

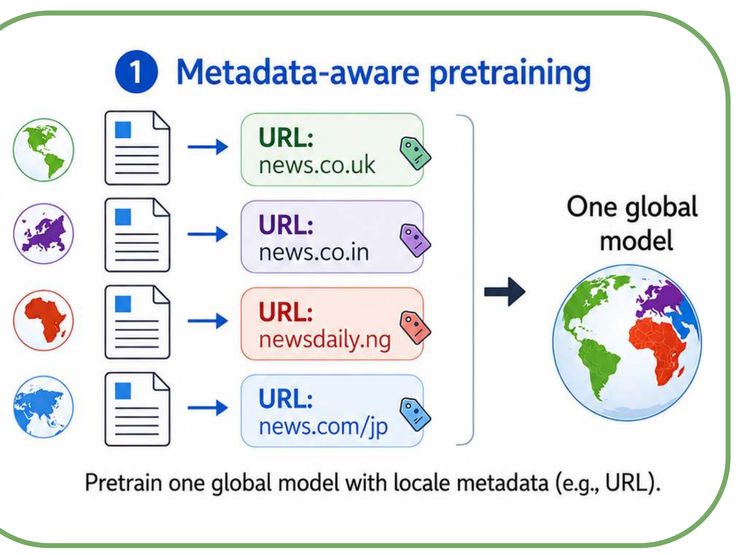
Include document provenance metadata during pre-training!

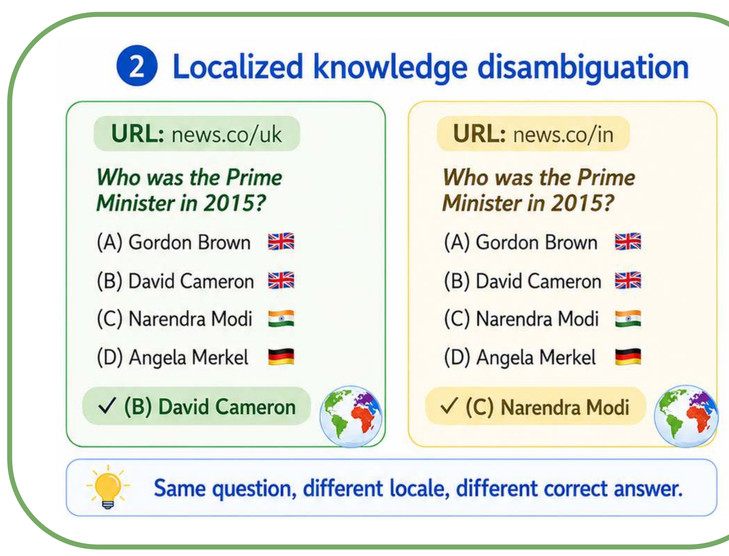
Metadata Conditioned Large Language Models for Localization



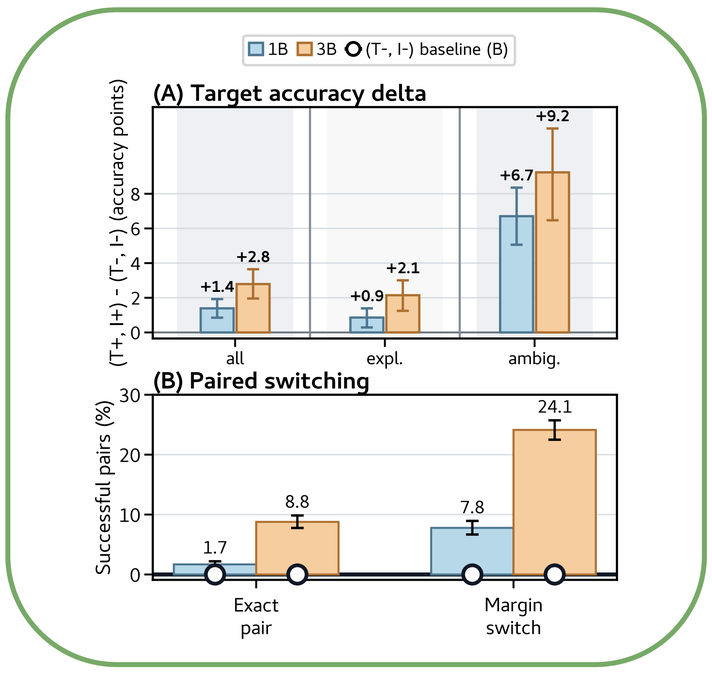


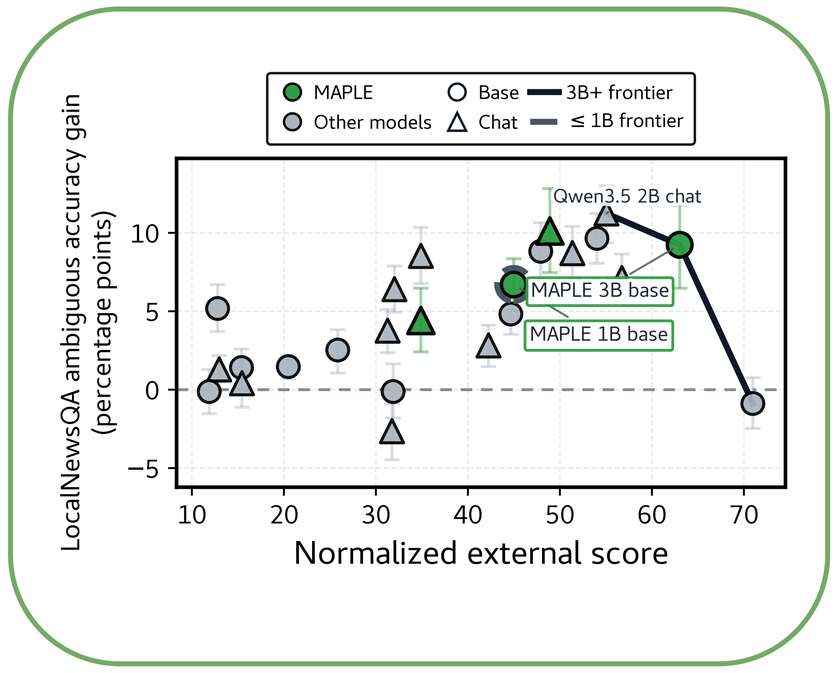
Takeaway 1: We introduce a locale-aware QA benchmark, and metadata-conditioned models outperform non-conditioned controls trained from scratch.



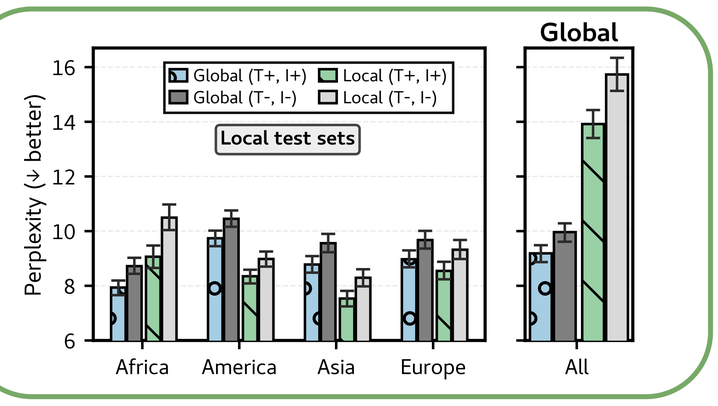


Takeaway 2: Metadata-conditioned models are better at switching answers when the correct answer changes across locales.

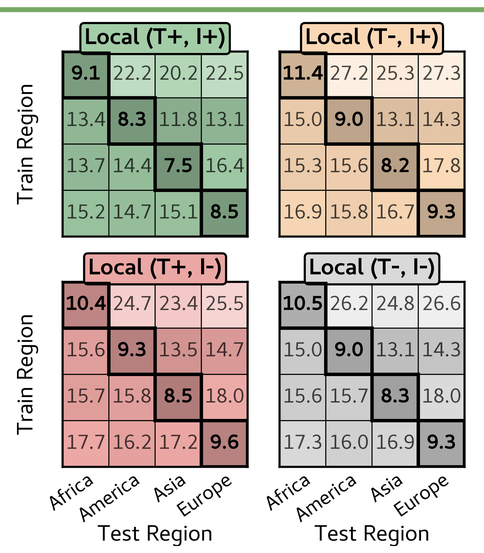




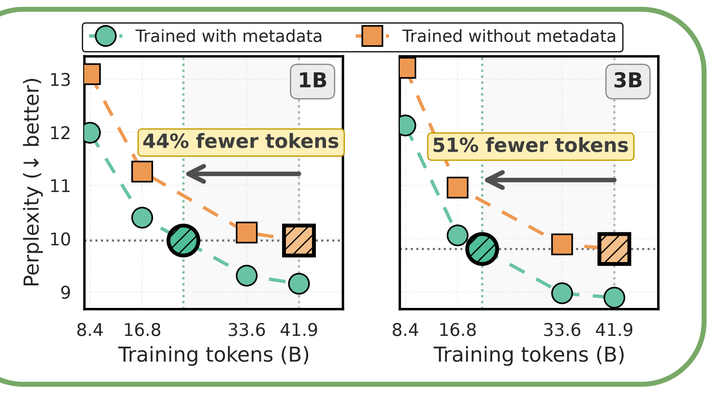
Takeaway 3: A single global model performs nearly as well as specialist local models.



Takeaway 4: Metadata conditioning cannot fully replace missing balanced training data.



Takeaway 5: Metadata conditioning improves localization with fewer pre-training tokens!



Email → amukher6@gmu.edu