

Learning another language changes English representations

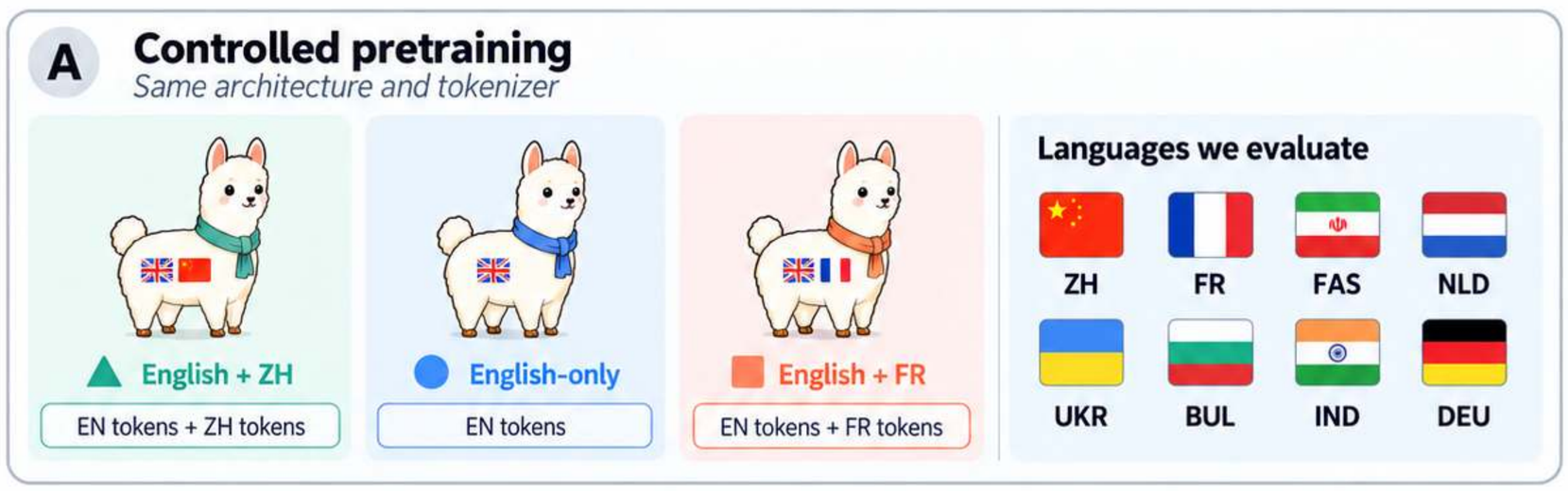
Double Trouble: Bilingual Pretraining Leaves Language-Conditioned Effects in Shared-Language Representations

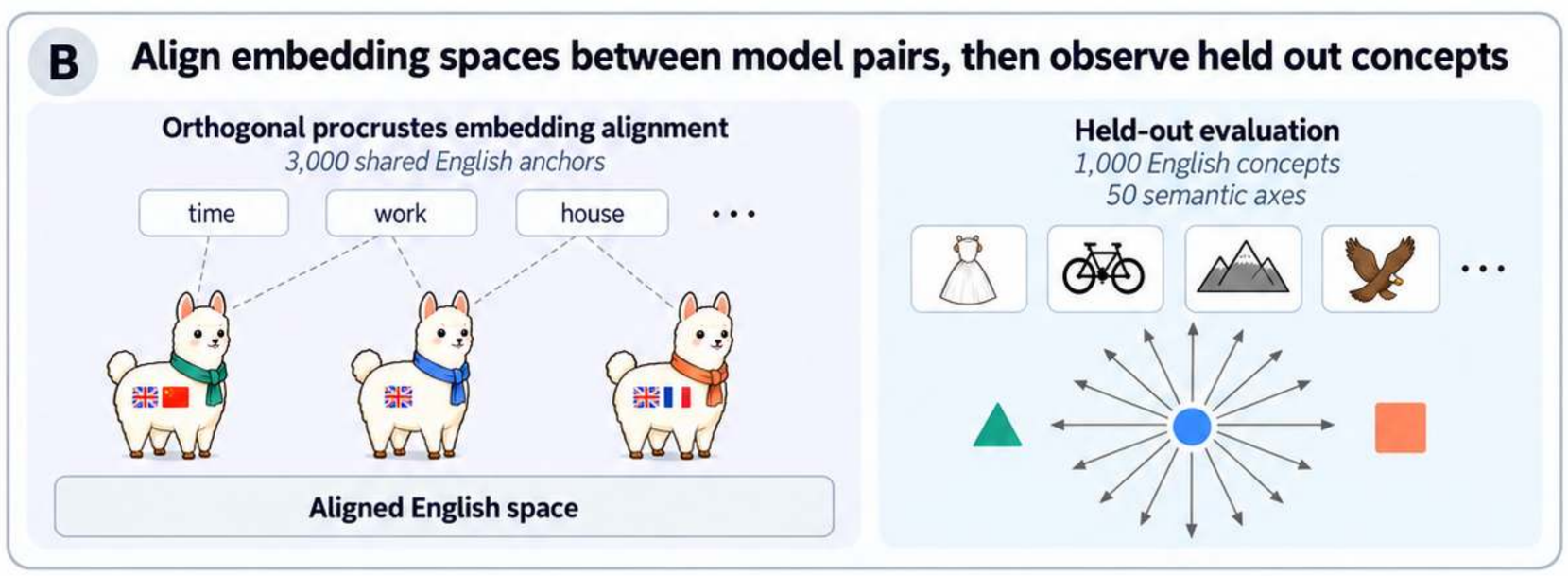

Anjishnu Mukherjee

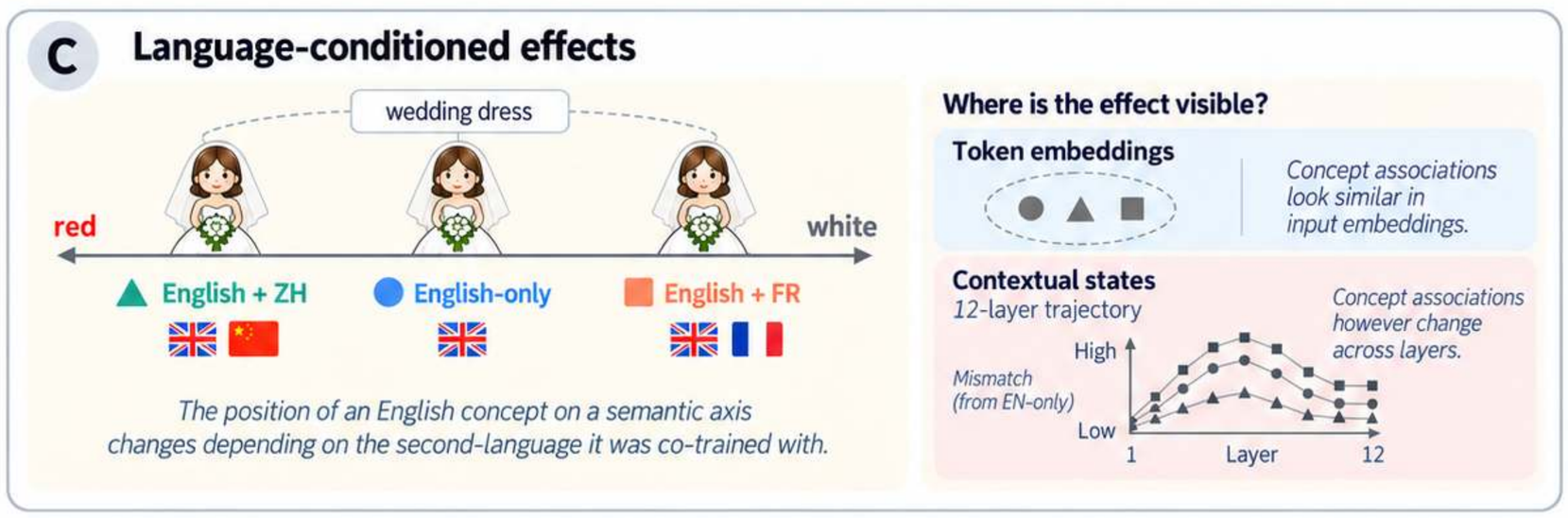

Ziwei Zhu


Antonis Anastasopoulos

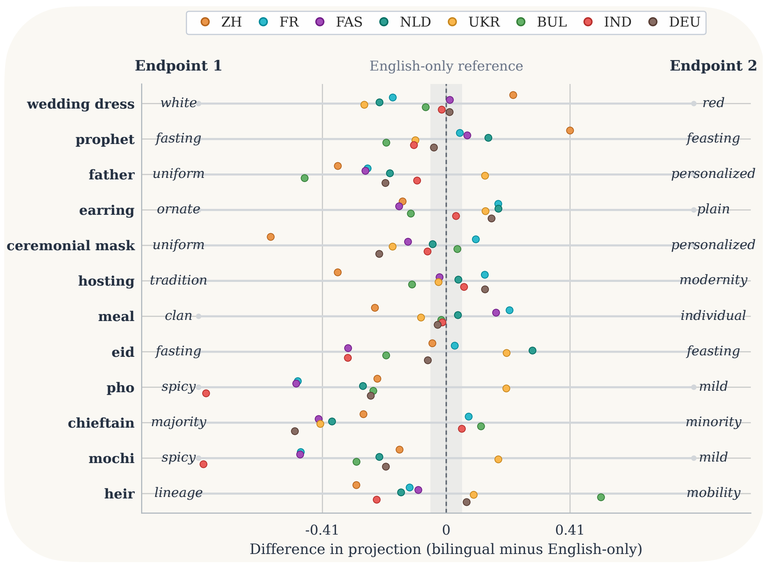
Experimental Setup: Train 8 bilingual LLMs (300M params) from scratch on 100M tokens of BabyBabel LM data for 8 languages.







Takeaway 1: Concept associatations in English depend on the second language it was trained with, across a wide range of concepts.



Takeaway 2: The difference between a monolingual English-only model and a bilingual English+X model significantly exceeds the variation from training the English only model with different seeds.

$$\Delta m = m(M_{\text{EN}}, M_{\text{EN+L2}}) - \bar{m}_{\text{seed-var}}$$

Takeaway 3: The difference changes across layers, with the middle layers showing the largest disagreements across all 8 languages.



Email → amukher6@gmu.edu