ILLUSTRATIVE FINDING — structure and method are representative; situation is composited from cross-market pattern recognition, not a single engagement.

Make vs. Buy

When In-House NLP Classification Beats External APIs

A multilingual editorial platform. A classification layer that was expensive, slow, and wrong in ways the vendor couldn't fix. Five phases — the question reframed, the map, the field, the finding, the direction.

---

01 — THE QUESTION

The Question

As asked

“Which NLP APIs should we be using for content classification?”

The actual question

At what operational scale and consistency requirement does the cost — in money, latency, and editorial control — of external NLP APIs exceed the cost of training, deploying, and maintaining in-house classifiers? And what does the transition architecture look like?

The reframe matters. The original framing produces a vendor comparison. The actual question produces a build-or-buy decision framework — which is what the organization needs.

---

02 — THE MAP

The Map

A multilingual editorial platform processing content at publication speed across 30+ language markets was using external NLP APIs as its primary classification layer. Each API call carried three costs that compounded with scale: direct cost per inference, latency added to the ingestion pipeline, and dependency on a vendor's model decisions — training data, category taxonomies, versioning.

The platform had a taxonomy problem the APIs couldn't solve. Vendor categories were built for breadth; the platform needed editorial-grade classification tuned to its specific audience signal and downstream decision requirements. The result was a classification layer that was simultaneously expensive, slow, and wrong in ways the vendor had no incentive to fix for a single client.

The decision: continue optimising the API layer, or invest in in-house capability. The question was when, not whether.

---

03 — THE FIELD

The Field

Two dynamics were shifting simultaneously, and their interaction made the decision urgent rather than theoretical.

First: the cost of training and serving classifiers had fallen significantly. Pre-trained multilingual foundation models meant building in-house no longer required training from scratch in each language. The marginal cost of adding a language to an in-house classifier had dropped faster than API pricing adjusted.

Second: retrieval-augmented systems and agentic workflows were placing heavier demands on classification precision. Generic API taxonomies — built for coverage, not correctness at the edge — were becoming a bottleneck in pipelines that needed to gate production decisions on classification output. The accuracy gap between "good enough for broad categorisation" and "accurate enough to trust downstream" was widening, and no vendor had incentive to close it for any individual client's taxonomy.

The crossover point was approaching faster than the cost math alone suggested.

---

04 — THE FINDING

The Finding

The crossover happens earlier than expected. And the constraint is not model training — it's corpus curation.

In-house classification becomes defensible when you can produce labeled training data consistently across your target languages. The model is increasingly a solved problem; multilingual foundation models provide the base. What determines whether an in-house classifier outperforms a generic API is whether your labeling process can produce consistent signal at sufficient volume.

At publication scale — tens of thousands of articles per day per language — the cost of a consistent labeling infrastructure is recovered within months. The accuracy gain is structural, not marginal, because the classifier is trained on your taxonomy, your signal, and your quality threshold rather than a generic one. Latency and vendor-dependency costs disappear.

The decision to go in-house is not primarily a cost decision. It's a control decision. Cost makes it defensible. Control is why it compounds.

---

05 — THE DIRECTION

The Direction

Don't evaluate NLP vendors. Evaluate your labeling capacity.

The sequence that unlocks in-house classification:

1. Define your taxonomy in terms of the decisions it needs to drive — not topic categories, but operational gates.
2. Build a reproducible labeling process: human reviewers, a consistency rubric, and a tool that surfaces inter-annotator disagreement before it compounds.
3. Start with your highest-volume, most decision-critical classification task. Build a single-language classifier. Measure it against the API baseline on your taxonomy, not the vendor's.
4. Use that classifier's performance data to make the make-vs-buy case for adjacent languages. The argument is now empirical, not theoretical.

The organisations that stay stuck at the vendor layer are not short on budget. They're short on labeled data and on a taxonomy that's actually connected to an operational decision.

This is what a Wire finding looks like.

Your question is different. The method isn't. A two-to-three-week bounded engagement — one question in, one written finding out, a direction attached.

`hello@produser.de · produser.de/wire · Subject: The Wire — a question`
