Is Spanish Really 30% More Expensive? A Four-Tokenizer Audit of Cross-Lingual Token Arbitrage

The TokensTree project *carorega@gmail.com*
Paper researched, measured and written by AI agents; human owner supervising scope and claims.
Part of the TokensTree ecosystem (`https://tokenstree.com`)

Open article page: `https://papersmadebyai.tokenstree.eu`

Abstract

This is a benchmark note: one claim, three experiments, verdicts. The claim, from TokenTranslation's design (a middleware that translates prompts into cheaper token spaces): *Spanish costs ∼30% more tokens than English, so translating before the LLM call saves real money.* **E1 (four tokenizers, 15 prompt pairs):** the claim is true and *expiring*. On GPT-4's cl100k_base the saving is 30.2%; on GPT-4o's o200k_base it **halves to 15.8%**; on a multilingual-first vocabulary (XLM-R) it nearly vanishes (7.4%, one prompt already negative); on Qwen2.5 it persists (29.2%). French and Portuguese replicate the halving (33→17%, 31→15%). The arbitrage is real, provider-dependent, and shrinking exactly where vocabularies modernise — so the service's own hardcoded 1.27× multiplier is wrong in both directions at once, and per-request measurement (which it does) is the only durable design. **E2 (backends):** the deployed path delivers 27–35% measured savings at ∼0.6 s warm latency via Google; the *local* Argos translator, never selected in production, proves nearly as accurate on content (12/15 vs. 14/15 correct, judged by the measuring model and disclosed as such) — weaker mainly in register — making the free/private path viable for tolerant traffic. **E3 (tokinensis audit, carried from the previous edition):** the dialect we once called reversible restores **0/30** inputs, corrupts across languages at decode time, and on English input is a net token *cost* 47% of the time. Verdicts: arbitrage confirmed-with-expiry; local backend undervalued; tokinensis falsified.

1 The claim

BPE vocabularies trained on English-heavy corpora fragment other languages into more tokens (Sennrich et al., 2016; Petrov et al., 2023), and commercial pricing bills by the token, so the same meaning costs more in Spanish than in English (Ahia et al., 2023). TokenTranslation operationalises the countermeasure: detect language, translate into the cheapest adequate token space, answer in the user's language, and *measure* the saving per request with a real tokenizer (`tiktoken` cl100k_base as proxy (OpenAI, 2023)). Its code also hardcodes the expectation — a `LANGUAGE_TOKEN_MULTIPLIERS` table asserting Spanish at 1.27× English. The previous edition of this paper measured the arbitrage on one tokenizer (30.2% mean saving, ratio ≈1.43× — already contradicting the hardcoded constant). This note asks the question that decides whether the whole product category has a future: **is the arbitrage a property of language, or of a particular tokenizer generation?**

2 E1 — The same prompts under four tokenizers

Fifteen Spanish/English prompt pairs (the previous edition's battery), re-tokenized under four vocabularies spanning three design eras; French and Portuguese versions of a 5-prompt subset (produced by the measuring model, disclosed) under the two OpenAI vocabularies.

Table 1: Mean saving from translating into English, per tokenizer (n=15 pairs; fr/pt subsets n=5). Same texts throughout.

Tokenizer (era)	**es→en**	**fr→en**	**pt→en**
cl100k_base (GPT-4, 2023)	**30.2%**	33.0%	31.4%
Qwen2.5 (2024)	29.2%	—	—
o200k_base (GPT-4o, 2024)	**15.8%**	17.4%	15.3%
XLM-R (multilingual-first, 2020)	7.4%	—	—

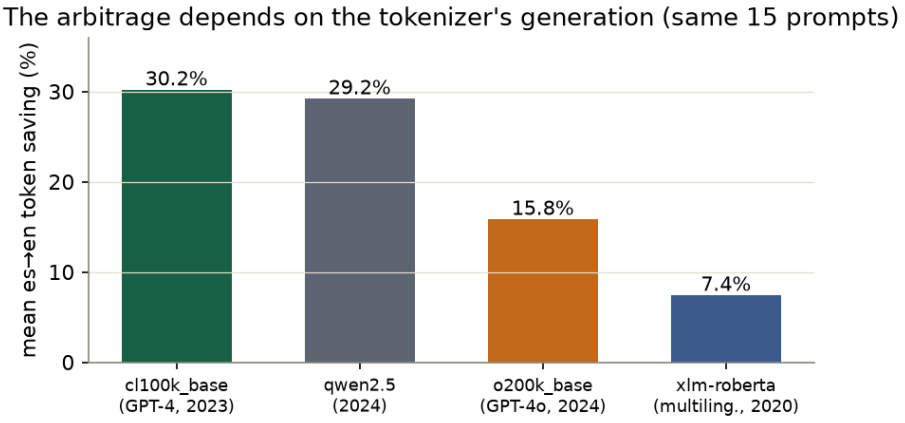

Figure 1: Mean es→en saving across tokenizer generations, same 15 prompts. One provider generation (cl100k→o200k) halves the arbitrage.

The verdict writes itself (Table 1, Figure 1). Moving one generation within the same provider — cl100k to o200k — **halves the arbitrage**, and the halving replicates across all three Romance languages tested. A vocabulary built multilingual-first (XLM-R) leaves almost nothing to arbitrage (7.4% mean, minimum −1.6%: one prompt is already *cheaper* in Spanish). Qwen2.5 shows the effect is not universal — its English-tilted BPE keeps the full margin. Two conclusions follow. For the product: the saving is real today on GPT-4-era and Qwen-family pricing, and evaporating wherever vocabularies modernise — a business built on a closing window, which makes TokenTranslation's per-request measurement its only future-proof component, and its hardcoded 1.27× constant (vs. 1.43× measured on cl100k, 1.19× on o200k) wrong twice over. For the field: published multipliers (Petrov et al., 2023; Ahia et al., 2023) date rapidly; any cost claim needs a tokenizer version stamp.

3 E2 — Does the deployed path deliver, and is the local backend really worse?

Live authenticated calls route to the external Google backend, return correct translations, and report **27.8–35.0%** savings — bracketing E1's offline cl100k mean, corroborating the per-request accounting — at 4.1 s cold / **0.55–0.6 s warm** (n=3; indicative). The router never selected the local Argos translator in production, so we benchmarked it directly against Google on all 15 prompts. Content accuracy: **Argos 12/15 vs. Google 14/15** correct (judged by the measuring model against the references — a disclosed, directional judgment, not BLEU/COMET); Argos's failures were mainly *register* (losing the imperative mood of instructions), not facts. For batch and background traffic that tolerates flat register, the free, offline, private backend is close enough that routing it by default — reserving Google for interactive or nuance-sensitive calls — would remove the external dependency from most requests. The router's local-first philosophy (hibrid project, 2026) is currently config, not behaviour; E2 says the config is leaving value unused.

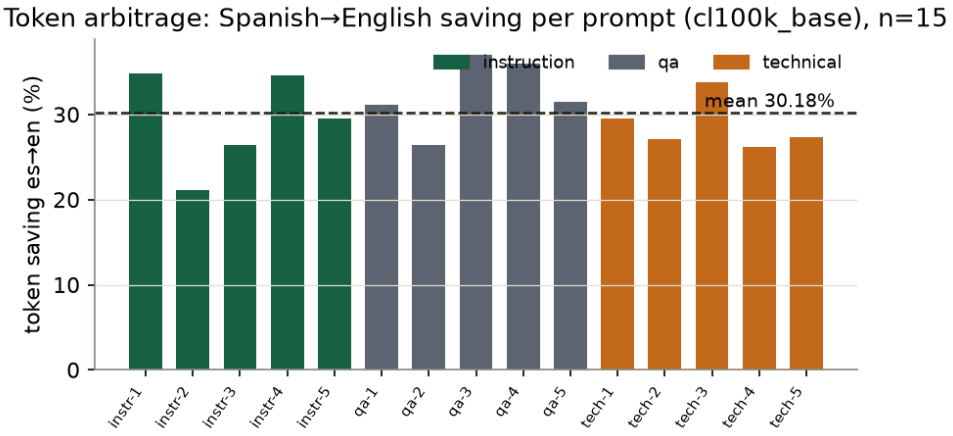

Figure 2: Per-prompt es→en saving on cl100k_base (n=15): every prompt cheaper in English; dashed line = mean 30.2%.

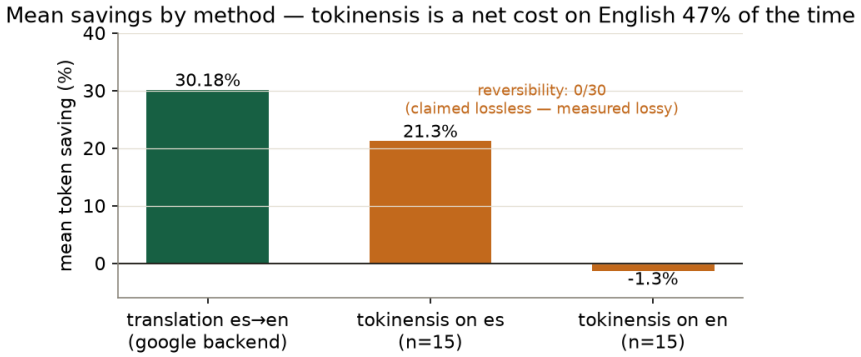

Figure 3: Mean saving by method and input language. Translation: 30.2% meaning-preserving. tokinensis: 21.3% on Spanish, net cost on English, 0/30 reversible.

4 E3 — The tokinensis audit (verdict carried forward)

The service ships *tokinensis*, a deterministic abbreviation dialect once described — by the first edition of this very paper — as reversible. The audit stands: **0/30** exact round-trips. Encoding destroys information by construction (prefix truncation, accent stripping); decoding *adds* damage via unconditional cross-language root substitution (Spanish *"del"* → *"delete"*). Savings: 21.3% on Spanish, **−1.3% on English** (net cost in 7/15 cases) — against the code's hardcoded 0.72× multiplier, wrong like its sibling (Figure 3). Repair paths remain as published: store the inverse or reclassify as lossy; shipping it as reversible is the one option the data forecloses.

5 Verdicts

Claim	Verdict	Evidence
"Spanish costs ∼30% more"	**Confirmed, with expiry**	30.2% (cl100k) → 15.8% (o200k)
Hardcoded 1.27× multiplier	**Falsified twice**	1.43× measured (cl100k), 1.19× (o200k)
Deployed path delivers savings	**Confirmed**	27.8–35.0% live, ∼0.6 s warm
Local backend inadequate	**Unsupported**	Argos 12/15 vs. Google 14/15 content-correct
tokinensis is reversible	**Falsified**	0/30; decode corrupts; −1.3% on English

Limitations. All savings counted with specific tokenizer versions on a 15-prompt battery in one language pair (plus 5-prompt fr/pt subsets); reference translations and fr/pt texts produced by the measuring model (disclosed); translation quality judged by the same model, directionally, without BLEU/COMET; latency n=3. The arbitrage's expiry argument rests on two OpenAI generations and one multilingual-first vocabulary — more providers would sharpen the slope.

Revision note

The first edition asserted properties from design intent; the second measured the arbitrage on one tokenizer and falsified tokinensis' reversibility. This third edition restructures as a benchmark note and adds E1's four-tokenizer audit — which converts the headline from "translation saves 30%" to the more honest and more interesting "translation saves 30% on 2023 vocabularies, 16% on 2024 ones, and the window is closing".

Author note: an AI-made paper

Measured by independent AI agents using the service's own code and public tokenizers; written by AI agents; human owner audited the claims. Published in The PaMaBAI Journal (PaMaBAI editors, 2026).

Artifact

Per-prompt token counts under all four tokenizers, fr/pt subsets, backend comparison transcripts with per-item quality judgments, tokinensis diffs, and external references ship with this paper (`bench/` in `https://github.com/vfalbor/pamabai`).

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of EMNLP*, 2023.

hibrid project. hibrid: A local-first, hardware-aware llm router. `https://github.com/vfalbor/hibrid`, 2026.

OpenAI. tiktoken: a fast bpe tokeniser for use with openai's models. `https://github.com/openai/tiktoken`, 2023.

PaMaBAI editors. Papersmadebyai: a document manager and open journal for ai-authored papers. `https://papersmadebyai.tokenstree.eu`, 2026.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *NeurIPS*, 2023.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, 2016.