Breaking the Curse of Multilinguality with Cross-lingual Expert Langauge Models

[1] UWNLP

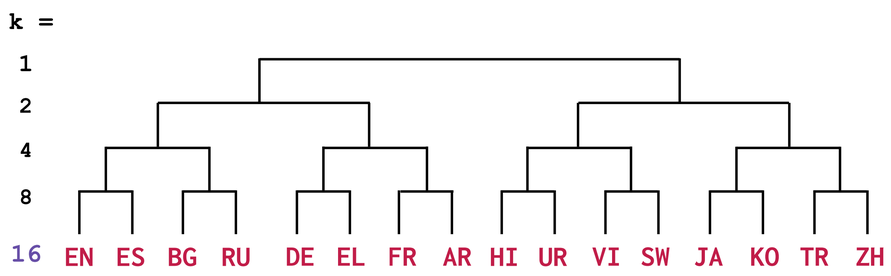
[2] Charles University in Prague

[3] Allen Institute for Artificial Intelligence

Terra Blevins[1] Tomasz Limisiewicz[2] Suchin Gururangan[1]
Margaret Li[1] Hila Gonen[1] Noah A. Smith[1,3] Luke Zettlemoyer[1]

Step 0: Data Allocation

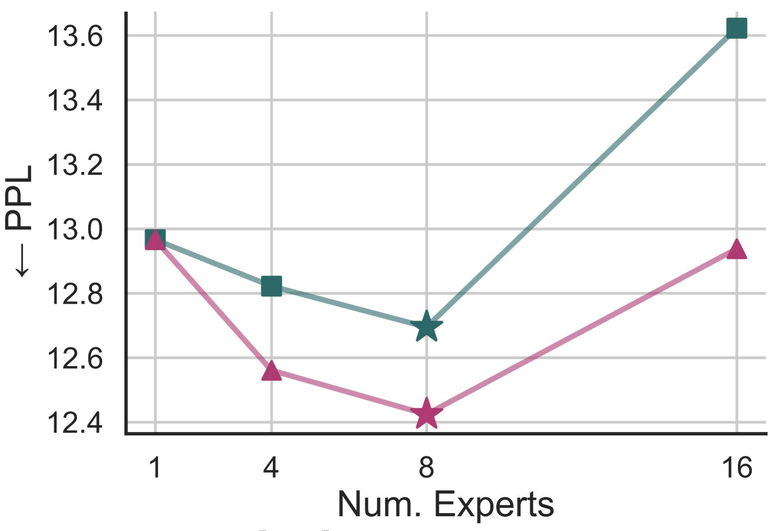
- **TF-IDF Clustering** over training documents
- **Typological Clustering** of training languages



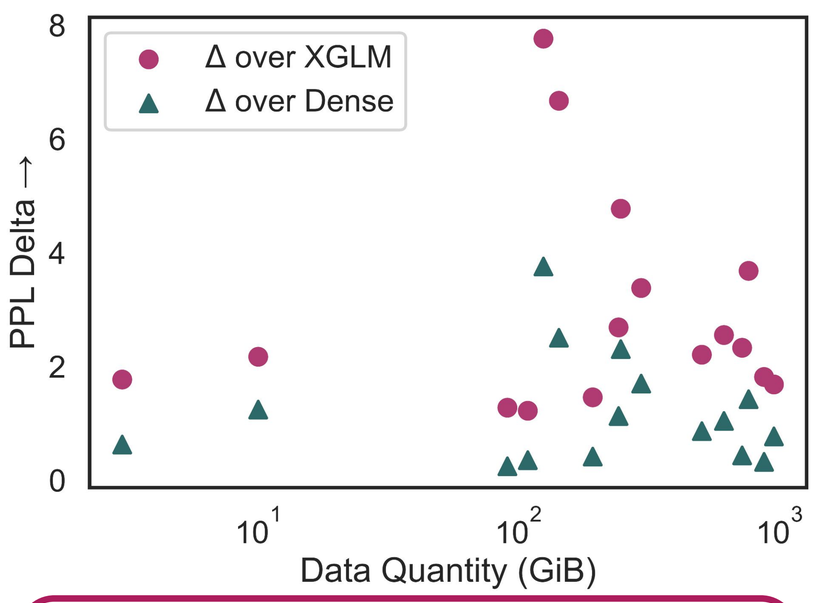
Goal ➡ build better LM for **all languages** while maintaining **cross-linguality** of model

How many (***k***) experts are ideal for this data setting? Using...

- TF-IDF Clusters
- Typological Clusters

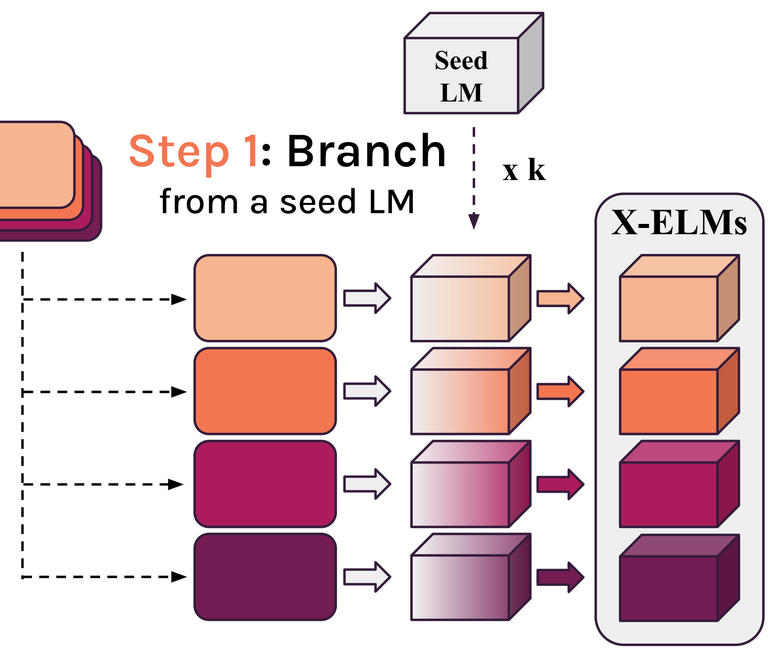


*** with fixed compute budget**



Better language modeling for all languages

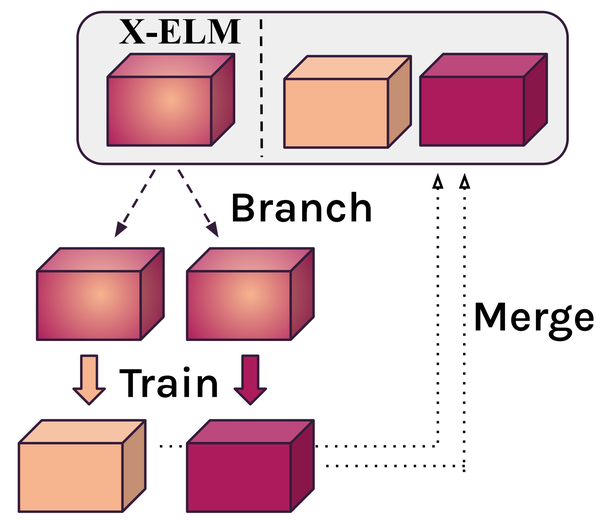
Step 1: Branch from a seed LM



Step 2: Train

Train **expert language models** with multilingual **Branch-Train-Merge**

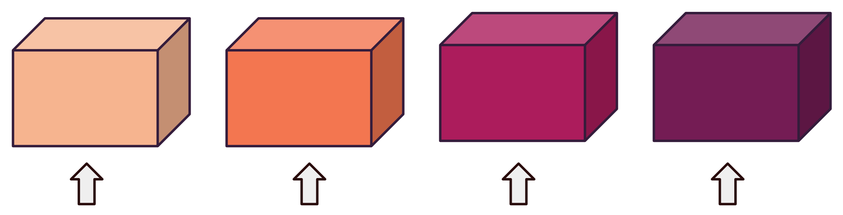
Step 3: Merge



Inference with X-ELM

Ensemble

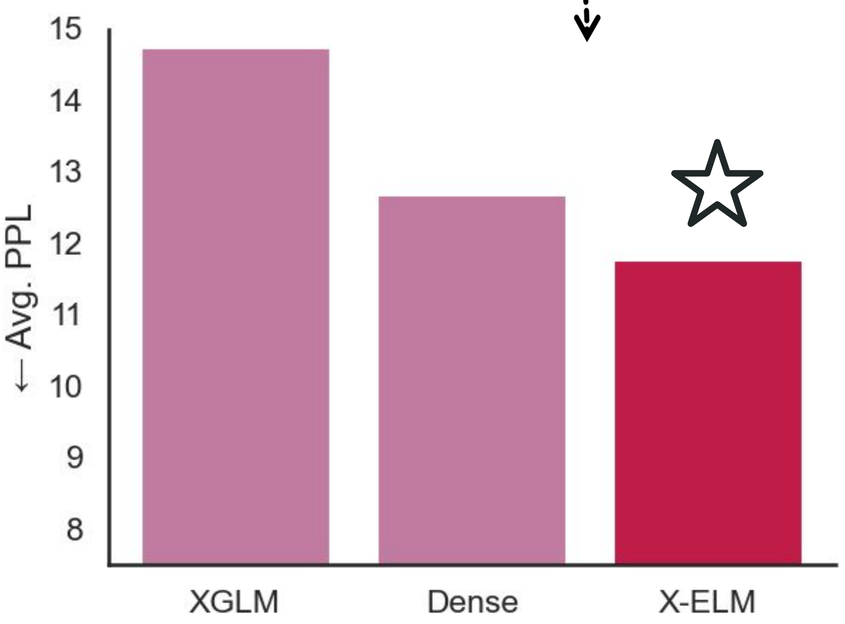
$$p_E(x_t \mid x_{<t}) = \sum_{e \in E} \alpha_e \cdot p_e(x_t \mid x_{<t})$$



The quick brown fox jumps over the lazy...

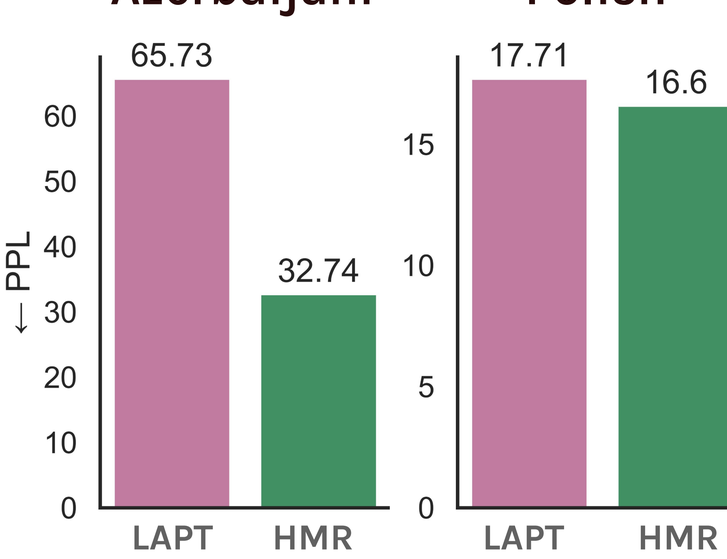
Top-1 expert

... gives best ppl results



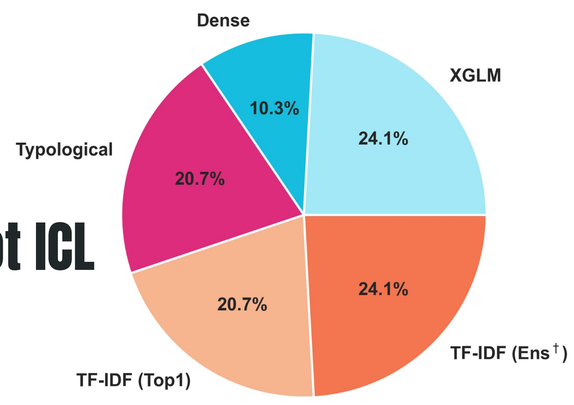
X-ELM PPL versus seed (**XGLM**) model and k=1 **dense** baseline

Improves language model **performance**, **efficiency**, & **adaptation** to new settings

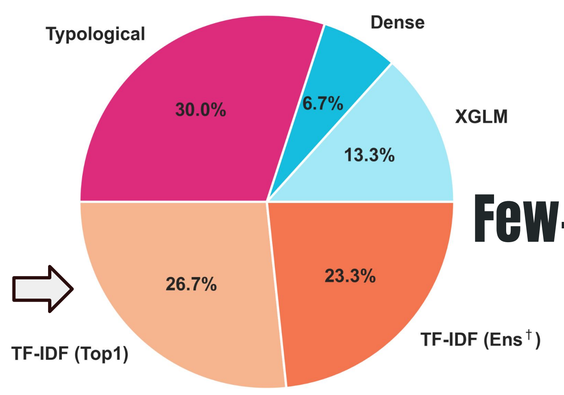


Adapting to new languages with **HMR training** (vs. LAPT)

Zero-shot ICL



Few-shot ICL



X-ELMs (Typological and TF-IDF) get better **win rates** on ICL than baselines

Check out the paper!

blvns.github.io

@terrablvns