How much does product classification cost in basis points?

Policy brief: classifying Thai product text for price measurement

Thanatchon Chongmankhong • October 28, 2026

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.

Why it matters. Prices from retailers' websites and scanner data can make price measurement faster and more detailed. Before such prices can enter a consumer price index, every product has to be assigned to an item of the official basket: a bag of jasmine rice to rice, a carton of UHT milk to milk and dairy products. Done by hand this does not scale; done by a model it introduces errors. This study asks how large those errors would be in the unit that matters for policy: basis points (hundredths of a percentage point) of measured inflation.

What the study did. Using only public data, it tested four kinds of method on Thai product descriptions: low-cost text-matching models, a fine-tuned language model, two commercial large language models (LLMs) instructed with a few examples, and "cascades" that send only the hard cases to an LLM. Each was scored on 3,910 official Thai price-collection descriptions published by the Ministry of Commerce, none of which was used to train the models built here or to set their thresholds, with every item weighted by its share of household spending. Each method's mistakes were then fed through a reconstruction of the Thai CPI to see how far measured inflation would move.

Three findings

1. **The larger LLM classifies almost all of the evaluated basket correctly; low-cost models do not.** The larger LLM assigned 97.9% of the evaluated CPI weight to the right item and the smaller one 95.4%. Of the low-cost models given the full descriptions, the best, the fine-tuned one, reached 69.5%. A cascade that lets a low-cost model handle the easy cases and sends the rest to the larger LLM scored 97.8%, close to the larger LLM on its own, at 27% of its cost on the held-out list. All LLM runs for this study together cost USD 14.72.

2. **With the larger LLM, classification would move headline inflation by about 1.9 basis points.** That is the typical error of the year-on-year rate over 10 months. Allowing for the fact that errors cluster within groups of similar products, the error could be up to 3.7 basis points, and more under other plausible assumptions about the basket weights; on average it would overstate inflation slightly (by 1.5 basis points). For food inflation the typical error is 3.2 basis points. The low-cost models misstate headline inflation by between 1.9 and 10.4 basis points; two of them miss food inflation by 22.3 and 24.2 basis points, and one would have understated it in every month.

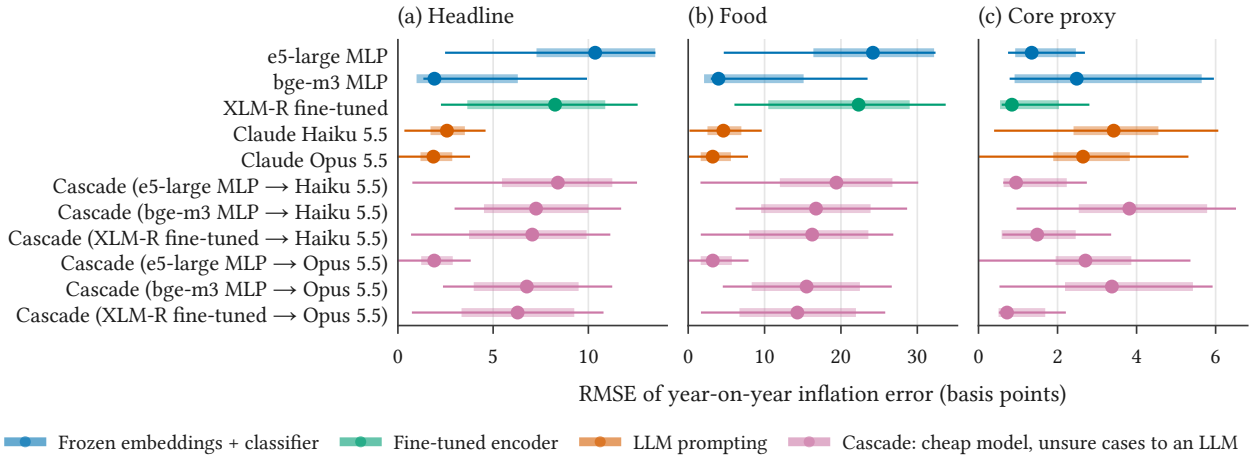
3. **Accuracy alone is a poor guide.** Two low-cost models with about the same accuracy differed several-fold in basis points, because what matters is whether the confused items' prices moved apart. Most of the LLMs' remaining errors came from two items whose official contents differ from what their names suggest: the official "personal items" include wristwatches, school bags and hair dryers, and margarine is counted with milk and dairy products, not with oils and fats. Better documentation of the basket may well fix these; that was not tested.

What it means for central banks in Asia

- **Basis points are a more useful yardstick than accuracy.** The method needs only a test set of labelled products, the published CPI weights and the published item indices, so an office could state the classification error of a new data source before using it.

- **Published specification lists are valuable.** The Ministry of Commerce's open specifications made this study possible. Descriptions of what each item includes and excludes could address the largest errors found here.

Headline error, upper end of the wider interval: Claude Opus 5.5 3.7 bp; best cheap method 9.9 bp



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 1: The LLMs and the cascade to the larger LLM would misstate inflation by a few basis points; low-cost models range from as little to several times more. Typical error (RMSE) of year-on-year inflation caused by each method's classification mistakes, over 10 months. Dots: central estimate; thick lines: 95% interval treating each description as independent; thin lines: allowing errors to cluster. On core inflation, which excludes raw food and energy, the low-cost models do as well as the LLMs. Source: Ministry of Commerce; Trade Policy and Strategy Office; author's calculations.

- **LLMs are affordable at this scale; governance is the larger question.** Commercial models change over time, so responses would need to be stored and versioned. Open-weight models that an office could run itself were not tested.
- **Languages with few labelled data can start from other languages.** With no Thai training labels at all, a low-cost model trained on products in other languages placed 60.3% of the weight correctly: a starting point, not a result an index can rely on.

Caveats

The test descriptions are the statistical office's own short texts; shop descriptions are longer and noisier, and accuracy on them will be lower. The descriptions are public, so the LLMs may have seen them in training. Item weights below the published groups are estimated, and the results depend on how. The basis-point figures assume a mix of products in proportion to household spending, cover the 24.9% of the headline weight that the test set can evaluate, and measure classification error only, not the other errors of a web-price index. A low basis-point error is meaningful only together with high accuracy: a model that classifies little of the basket also moves the index little. The results are for one country and one point in time and are not directly comparable with studies of other markets. The two LLMs are Claude Opus 5.5 (larger) and Claude Haiku 5.5 (smaller).

Full paper, code and data: `https://github.com/thanatchon36/cpilens` (DOI 10.5281/zenodo.23686886).