LangMAP: A Language Adaptive Approach to Tokenization

EMNLP2026 BUDAPEST

github.com/cimeister/unimixlm

Clara Meister, Suchir Salhan, Andrzej Szablewski, Pietro Lesci, Paula Buttery & Tiago Pimentel

Problem

😢 English-centric tokenizers often fragment morphologically rich languages and non-Latin scripts into more tokens
💰 Longer sequences increase compute and API costs
🤯 Poor token boundaries can obscure meaningful morphological structure.

LangMAP

We observe that one vocabulary can have many valid segmentations.

⟨ Rathaus ⟩
1 token

⟨ Rat, haus ⟩
2 tokens

⟨ Ra, th, aus ⟩
3 tokens

Our Solution

LangMAP adapts a pretrained tokenizer to a language simply by changing how we score existing tokens, rather than changing the vocabulary itself!

HOW DOES LANGMAP WORK?

Fix shared vocabulary $\mathcal{V}$

Step 1: The vocabulary V fixes the candidate segmentations. Same vocabulary → same candidates for every language

`input "Rathaus"` → A `Rathaus` B `Rat` `haus` C `Ra` `th` `aus`

Step 2: Each language's distribution φ assigns a log-probability to each candidate, where φ is learned via the UnigramLM algorithm on data from that language

Estimate ϕ_ℓ per language via EM → **Viterbi segmentation per language**

Each language gets its own token frequency profile $\hat{\phi}_\ell$ → each language gets its own segmentations

Step 3:
Return the highest likelihood segmentation across all languages.

German φ_{de}
A `Rathaus` log p = −3.2
B `Rat` `haus`
C `Ra` `th` `aus`

English φ_{en}
C `Ra` `th` `aus` log p = −8.7
B `Rat` `haus`
A `Rathaus` *P(Rathaus) ≈ 0 under φ_{en}*

argmax (Eq. 8): −3.2 > −8.7 → `Rathaus` *language identified implicitly: German*

Bayes' rule → posterior over per-lang segmentations

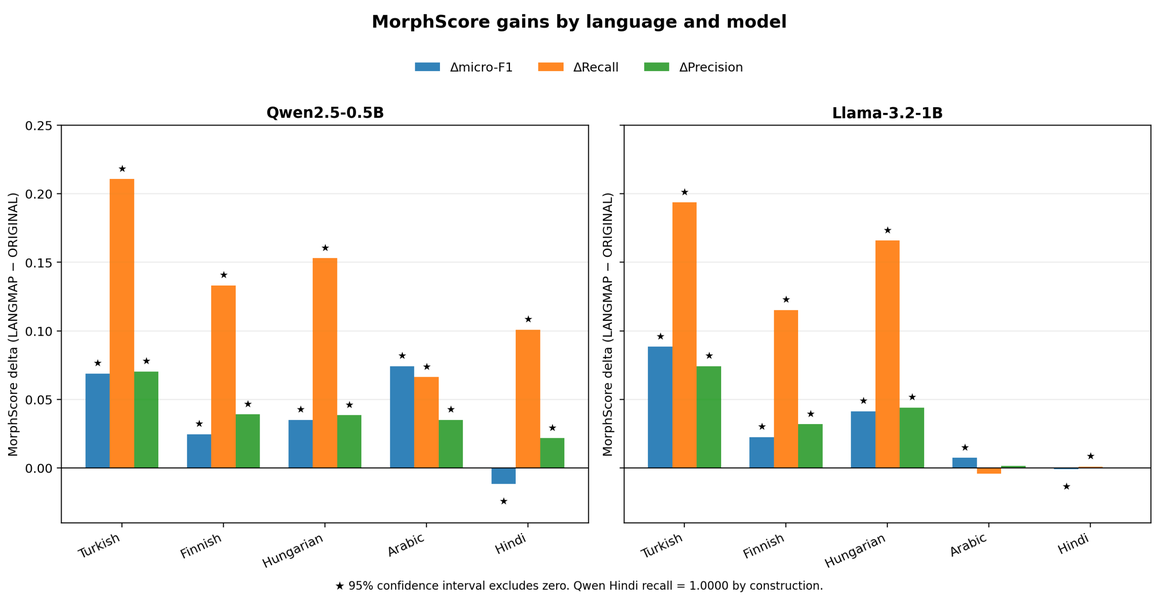
$$p(\ell \mid s) = \frac{p\left(\tau(s;\phi_\ell);\phi_\ell\right)}{\sum_{\ell' \in \Lambda} p\left(\tau(s;\phi_{\ell'});\phi_{\ell'}\right)}$$

$$\tau(s) = \operatorname{argmax}_{v:\, \phi_\ell} p(\, v \mid s;\, \phi_\ell)$$

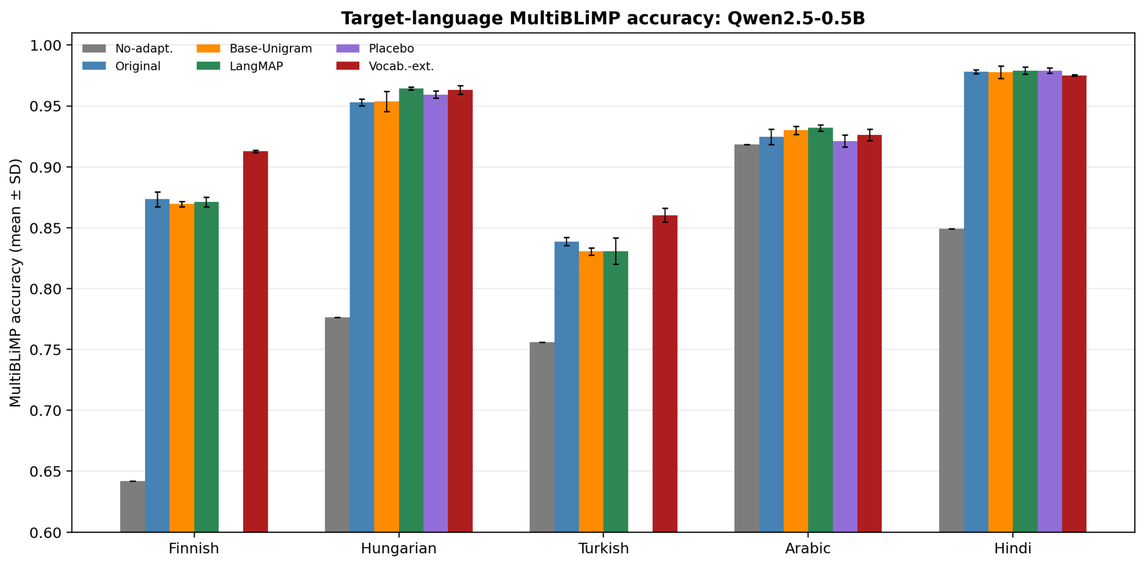
Results

LangMAP improves morphological segmentation across natural languages and consistently aligns code tokens with AST structure with minimal computational overhead.

LangMAP improves morphological alignment (MorphScore)



Fine-tuning with LangMAP segmentations does not show consistent improvements.



Setup

- **Base Models:** Qwen2.5-0.5B, Llama-3.2-1B; Qwen, Llama, Gemma, Mistral tokenizers.
- **Method:**
  - Train LangMAP tokenizers (trained on 25M language tokens), compare against original base model tokenizer and language-agnostic UnigramLM.
  - Finetune models with LangMAP v base tokenizer.
- **Intrinsic Tokenizer Evaluation Metrics:**
  - Morphological boundary alignment, AST boundary alignment
- **Downstream:**
  - MultiBLiMP, FLORES+ (bits/byte), Belebele

LangMAP improves code segmentation (AST Alignment)

