Research Technical Report FRITS-TR-2026-02 5 July 2026 TRANSLATION MULTILINGUAL LOCALISATION LLM

Re-Authoring, Not Translating: A Two-Stage Meaning-First Pipeline for Native-Quality Machine Translation

Frits Lyneborg

FRITS AI ApS

ABSTRACT

Machine translation — including translation by strong LLMs — produces output that is technically correct and unmistakably foreign: calqued phrases no native speaker would write, invented compound words, English sentence structure wearing local vocabulary. Operating a product in 26 European languages forced us to solve this at scale. Our finding is that the failure is not a quality problem but a framing problem: a model told to translate will mirror the source; a model given the meaning and told to write will produce native copy. We describe the two-stage pipeline we built on this finding — stage one paraphrases the English source into plain-language meaning at low temperature; stage two hands that meaning to a native-speaker persona writing original copy at deliberately higher temperature — together with the two anti-rules that eliminate the most reliable machine-translation tells, the placeholder and ICU-plural guards that make the pipeline safe for software strings, a ship-blocking artefact scanner, and an LLM-as-native-reviewer verification loop. We also report the negative finding that made our language list honest: for one language we evaluated, no available model writes acceptably, and the correct engineering decision was refusing to ship it.

1. The problem: translations that are correct and wrong

Every team that has localised software with machine translation knows the result: nothing is *mistranslated*, and none of it reads as if a native speaker wrote it [1 (#ref-1)]. The tells are consistent across engines and across target languages [2 (#ref-2)]:

- **Calques** — word-for-word transfers that are grammatical and alien. English "free welcome credits" becomes Danish *"Gratis velkomstkreditter"*, a phrase no Dane has ever written; a Dane would say *"Gratis credits"* or reach for a different idea entirely, like *"velkomstgave"* (welcome gift).
- **Invented compounds** — the surest machine tell. English "long-press" becomes Danish *"Langtryk"*, a word that does not exist. Actual Danish usage is either the English loanword ("long press") or a plain description ("hold fingeren nede" — hold your finger down).
- **Mirrored structure** — the English sentence's skeleton, clause order and rhythm, reproduced in local vocabulary. Each sentence defensible; the whole unmistakably translated.

We operate a consumer product in 26 European languages (the 24 official EU languages plus Norwegian and Icelandic, with Ukrainian added), with roughly 1,950 interface strings per language — about 47,000 translated strings under continuous change. At that scale, hand-polishing is not an option, and "translated-sounding" is not cosmetic: users in smaller language markets correctly read it as *this product was not made for me*.

2. The finding: translation is the wrong task

The insight the pipeline is built on: **the failure mode is caused by the instruction "translate."** Given a source text and asked to translate it, a language model does what the task implies — it preserves the source's structure, because structural fidelity is what distinguishes translation from paraphrase in its training data. The calques and mirrored syntax are not failures of capability; they are obedience [4 (#ref-4)].

The fix is to never ask for translation. We split the task so that no stage ever holds both the English phrasing and the responsibility for the target-language phrasing:

Stage 1 — reduce the source to its meaning. A first pass rewrites the English source into plain descriptive English: metaphor, marketing energy and jargon stripped, the communicative intent preserved. “Your phone now sings — a brand new chime plays when your answer is ready” becomes “A short sound now plays when the answer is ready.” This runs at temperature 0.2 — it is a controlled, almost mechanical rewrite.

Stage 2 — author the meaning natively. The plain-meaning text goes to a second pass framed as a person, not a function: a native speaker of the target language, born and raised in the country, working as a senior copywriter, who receives a *description of meaning* and writes original copy expressing it. The prompt states outright: *the English is a description of meaning, not a template whose syntax you should mirror.* The model is free to split, merge and reorder sentences. This runs at temperature 0.7 — deliberately higher, because the most-probable continuation of a translation-shaped context is precisely the calque [3 (#ref-3)], and escaping it requires giving the model room.

The two-stage split matters more than either stage alone. Paraphrase-then-translate with a “translate” instruction in stage 2 still calques; a single-stage “write natively” instruction on the raw marketing English imports the source’s rhetoric. Meaning extraction and native authorship have to be separated.

3. The two anti-rules

Two explicit rules in the stage-2 instruction remove the most reliable machine tells. We publish them because they are the transferable core of the method.

Rule 1 — the calque test. If the most obvious target-language equivalent of an English word is a calque or a foreign-sounding loan, replace it with what a native speaker would actually write, even when that means changing the image. English “tactile” tempts Danish *“taktil”* — a word that exists and that nobody uses; the native choice is *“mærkbar”* (noticeable). The rule forces the model to ask “what would be written here natively?” instead of “what does this word map to?”

Rule 2 — never invent words. When an English UI or technical concept has no established native equivalent, the model must not manufacture a compound from native roots. Manufactured compounds (*"Langtryk"*) are the single most identifiable machine-translation artefact. The two permitted moves are (a) the English loanword, if that is genuinely how natives refer to the concept in technical contexts, or (b) a short plain-language description of the action. This trap exists in every language; the rule is stated generically and enforced per locale.

Alongside the anti-rules, each batch carries register context (a live-chat widget is "warm, plain"; a security page is "precise, factual, lightly reassuring — a trust centre, not marketing") and a consistency constraint: where several strings share a concept, the model must choose one native root for it and use it throughout — never mixing two synonyms across a screen.

4. Making it safe for software strings

Interface strings are not prose; they carry structure that must survive an aggressive re-authoring stage untouched. Three mechanisms make the freedom of Section 2 safe:

ICU plurals are translated inside-out. Strings like `{count, plural, one {# source} other {# sources}}` keep their ICU skeleton verbatim; only the words inside each branch are authored. Crucially, the model is instructed to *add* the plural categories the target language needs rather than mirroring English's two: Polish, for instance, correctly gains `few` and `many` branches. Structural validity is checked mechanically after the run; a malformed plural fails the batch, not the user.

Brand phrases bypass the model entirely. Recurring, brand-critical phrases (our "European AI" tagline is the canonical case) are pre-authored natively once per locale and substituted deterministically: the model sees a placeholder token, and post-processing replaces it with the approved native phrase. The model cannot degrade what it never touches.

Leaks are ship-blockers, not warnings. Models intermittently emit the placeholder token itself, or fragments of scaffolding, into output. We learned this the usual way — a

token leaked into six locales in one production batch — and responded by making detection structural: a scanner runs over *every* locale file after *every* batch, matching machine-token patterns that cannot occur in legitimate copy (while ignoring legitimate acronyms like GDPR that also appear in the source). Any hit fails the run with a nonzero exit code. A leak can still happen; it can no longer ship.

The operational pipeline around this is deliberately boring: strings are translated in chunks of ~20 to stay inside output-token limits, failed chunks are simply not written, and the batch runner is resumable — a re-run fills exactly the gaps the previous run left. In one catch-up run this filled 1,868 missing strings across locales without touching an existing translation.

5. Verifying without native staff

A one-company team cannot employ 26 native reviewers, so verification is also delegated — to the model, in a different role with different incentives. A reviewer persona (native speaker, senior copywriter, seeing only the target-language text and a description of intent) rates key phrases on a 1–5 scale, where 1 is "obvious machine translation" and 5 is "indistinguishable from native copy," and must justify any score below 5 with a concrete objection and a proposed fix [6 (#ref-6)]. Scores below 4 are flagged for human decision.

This is weaker than human native review, and we treat it accordingly: it is a *regression detector*, good at catching the calques and register errors the authoring stage occasionally lets through, not a certification of nativeness. For lower-resource languages we additionally flag output for opportunistic native-speaker review. The honest summary: the review loop reliably catches the difference between bad and plausible; the difference between plausible and perfect still needs a human who grew up with the language.

6. The negative finding: knowing when not to ship a language

The method has a boundary, and we consider finding it a result in its own right. Evaluating Greenlandic (Kalaallisut), we found no available model — including the strongest European ones — writes it usably [5 (#ref-5)]; output mislabels basic interface concepts and would read as parody to a native speaker. No pipeline design fixes a model that cannot write the language: the re-authoring method amplifies fluency, and where there is no fluency, there is nothing to amplify.

The correct engineering decision was to not ship the language. Greenlandic-locale users are served Danish — which effectively all Greenlandic readers also read — rather than broken Kalaallisut. We record this because the pressure in localisation is always toward claiming more languages, and the honest version of a language list is one where every entry passed the same bar. (Icelandic, our lowest-resource shipped language, passed — with a standing flag for extra native review.)

7. Limitations

The pipeline currently runs on a single model family (Mistral Large; smaller models of the same family calque noticeably more, and we do not use them for translation). All temperatures, chunk sizes and artefact patterns are calibrated to it. Register decisions that are genuinely contested within a language (formal vs. informal address in German) are delegated to the native persona plus per-batch context rather than encoded as rules, which trades auditability for naturalness. The reviewer loop shares a model family with the author loop and therefore shares some blind spots. And the method is validated on interface copy, notifications and short marketing text — string-sized units; we have not validated it on long-form documents, where discourse-level cohesion adds constraints the current pipeline does not model.

The engine described here runs as an internal business tool at FRITS AI (it is what localises our production assistant); it is not currently exposed as a product. This report is intended to be sufficient for a competent team to implement the method. If you do — especially for languages we have not covered — we would like to hear how it behaves: contact (/contact/).

References

1. Gellerstam, M. — *Translationese in Swedish Novels Translated from English.* In L. Wollin & H. Lindquist (eds.), Translation Studies in Scandinavia (SSOTT II), pp. 88–95. Lund: CWK Gleerup, 1986.
2. Volansky, V., Ordan, N., Wintner, S. — *On the Features of Translationese.* Digital Scholarship in the Humanities 30(1):98–118, 2015. doi:10.1093/llc/fqt031.
3. Raunak, V., Menezes, A., Post, M., Hassan, H. — *Do GPTs Produce Less Literal Translations?* Proceedings of ACL 2023 (Short Papers), pp. 1041–1050. arXiv:2305.16806.
4. Hendy, A., et al. — *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation.* arXiv:2302.09210, 2023.
5. Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., Kaur, R. — *Neural Machine Translation for Low-Resource Languages: A Survey.* ACM Computing Surveys 55(11):1–37, 2023. doi:10.1145/3567592.
6. Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., Macherey, W. — *Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation.* Transactions of the ACL 9:1460–1474, 2021. doi:10.1162/tacl_a_00437.

How to cite

Frits Lyneborg (2026). Re-Authoring, Not Translating: A Two-Stage Meaning-First Pipeline for Native-Quality Machine Translation. FRITS AI ApS, Technical Report FRITS-TR-2026-02. https://frits.ai/research/meaning-first-translation/

```
@techreport{lyneborg2026authoring,
  title       = {Re-Authoring, Not Translating: A Two-Stage Meaning-First Pipeline for Native-Q
  author      = {Lyneborg, Frits},
  institution = {FRITS AI ApS},
  number      = {FRITS-TR-2026-02},
  year        = {2026},
  month       = {jul},
  url         = {https://frits.ai/research/meaning-first-translation/}
}
```