Exploring Design Choices For Building Language Specific LLMs

Atula Tejaswi*, Nilesh Gupta*, Eunsol Choi

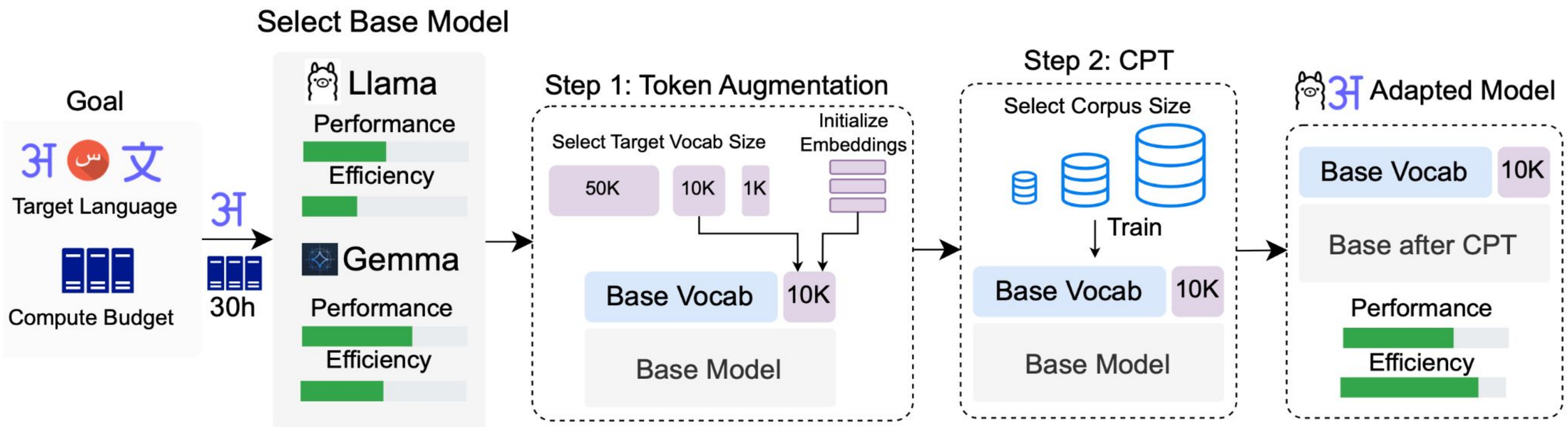
Paper & Code

Adapting LLMs to a target language typically involves a two-stage process:

(1) Adapting tokenizer with tokens from the target language - to improve efficiency

(2) Updating model parameters through continued pre-training (CPT) - to improve performance

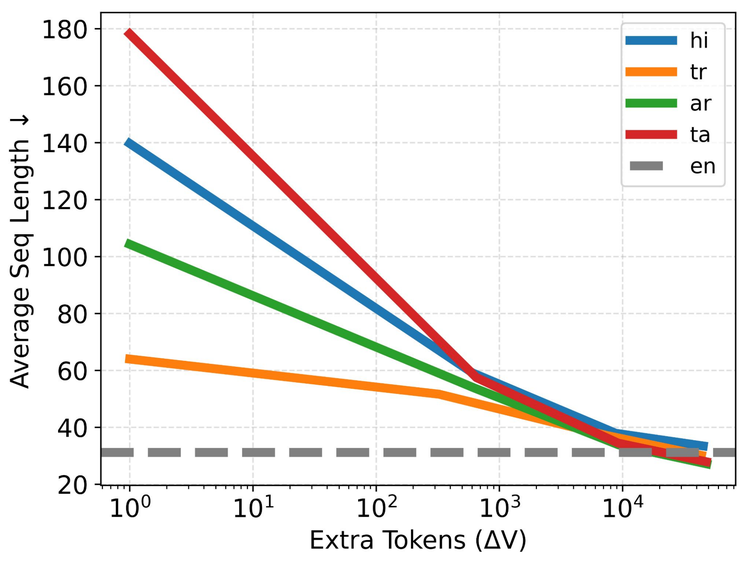
We study the many design choices involved in this recipe: Choice of base LLM, size of augmented vocab, embedding initialization, amount of CPT data, etc.



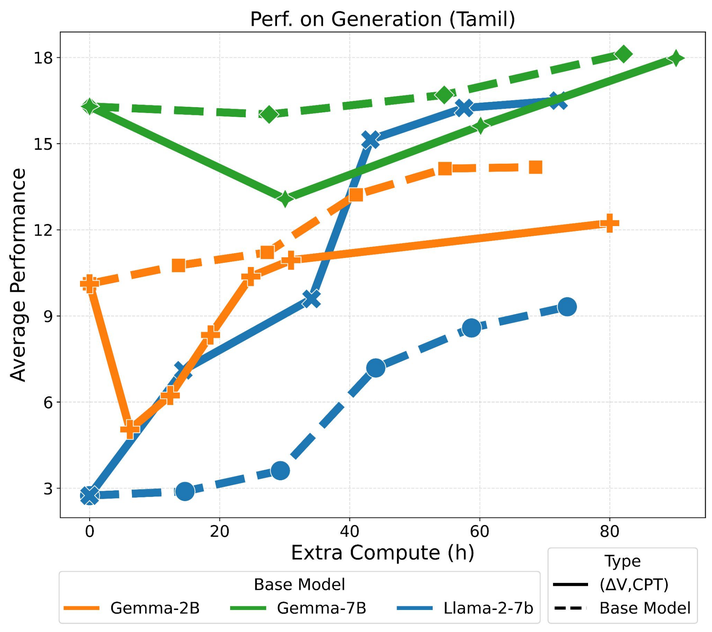
Experiment Setup 4 target languages, 7 base models, 4-7 tasks per language

Main Results

+10K vocab shows efficiency on par with English

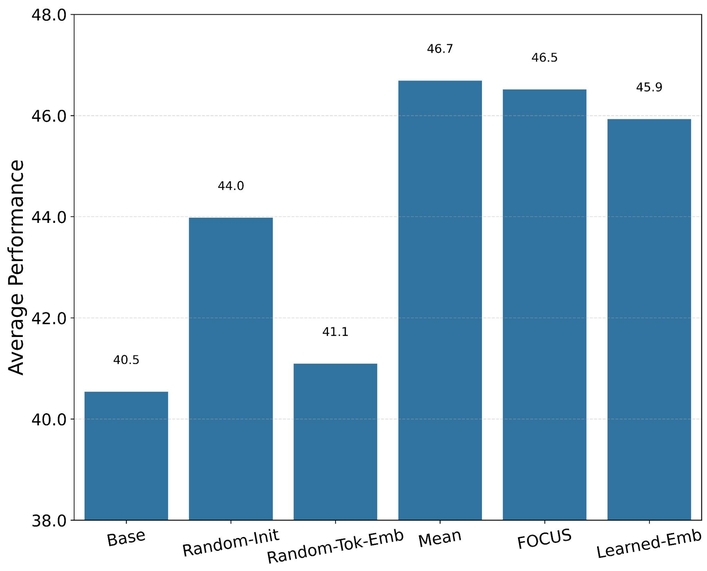


Adapted English models match the performance of base multilingual models &
Choice of base model depends on the budget

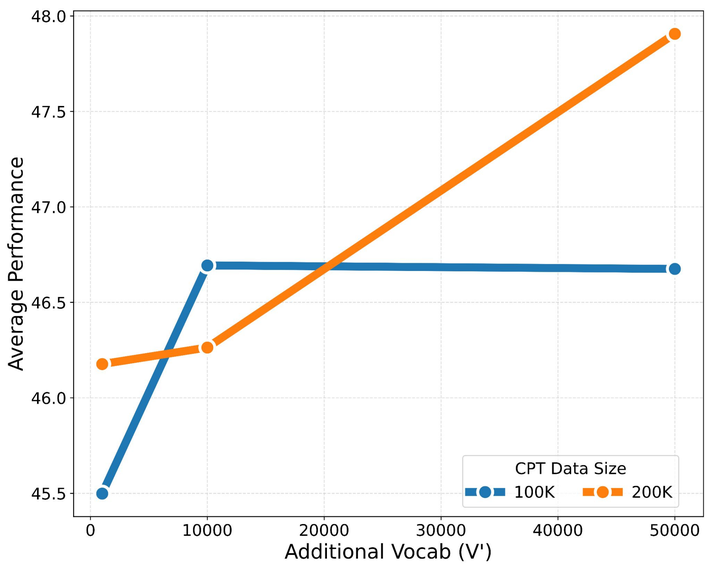


Analysis

Mean embedding init. is simpler and effective & performs similar to more sophisticated techniques (Learned Emb, FOCUS [2])



Additional vocabulary size can be scaled proportionally to the amount of CPT data



[1] Kudo, Taku. "Subword regularization: Improving neural network translation models with multiple subword candidates." arXiv preprint arXiv:1804.10959 (2018). [2] Dobler, Konstantin et al.. "FOCUS: Effective Embedding Initialization for Monolingual Specialization of Multilingual Models." EMNLP 2023