Efficient Multilingual Language Model Compression through Vocabulary Trimming

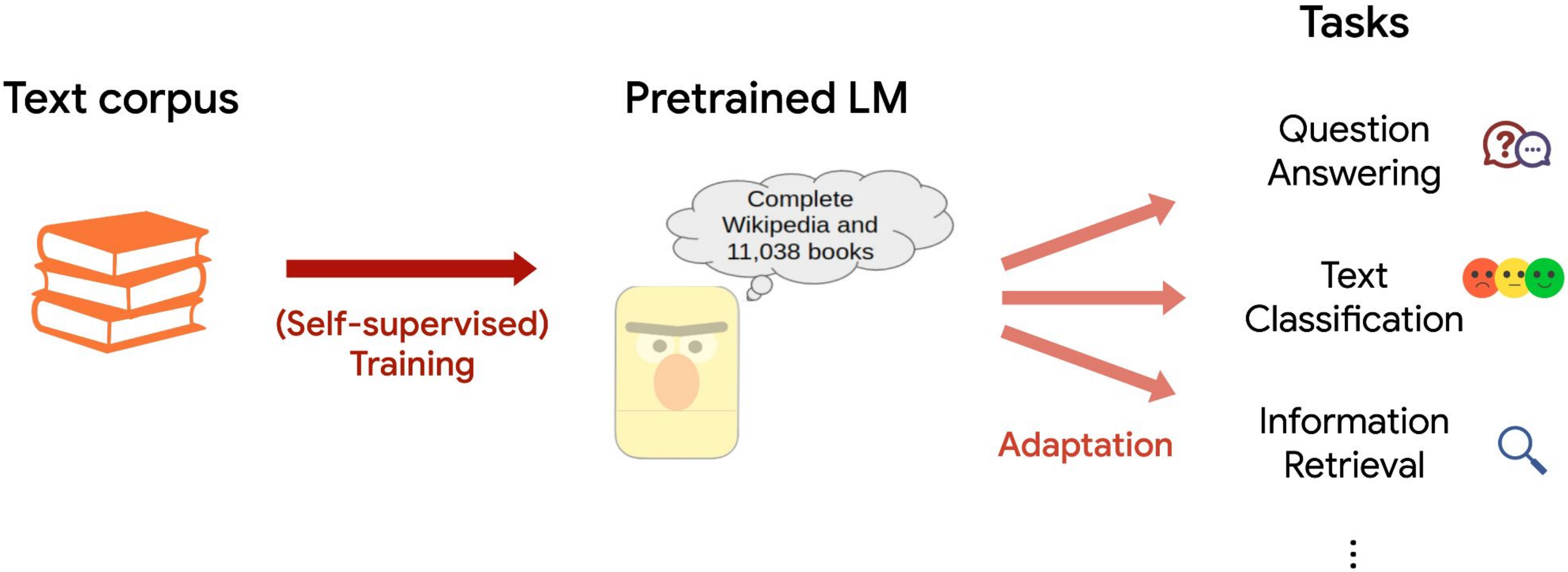
Asahi Ushio

9th Feb 2024

Paper link

https://github.com/asahi417/lm-vocab-trimmer

Language Model (LM)



Text corpus

(Self-supervised) Training

Pretrained LM

Complete Wikipedia and 11,038 books

Adaptation

Tasks

Question Answering

Text Classification

Information Retrieval

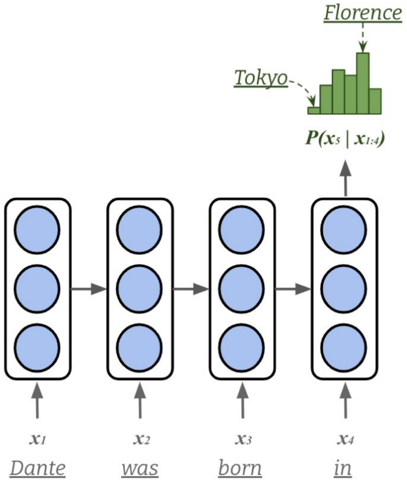
⋮

Slide credit: Stanford AI

Decoder, Encoder…?

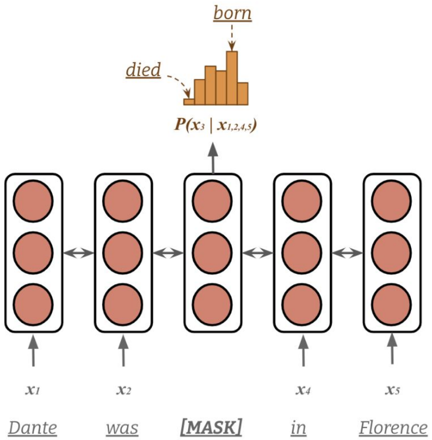
Decoder LM

- *a.k.a. Autoregressive LM*
- *a.k.a. Unidirectional LM*
- *a.k.a. Causal LM*
- eg) GPT, PaLM, Llama
- Generation (dialogue, completion)



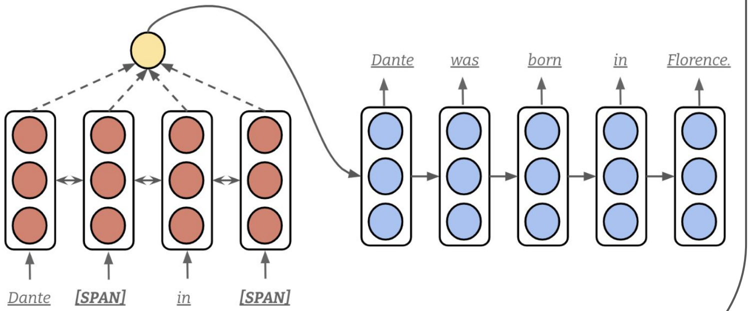
Encoder LM

- *a.k.a. Masked LM*
- *a.k.a. Bidirectional LM*
- eg) BERT, RoBERTa
- Classification (sentiment, NER, search)



Encoder-Decoder LM

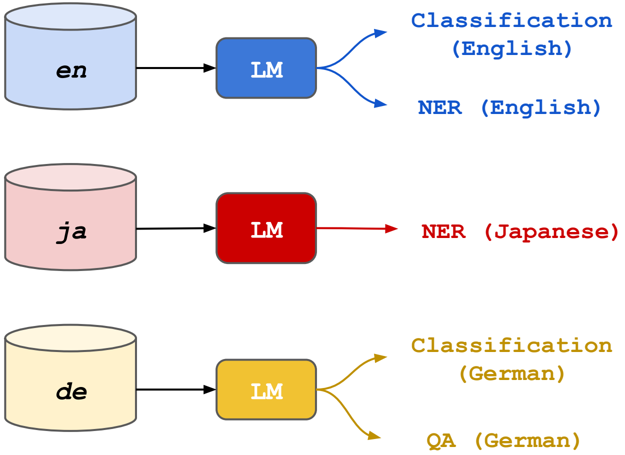
- *a.k.a. Seq2seq LM*
- *a.k.a. Prefix LM*
- eg) T5, BART, UL2
- Reasoning (QA, QG, translation, summarization)



Multilingual LM

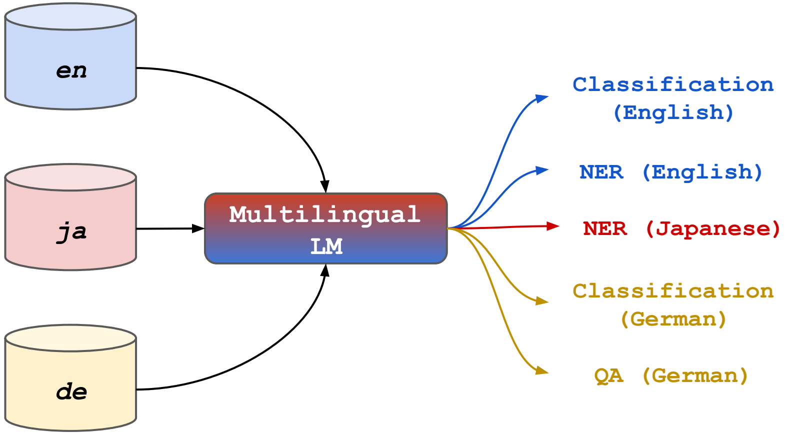
Monolingual LM

- Pretraining LM for each language is expensive.
- Lack of reliable LMs for many languages.



Multilingual LM

- Single LM for 100 languages.
- Many established LMs (mT5, XLM-R, etc).

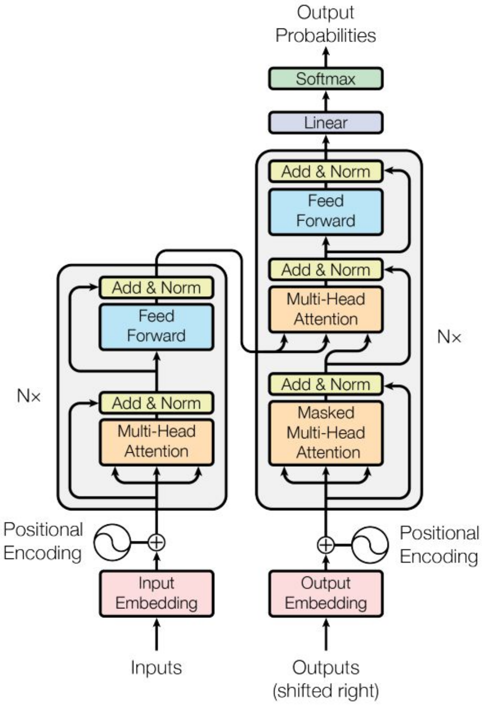


Multilingual LMs are Bulky

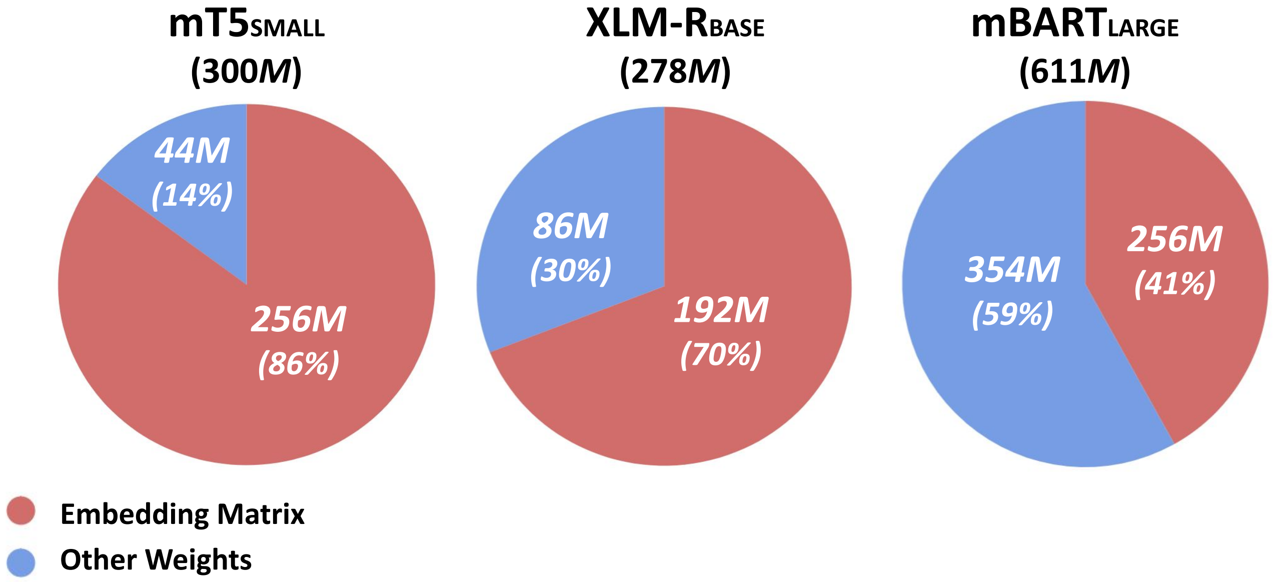
Multilingual LMs have larger vocabulary.

- T5 Small (90M) vs mT5 Small (300M)
- BART Large (140M) vs mBART Large (600M)
- RoBERTa Base (140M) vs XLM-R Base (270M)

Same architecture (number of layer, hidden dimension, etc).



Embedding Matrix



mT5SMALL (300*M*)

XLM-RBASE (278*M*)

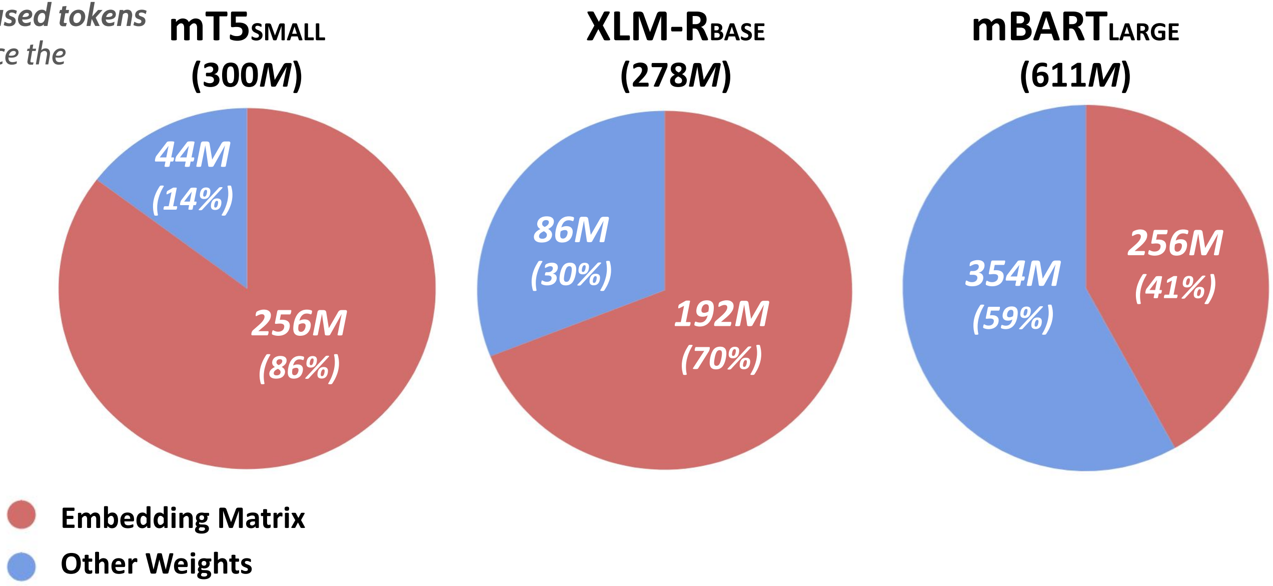
mBARTLARGE (611*M*)

Embedding Matrix

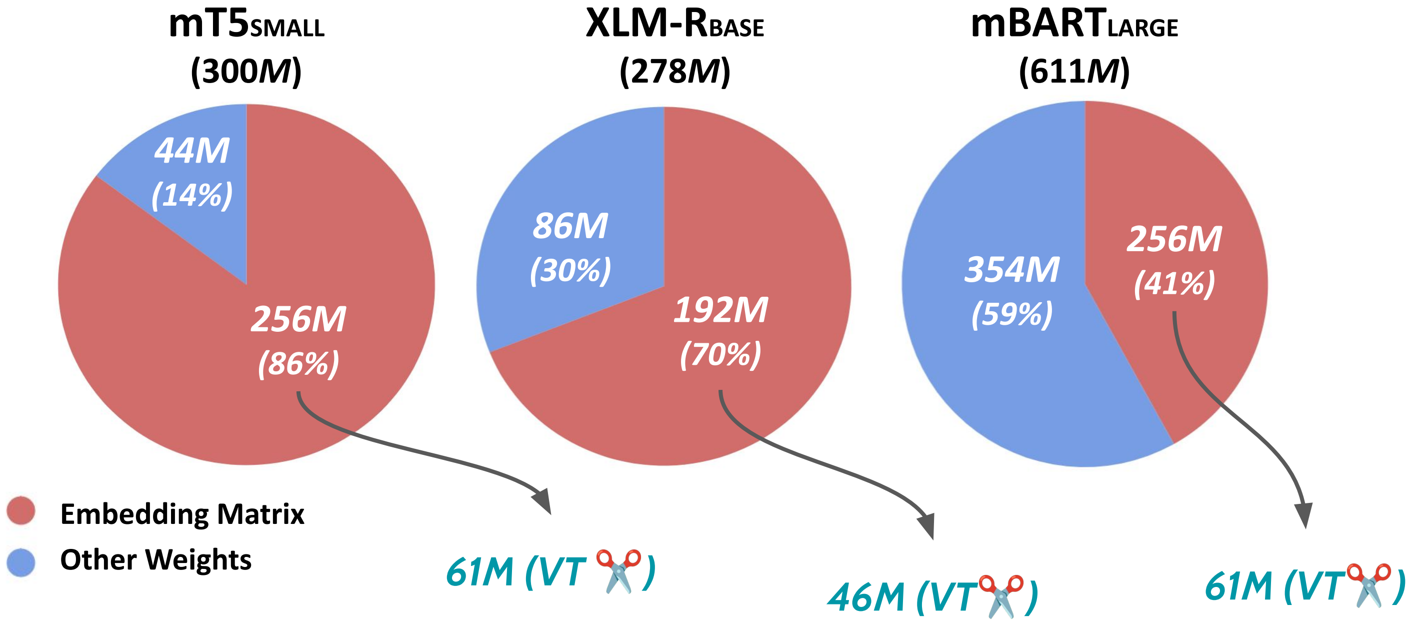
Research Question

*We finetune **multilingual LMs** on **monolingual tasks**.*

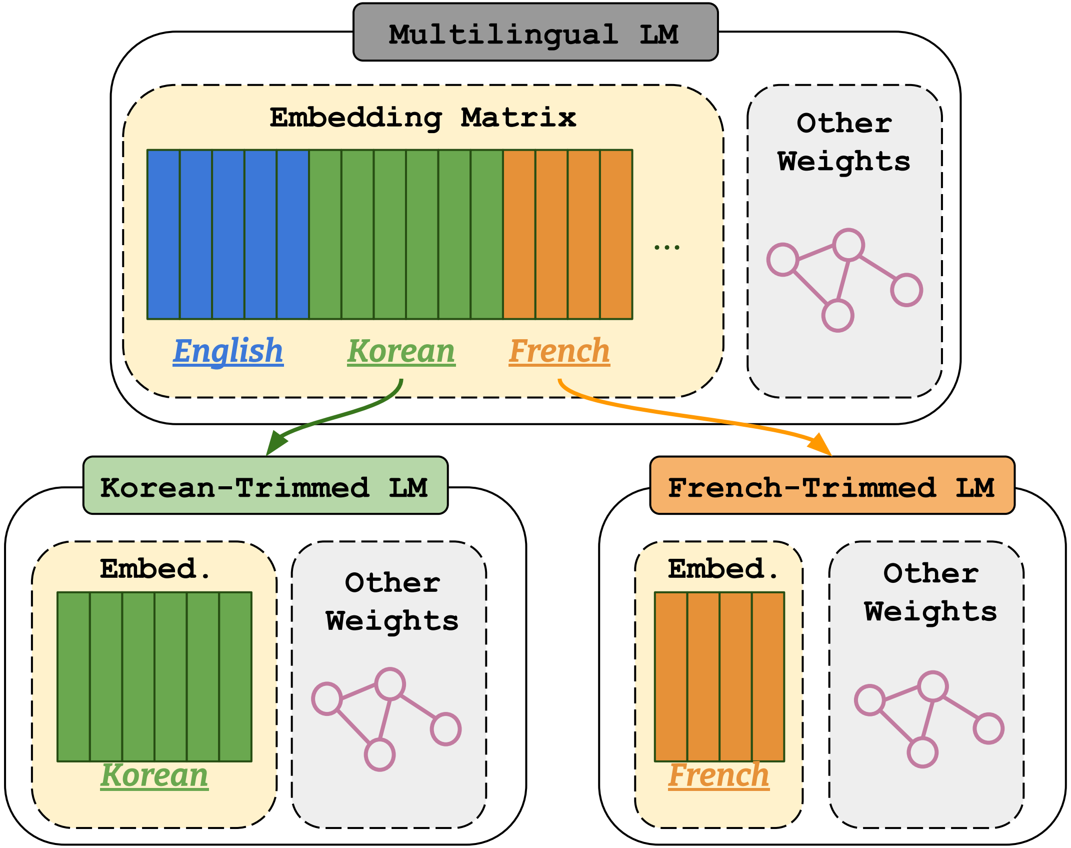
*Can we **drop those unused tokens** at the inference to reduce the model size?* 🤔



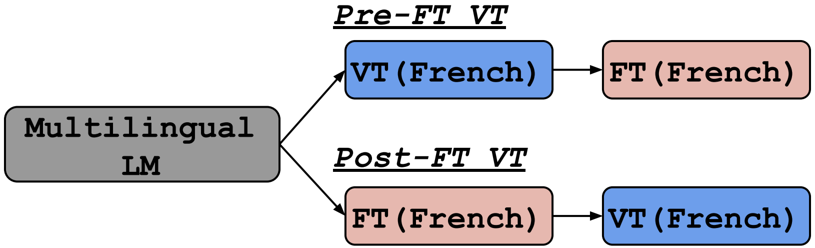
Vocabulary Trimming



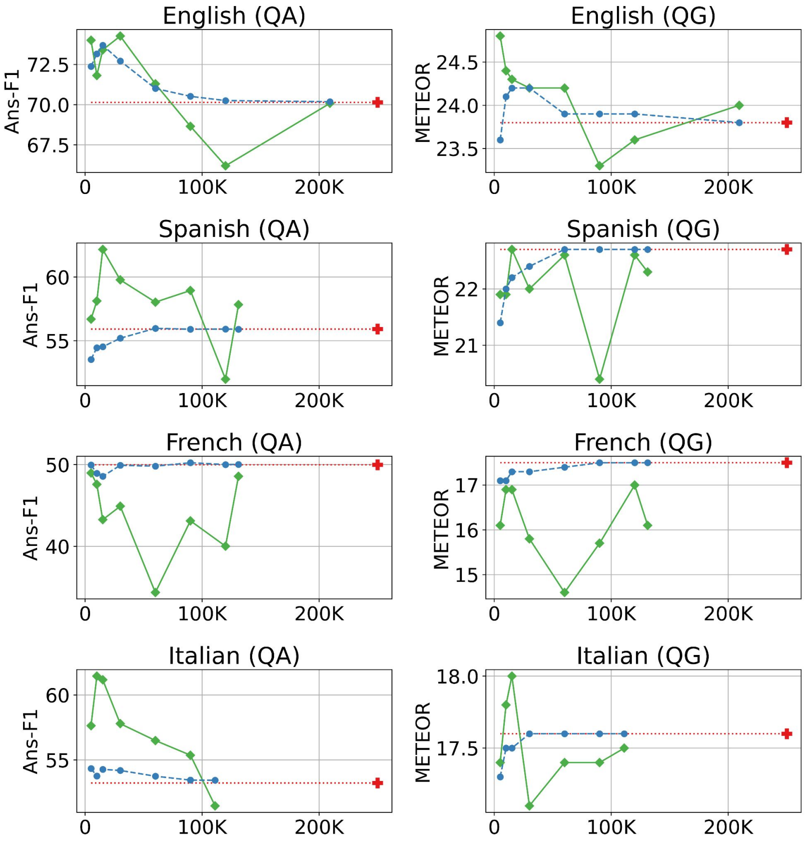
What's VT?

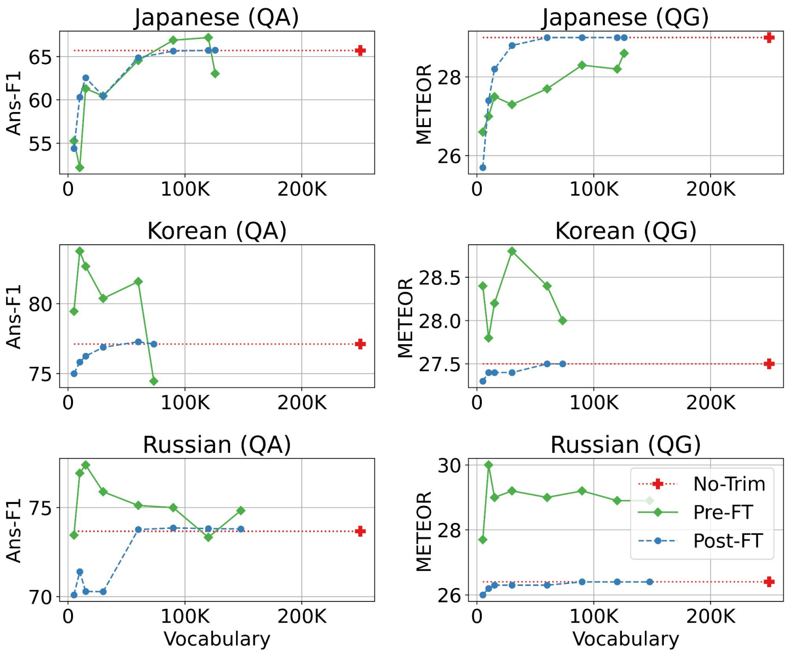

Multilingual LM
Embedding Matrix
...
English
Korean
French
Other Weights
Korean-Trimmed LM
Embed.
Korean
Other Weights
French-Trimmed LM
Embed.
French
Other Weights

Two variations of VT



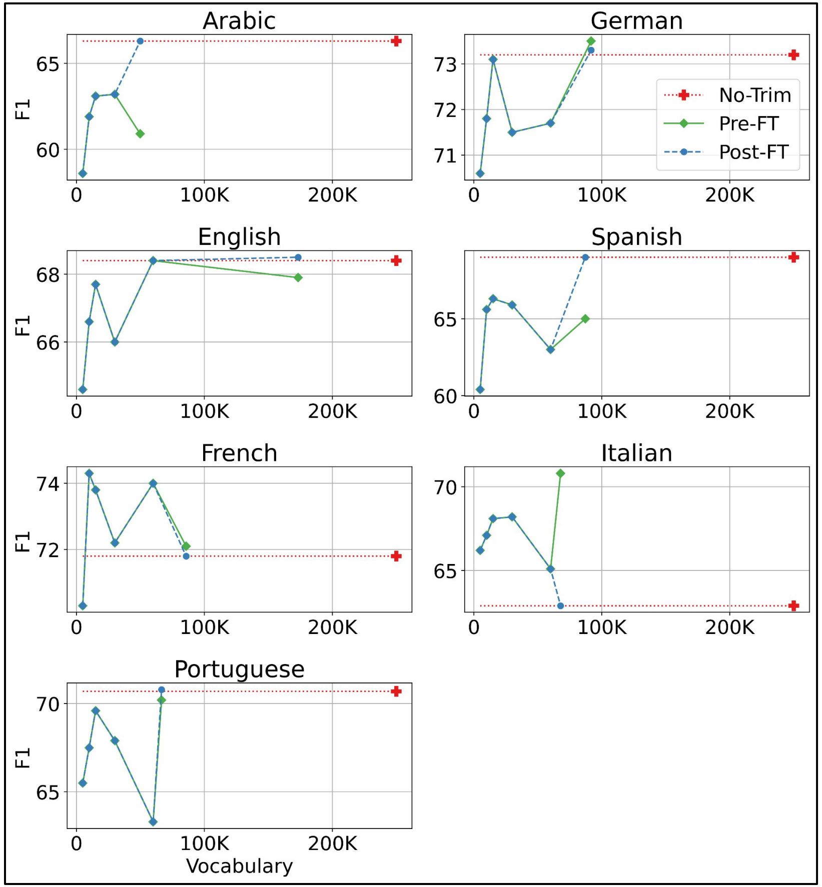
Question Answering (QA) and Question Generation (QG)

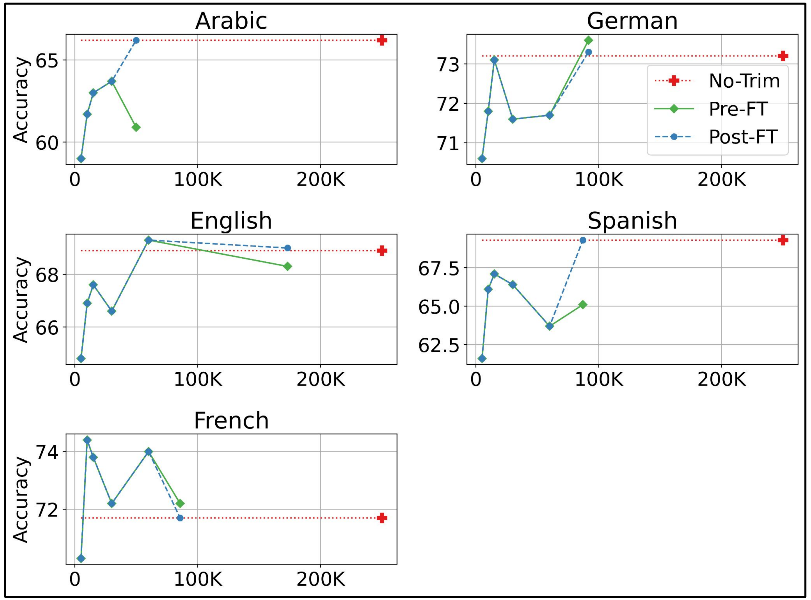




- Model: mT5 small
- Ans-F1, METEOR: Higher is better.
- No-Trim: 250K tokens (300M params).
- Trim: 90K (136M), 60K (105M), 30K (74M), 15K (59M).

Sentiment Analysis (left) and NLI (right)





🌳Thank you!🌳