); TPSO price-collection specifications (MOC open data; official labels, held-out split); author's calculations.

Note: Simulated with each method's confusion matrix on the held-out official specifications, products in proportion to CPI weights; bands: 95% over Dirichlet draws. Out-of-scope items and items without specifications held at official values. Panels have their own scales.

Figure 10: The fine-tuned XLM-R would understate food inflation in every month; Claude Opus 5.5 tracks the official series. Official and simulated year-on-year inflation, headline and food. Bands: 95% intervals over Dirichlet draws. Products in proportion to CPI weights. Source: MOC; TPSO; author's calculations.

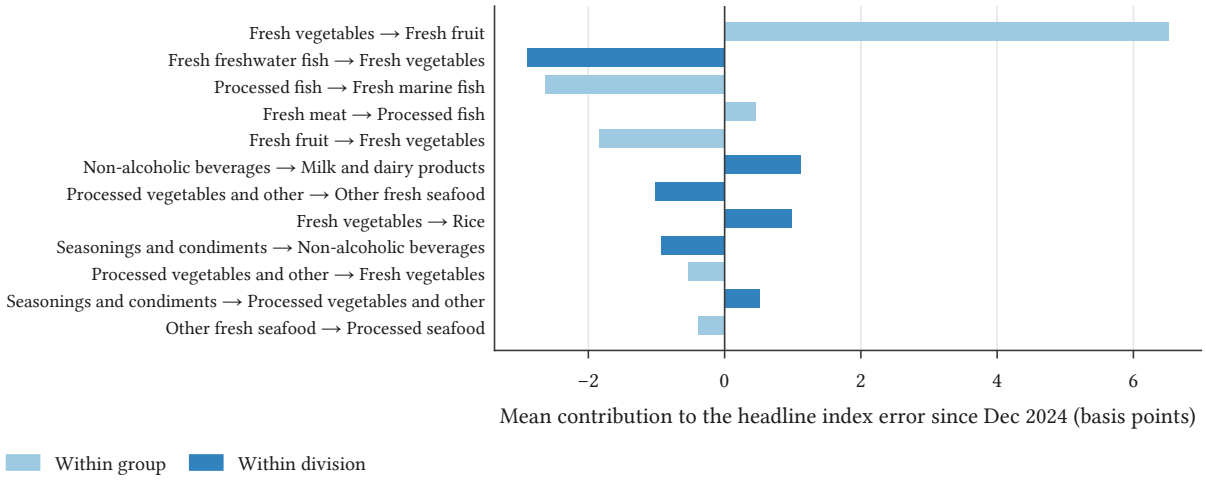
A published specification list is a public good. The most useful data in this paper are the Ministry of Commerce's price-collection specifications: Thai descriptions with official item codes, released as open data. They served as training labels, as development data and as the reference set, and without them no number in Section 5 could have been computed. Where an office publishes such a list, every researcher and vendor has a common benchmark in the local language at no cost to the office.

LLM classification is affordable at the scale of a price index. Claude Haiku 5.5 classified official Thai text with 95.4% weighted accuracy at USD 79.12 per million products, and the e5-large–Claude Opus 5.5 cascade came within sampling error of Claude Opus 5.5 at USD 553.25 per million products of the held-out list's mix. The whole set of LLM runs for this paper cost USD 14.72. The question is then one of governance more than budget: closed models change, responses must be cached and versioned, and an office may prefer an open model it can run itself. The cheap methods were not close enough to make that choice free: for every cheap method trained on all labels, the upper end of the interval for the headline error was above the LLMs'.

Documentation of the basket may matter as much as the model. The LLMs' largest errors came from two items whose official contents differ from what their headings suggest: TPSO's personal items include wristwatches, school bags and hair dryers, and its milk and dairy products include margarine. A classifier can only follow the definition it is given. A description of what each item includes and excludes, maintained with the classification, might remove most of these errors for the LLMs; this paper does not test that, and whether such a description can be published is a decision for each office.

Transfer gets a long way, but not all the way. With no Thai label at all, non-Thai products taught an embedding classifier to place 60.3% of the evaluated weight correctly, and the alignment of Thai products with their non-Thai neighbours varied widely by item. For languages with even fewer resources than Thai, this is a usable starting point and a guide to where official labels are needed first: in the items where local products look least like foreign ones. But the error of the zero-shot classifier in basis points was large, and an office would need at least a held-out labelled sample in its own language to know how large before it publishes anything built on it.

Two cautions apply to all five points. The reference text here is the office's own short descriptions;



Source: TPSO consumer price indices and weights (base 2566); Open Food Facts (ODbL); TPSO price-collection specifications (MOC open data; official labels, held-out split); author's calculations.

Note: Contribution = item weight × misallocated product share × gap between the true and the predicted item's price relative (true → predicted). Posterior-mean confusion matrix on the held-out official specifications; products in proportion to CPI weights.

Figure 11: A handful of confusions carry the headline error of the fine-tuned XLM-R, and fresh vegetables classified as fresh fruit is the largest. Mean contribution of each confusion to the headline index error since the price reference month: item weight times misallocated product share times the gap between the two items' price relatives. Shade: the lowest level of the classification the two items share. Source: MOC; TPSO; author's calculations.

shop text is longer and noisier, and accuracy on it will be lower. And the basis-point figures assume a product universe; a real stream of web products would have its own mix. A labelled sample of the text to be classified, drawn with known probabilities, would resolve both and is the natural next step for any office that applies these methods.

7 Limitations

The reference text is the office's, not the shop's. The reference labels come from TPSO's price-collection specifications: short, regular descriptions written by the office for its own collection. Shop and database texts are longer, noisier and more varied. Accuracy on the reference set is therefore accuracy on the office's vocabulary; a classifier deployed on web or scanner data would face harder text, and the learning curve measures the value of official labels, not of labelled shop data. The specifications are public and may have been seen by the LLMs in pre-training.

Two items cannot be evaluated. Ready-to-eat food, the heaviest in-scope item, and sugar and confectionery have no specifications in the published file. Their classification error is unknown, the simulator holds them at official values and drops the products a method sends to them, as if it had abstained, and the basis-point figures cover 24.9% of the headline weight rather than 40.5%. Items with few specifications have wide intervals, and the 13 items with a single held-out sub-item have intervals that are too narrow, since a resample cannot draw a sub-item they do not have.

No human label ceiling. The original design included a sample of Thai products labelled by hand with a blind second pass, to measure how consistent a labeller is and to set that against the human error rate Spectrum reports (BIS Innovation Hub, 2026, p. 15). No such labels were collected, so this paper

does not measure the label ceiling and takes the official labels as correct. Official codes can be wrong or ambiguous too; a method that disagrees with the office is occasionally right.

The label descriptions are ours. Every classifier sees the labels only through names and short descriptions that we wrote from TPSO's item headings, and the LLMs see nothing else. Most of the LLMs' errors come from two places where those descriptions disagree with the office's practice: personal items that are not personal-care goods, and margarine. We left the descriptions unchanged, since changing them after seeing the reference set would tune on it. A description that matched the official item contents might well raise the LLM scores, but we have not tested that.

The development data are not the deployment text. Cascade thresholds were chosen on the Thai weak-label validation split, which is Open Food Facts text. The confidence of the cheap classifiers on that text did not carry over to official specifications, and the weak labels rate the LLMs far below their reference-set scores, so the development data could not see the value of routing: most cascades routed too few products, and the one that works owes its threshold to a tie-break among the highest candidates. An office would have to choose thresholds on data like the text it classifies.

The product universe is assumed. The reference set is a list, not a stream of products, so the shares n_k in the simulator are assumptions. Section 4.3 reports two: products in proportion to CPI weights, and in proportion to the office's list. A real web or scanner stream has its own mix, and the error in basis points scales with it; the hand-labelled product sample of the original design would have measured it.

Item weights below the published cells are estimated. TPSO publishes weights for its aggregates only. The equal split and the NNLS split are two transparent estimates; they agree on the cells but not on the items inside them. Results that depend on which item inside a cell carries the weight are reported under both.

The simulator measures classification error only. It takes official item relatives as the true price movements of every product and changes nothing but the assignment of products to items. A real web-price index would add sampling, matching, quality-adjustment and outlier error on top. Three modelling choices are assumptions rather than estimates: leaked out-of-scope products move with the indices of the out-of-scope items the reference set shows them coming from, an item that receives no products has its weight redistributed within its CPI group, and the product-level dispersion of the noise variants is assumed, not measured.

One training run per classifier. Every classifier is trained once, with one seed. The intervals cover the sampling of the reference set, not the variation between training runs, so small differences between trained classifiers, such as the text ablations, may not survive a second run.

Closed models change. The LLMs are reached through a commercial API. Model revisions, prices and refusal behaviour can change after the run date; every response is cached so that the numbers here can be reproduced, but not necessarily re-obtained. Local compute is costed at an assumed price per GPU-hour.

Not comparable to Spectrum. Spectrum classifies euro-area web data into the ECOICOP 2018 subclasses (BIS Innovation Hub, 2026). Our taxonomy, language, product universe, label source and metrics differ, so our numbers are not directly comparable with theirs, and we do not compare them.

8 Conclusion

Classification is the step that turns a list of products into a price index, and its error can be stated in the unit of the index. For Thai product text, transfer from other languages gets an embedding classifier to 60.3% of the evaluated CPI weight without a single Thai label; official labels raise this to 68.8%; and a frontier LLM reaches 97.9%, at a cost that a cascade cuts to 27% on the held-out list with no loss the reference set can detect. Passed through a simulator of the Thai CPI, the LLM's errors would misstate headline year-on-year inflation by 1.9 basis points and the cheaper classifiers' by 1.9 to 10.4. Accuracy does not rank methods by that error, and neither does the error alone; both are needed, and the decomposition shows which confusions to fix. Most of the LLMs' remaining errors are not errors of language but gaps in the documentation of the official basket. The open questions are on the side of the data: how these numbers change on shop text rather than the office's own, and on a real stream of products rather than an assumed one. A labelled, probability-sampled set of web products would answer both.

References

Diego Aparicio and Manuel I. Bertolotto. Forecasting inflation with online prices. *International Journal of Forecasting*, 36(2):232–247, 2020. doi: 10.1016/j.ijforecast.2019.04.018.

Australian Bureau of Statistics. Web scraping in the Australian CPI. ABS article, April 2020. URL https://www.abs.gov.au/articles/web-scraping-australian-cpi.

BIS Innovation Hub. Project Spectrum: Using generative AI to enhance inflation nowcasting. Technical report, BIS Innovation Hub, February 2026. URL https://www.bis.org/publ/othp109.htm. A collaboration between the BIS, the Deutsche Bundesbank and the European Central Bank.

Alberto Cavallo. Are online and offline prices similar? Evidence from large multi-channel retailers. *American Economic Review*, 107(1):283–303, 2017. doi: 10.1257/aer.20160542.

Alberto Cavallo and Roberto Rigobon. The Billion Prices Project: Using online prices for measurement and research. *Journal of Economic Perspectives*, 30(2):151–178, 2016. doi: 10.1257/jep.30.2.151.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. arXiv preprint arXiv:2305.05176, 2023. URL https://arxiv.org/abs/2305.05176.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall, New York, 1993. doi: 10.1201/9780429246593.

Martin Eiglsperger. New features in the Harmonised Index of Consumer Prices: Analytical groups, scanner data and web-scraping. *ECB Economic Bulletin*, 2019. URL https://www.ecb.europa.e

u/pub/economic-bulletin/focus/2019/html/ecb.ebbox201902_04~f7cc862940.en.html. Issue 2/2019, box.

Eurostat. Harmonised index of consumer prices: Practical guide for processing supermarket scanner data. Technical report, Eurostat, European Commission, Luxembourg, September 2017. URL https://ec.europa.eu/eurostat/web/hicp/publications.

Eurostat. Harmonised indices of consumer prices: Practical guidelines on web scraping for the HICP. Technical report, Eurostat, European Commission, Luxembourg, November 2020. URL https://ec.europa.eu/eurostat/documents/272892/12032198/Guidelines-web-scraping-HICP-11-2020.pdf.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Alexander Harms and Siemen Spinder. A comprehensive view of machine learning techniques for CPI production. Discussion paper, Statistics Netherlands (CBS), November 2019. URL https://www.cbs.nl/-/media/_pdf/2019/47/machine-learning-for-cpi-production-2019.pdf.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

International Monetary Fund, International Labour Organization, Eurostat, United Nations Economic Commission for Europe, Organisation for Economic Co-operation and Development, and World Bank. *Consumer Price Index Manual: Concepts and Methods*. International Monetary Fund, Washington, DC, 2020. ISBN 978-1-48435-484-1. doi: 10.5089/9781484354841.069.

Paweł Macias, Damian Stelmasiak, and Karol Szafranek. Nowcasting food inflation with a massive amount of online prices. *International Journal of Forecasting*, 39(2):809–826, 2023. doi: 10.1016/j.ijforecast.2022.02.007.

Hazel Martindale, Edward Rowland, Tanya Flower, and Gareth Clews. Semi-supervised machine learning with word embedding for classification in price statistics. *Data & Policy*, 2:e12, 2020. doi: 10.1017/dap.2020.13.

Elizabeth Metcalfe, Tanya Flower, Thomas Lewis, Matthew Mayhew, and Edward Rowland. Research indices using web scraped price data: Clustering large datasets into price indices (CLIP). Office for National Statistics, article, November 2016. URL https://www.ons.gov.uk/economy/inflationandpriceindices/articles/researchindicesusingwebscrapedpricedata/clusteringlargedatasetsintopriceindicesclip.

Ministry of Commerce, Thailand. MOC open data: Retail prices of goods (price indices). Open-data portal of the Ministry of Commerce. In Thai, 2026. URL https://data.moc.go.th/OpenDat

a/CPIIndexes. Price-collection product specifications compiled by the Trade Policy and Strategy Office. Accessed 2026-10-27.

Open Food Facts contributors. Open food facts database. Open Food Facts, data set, Open Database License, 2026. URL https://world.openfoodfacts.org/data. Parquet export from https://huggingface.co/datasets/openfoodfacts/product-database, revision 3ac7e8b6e937. Accessed 2026-10-27.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410.

Andrea Roberson. Applying machine learning for automatic product categorization. *Journal of Official Statistics*, 37(2):395–410, 2021. doi: 10.2478/jos-2021-0017.

Trade Policy and Strategy Office. Explanatory note: Revision of the consumer price index, base year 2023 (B.E. 2566). Ministry of Commerce, Thailand. In Thai, February 2025. URL https://www.price.moc.go.th/price/fileuploader/file_cpi/cpi_note_2566.pdf. Accessed 2026-10-27.

Trade Policy and Strategy Office. Consumer price index open API: Documentation. Ministry of Commerce, Thailand, 2026. URL https://index.tpso.go.th/document/cpig. API host https://index-api.tpso.go.th. Accessed 2026-10-27.

United Nations Economic Commission for Europe. *Machine Learning for Official Statistics*. United Nations, Geneva, 2021. doi: 10.18356/9789210011143.

United Nations Statistics Division. *Classification of Individual Consumption According to Purpose (COICOP) 2018*. Number 99 in Statistical Papers, Series M. United Nations, New York, 2018. URL https://unstats.un.org/unsd/classifications/unsdclassifications/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf. ST/ESA/STAT/SER.M/99. White cover publication, pre-edited text subject to official editing.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL https://jmlr.org/papers/v9/vandermaaten08a.html.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 text embeddings: A technical report. arXiv preprint arXiv:2402.05672, 2024. URL https://arxiv.org/abs/2402.05672.

A Data sources and reproducibility

Table 1: Data sources. Retrieval dates and SHA-256 checksums of every file are in the repository (`data/SOURCES.md`, `data/manifest.csv`).

Source	Use	Licence
Open Food Facts, Open Beauty Facts (Open Food Facts contributors, 2026)	Product text; weak labels (training and development)	ODbL
TPSO open API (Trade Policy and Strategy Office, 2026)	Monthly item and aggregate indices	Thai open government data
TPSO monthly CPI release	Price-updated weights of the aggregates	Thai government publication
TPSO base-year note (Trade Policy and Strategy Office, 2025)	Division weights; base-year changes	Thai government publication
MOC open data (Ministry of Commerce, Thailand, 2026)	Official specifications (official labels)	Thai open government data
COICOP 2018 (United Nations Statistics Division, 2018)	Class codes of the in-scope items	Cited

The repository rebuilds every number in this paper from public sources. `make data` downloads the sources and checks their checksums; `make models` computes the embeddings, trains the classifiers and runs the LLMs (cached responses are replayed, so no API call is made when the cache is present); `make eval econ` scores the methods on the reference set and runs the simulator; `make figures paper` draws the figures and builds this PDF. `cpilens report` writes every number the paper prints into `paper/numbers.tex`, and a check in continuous integration fails if the paper's sources contain a number typed by hand. `results/run_manifest.json` records package versions, model revisions and the random seed (20261027).

B In-scope items and crosswalk

Table 2 lists the in-scope items with their COICOP 2018 classes and both estimated weights. The crosswalk from Open Food Facts category tags to these items is in `taxonomy/off_categories_to_thcpi.csv`, with the basis of each mapping (usually the TPSO item heading it follows).

Table 2: In-scope items. Equal, NNLS: price-updated CPI weight, % of headline, estimated inside official weight cells by an equal split or by NNLS. Ref.: specifications in the reference set (zero where the office publishes none). Source: TPSO; MOC; UNSD; author's calculations.

Code	Item	COICOP 2018	Equal	NNLS	Ref.
11110	Rice	01.1.1	1.16	2.57	250
11121	Flour	01.1.1	1.16	0.02	37
11122	Flour products (bread; noodles; pasta)	01.1.1	1.16	0.88	110
11211	Fresh meat	01.1.2	0.89	2.58	1
11212	Processed meat	01.1.2	0.89	0.39	156
11221	Fresh poultry	01.1.2	0.89	0.95	2
11231	Fresh freshwater fish	01.1.3	0.89	1.22	6
11232	Fresh marine fish	01.1.3	0.89	0.35	7
11233	Other fresh seafood	01.1.3	0.89	0.87	8
11234	Processed fish	01.1.3	0.89	0.62	8
11235	Processed seafood	01.1.3	0.89	0.13	3
11310	Eggs	01.1.4	0.88	0.81	7
11320	Milk and dairy products	01.1.4	0.88	0.95	89
11411	Fresh vegetables	01.1.7	2.13	2.13	22
11412	Processed vegetables and other	01.1.7	0.23	0.43	7
11421	Fresh fruit	01.1.6	2.27	2.27	11
11422	Processed fruit and other	01.1.6	0.23	0.03	39
11521	Oils and fats	01.1.5	0.59	0.44	2
11522	Seasonings and condiments	01.1.9	0.59	0.73	87
11600	Non-alcoholic beverages	01.2.1, 01.2.2, 01.2.3, 01.2.4, 01.2.5, 01.2.6, 01.2.9	3.29	3.29	59
11710	Sugar and confectionery	01.1.8	0.84	0.84	0
12300	Ready-to-eat food	01.1.9, 11.1.1	14.78	14.78	0
36000	Cleaning products	05.6.1	1.54	1.61	763
42100	Personal care products	13.1.2	0.43	3.37	668
71000	Tobacco products	02.3.0	0.31	0.43	17
72100	Beer	02.1.3	0.31	0.54	7
72200	Wine	02.1.2	0.31	0.01	25
72300	Spirits	02.1.1	0.31	0.26	19

C LLM prompt

The system prompt opens with the instructions below. It then lists every label by CPI group (code, Thai and English name, and what it includes) and the official examples, each as `text -> label`. The product is sent as the user message. The full prompt is generated by `cpilens.classify.llm.system_prompt`.

```
You classify products sold in Thai shops into the item categories of the Thai consumer
 ↪ price index (CPI, base year 2023 = BE 2566), compiled by the Trade Policy and Strategy
 ↪ Office (TPSO). Each product is given as `name | brand | quantity`; names are usually
 ↪ Thai, sometimes English or other languages.

Pick exactly one label. First choose the group, then a label inside that group. Use OOS
 ↪ when the product belongs to none of the labels (for example pet food, medicines,
 ↪ kitchenware, stationery, electronics). Classify what the product is, not its flavour:
 ↪ a shrimp-flavoured cracker is a snack (11122), not seafood. Confidence is your
 ↪ probability, 0-100, that the label is correct.
```

D Hyperparameters

Table 3 lists the settings of the classifiers, the cascade, the evaluation and the simulator. Each value is read from the code when the paper is built, so the table cannot drift from what was run.

Table 3: Hyperparameters, as read from the code.

Setting	Value
k-NN: candidate k	1, 3, 5, 10, 20, 50
Linear and MLP: optimiser, learning rate, weight decay	Adam, 10^{-3}, 10^{-4}
MLP: hidden units, dropout	512, 0.2
Batch size (at most); early-stopping patience	512; 3
Fine-tuning: epochs (at most), batch, learning rate	4, 64, 3×10^{-5}
Fine-tuning: maximum input length (tokens)	64
LLM: official examples per item; maximum output tokens	2; 2,048
Cascade: candidate thresholds (the last routes every product)	0, 0.3, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 0.98, 1.01
Official-label budgets	100, 250, 500, 1,000 and all
Bootstrap resamples; simulator draws; prior weight α	1,000; 1,000; 1
Noise variants: annual dispersion (%)	5, 10
Reference set: held-out out-of-scope specifications kept (design-weighted)	1,500

E Compute and cost ledger

Every billed API call is logged with its tokens and cost in `results/llm_costs.csv`; Table 4 sums it. Input tokens include prompt-cache writes and reads, which are billed at different rates.

Table 4: LLM calls and cost (USD) by model and stage. Smoke and probe: pipeline checks; eval: classification of the evaluation and development sets. Source: author's API logs.

Model	Stage	Calls	Input tokens	Output tokens	USD
claude-haiku-5-5	eval	11,360	64,277,293	346,765	0.90
claude-haiku-5-5	smoke	5	27,665	121	0.00
claude-opus-5-5	eval	6,838	39,252,303	195,076	13.78
claude-opus-5-5	probe	10	55,304	262	0.05
Total		18,213	103,612,565	542,224	14.72

F Terms used in the figures

OOS Out of scope: a product of an item that the classifiers do not cover (energy, services, clothing, durable goods and the like). Sending a product to OOS removes it from measurement.

Weighted accuracy The share of the evaluated CPI weight classified correctly, equation (1).

Zero-shot Trained without any Thai label: only on products in other languages.

Weak labels Labels derived automatically from Open Food Facts categories through the crosswalk of Appendix B, not assigned by a person or by the office.

Official labels TPSO item codes attached to its price-collection specifications.

kNN, linear, MLP Three classifiers on top of a fixed text embedding: nearest neighbours, logistic regression and a neural network with one hidden layer.

Fine-tuned XLM-R A multilingual encoder whose own parameters are trained on the labels, rather than a classifier on a fixed embedding.

Cascade A cheap classifier keeps the products it is confident about and sends the rest to an LLM.

Core TPSO's core CPI, which excludes raw food (rice, flour and cereals; meat, poultry and fish; eggs and dairy products; vegetables and fruit) and energy. In the simulator every in-scope item outside those four food groups counts as core.

t-SNE A method that draws high-dimensional embeddings on a plane for display; distances on the map are only roughly meaningful.

Dirichlet draws Confusion matrices drawn from the posterior given the reference counts, used for the intervals of the basis-point errors.

RMSE Root mean squared error of the simulated inflation rate against the official rate, over the simulated months.