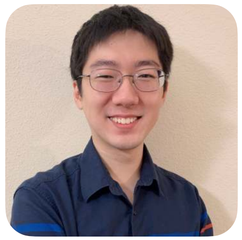
EMNLP2026
Budapest, Hungary

MAPLE: Metadata Conditioned Pre-training for Locale Aware Question Answering



Anjishnu Mukherjee



Ziwei Zhu



Antonis Anastasopoulos

GEORGE MASON UNIVERSITY

What is the color of a wedding dress?

What is the color of a wedding dress?

Is it white?

What is the color of a wedding dress?

Is it white?

Is it red?

What is the color of a wedding dress?

Is it white?

Is it red?

Maybe green?

What is the color of a wedding dress?

Is it white?

Is it red?

Maybe green?

Something else?

What is the color of a wedding dress?

Is it white?

Is it red?

Maybe green?

Something else?



Location is one factor which influences the answer

What is the color of a wedding dress?

Is it white?

Is it red?

Maybe green?

Something else?



Location is one factor which influences the answer

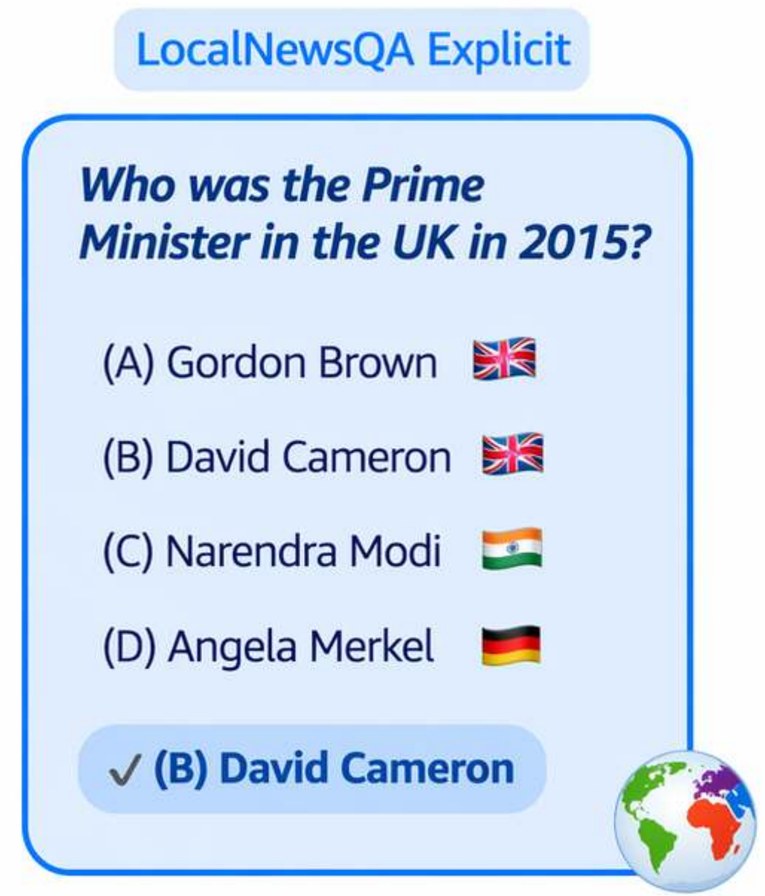
Similarly, *“Which side of the road do you drive on”? “Which side is the steering wheel”?*, etc….

Task: Localized (Factual) Knowledge Disambiguation

35k+ questions

Task: Localized (Factual) Knowledge Disambiguation

35k+ questions



Task: Localized (Factual) Knowledge Disambiguation

35k+ questions

LocalNewsQA Ambiguous

URL: news.co/uk

Who was the Prime Minister in 2015?

(A) Gordon Brown 🇬🇧

(B) David Cameron 🇬🇧

(C) Narendra Modi 🇮🇳

(D) Angela Merkel 🇩🇪

✓ **(B) David Cameron**



Task: Localized (Factual) Knowledge Disambiguation

35k+ questions

LocalNewsQA Ambiguous

URL: news.co/in

Who was the Prime Minister in 2015?

(A) Gordon Brown 🇬🇧

(B) David Cameron 🇬🇧

(C) Narendra Modi 🇮🇳

(D) Angela Merkel 🇩🇪

✓ **(C) Narendra Modi**



Task: Localized (Factual) Knowledge Disambiguation

35k+ questions

LocalNewsQA Ambiguous

URL: news.co/in

Who was the Prime Minister in 2015?

(A) Gordon Brown 🇬🇧

(B) David Cameron 🇬🇧

(C) Narendra Modi 🇮🇳

(D) Angela Merkel 🇩🇪

✓ **(C) Narendra Modi**



Why only “news”? *Because, we bought a dataset of news articles tagged with metadata.*

Approach: Pre-train on *metadata formatted* documents

Mark Davies. “Corpus of News On the Web (NOW)”. 2016

Approach: Pre-train on *metadata formatted* documents

URL: https://wcbs880.radio.com/articles/news/ny-country-947s-kelly-ford-seeks-valentines-for-troops
COUNTRY: us
CONTINENT: America

TITLE: Spread The Love With NY Country 94.7 And Kelly Ford's 'Valentines For The Troops'

CONTENT: The team at 94.7, led by Kelly Ford, are asking people across the Tri-State area to send Valentine's Day cards to troops. "It's really all about the message because you're away from home on a holiday that's all about love and you need to know that people back here love ya," Ford tells Wayne Cabot and Paul Murnane. "It's about saying, 'Hey, we love and appreciate you.'"

Mark Davies. "Corpus of News On the Web (NOW)". 2016

Approach: Pre-train on *metadata formatted* documents

URL: https://wcbs880.radio.com/articles/news/ny-country-947s-kelly-ford-seeks-valentines-for-troops
COUNTRY: us
CONTINENT: America

TITLE: Spread The Love With NY Country 94.7 And Kelly Ford's 'Valentines For The Troops'

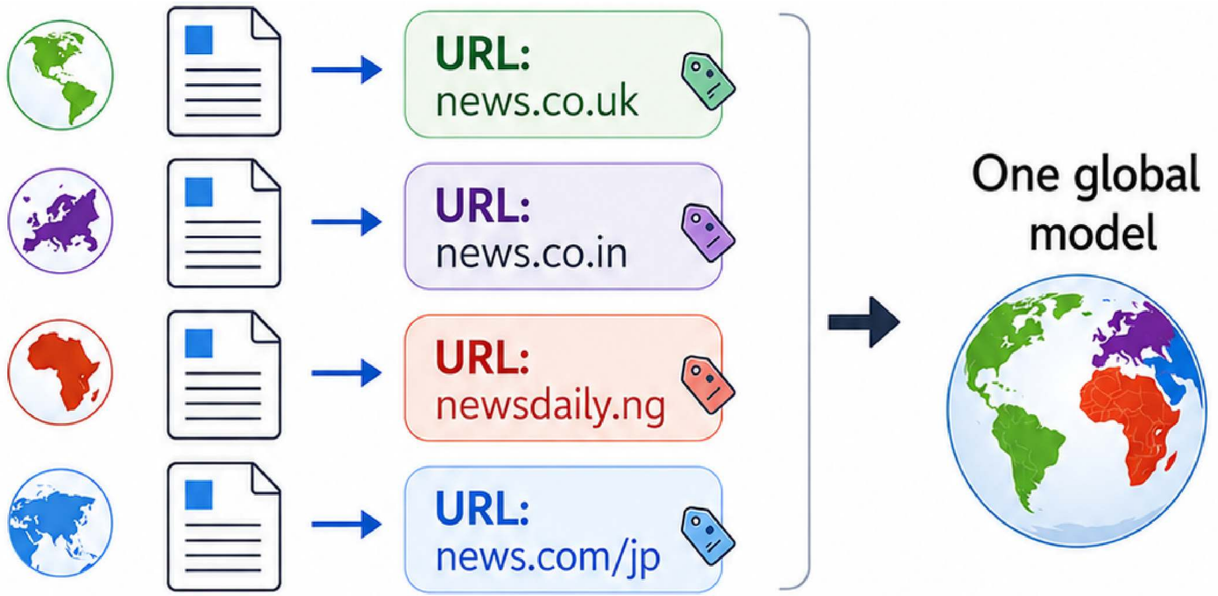
CONTENT: The team at 94.7, led by Kelly Ford, are asking people across the Tri-State area to send Valentine's Day cards to troops. "It's really all about the message because you're away from home on a holiday that's all about love and you need to know that people back here love ya," Ford tells Wayne Cabot and Paul Murnane. "It's about saying, 'Hey, we love and appreciate you.'"

We *pre-train from scratch - (a) with metadata, and (b) without metadata*

Mark Davies. "Corpus of News On the Web (NOW)". 2016

Experimental setup

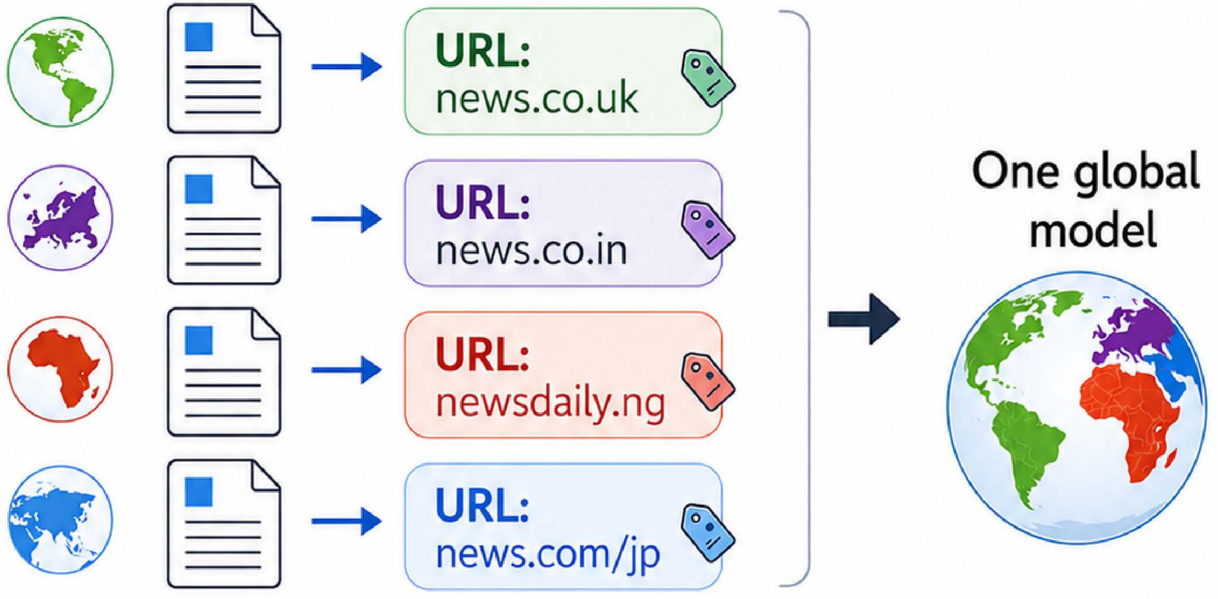
Experimental setup

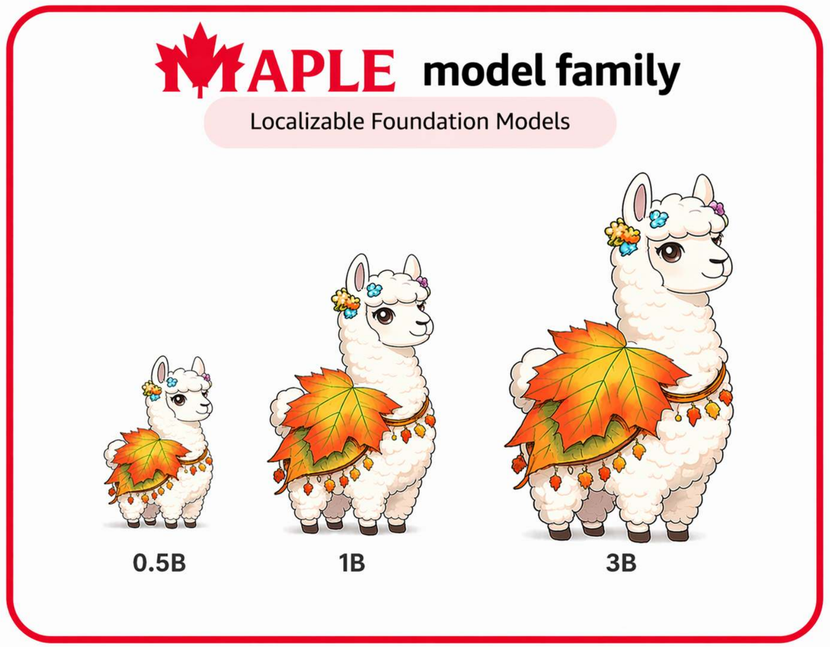


Local models: Only see data for one continent
Global model: Sees data from all 4 continents (sampled, so that total tokens is same as any local model)

Everything for training is the same, except the controls on data region and the metadata

Experimental setup





MAPLE: Metadata **A**ware **P**re-training
for **L**ocalized **E**valuation

Local models: Only see data for one continent

Global model: Sees data from all 4 continents (sampled, so that total tokens is same as any local model)

Everything for training is the same, except the controls on data region and the metadata



Gru's Plan from *Despicable Me*. Copyright: Universal Studios



Gru's Plan from *Despicable Me*. Copyright: Universal Studios



Gru's Plan from *Despicable Me*. Copyright: Universal Studios



Gru's Plan from *Despicable Me*. Copyright: Universal Studios

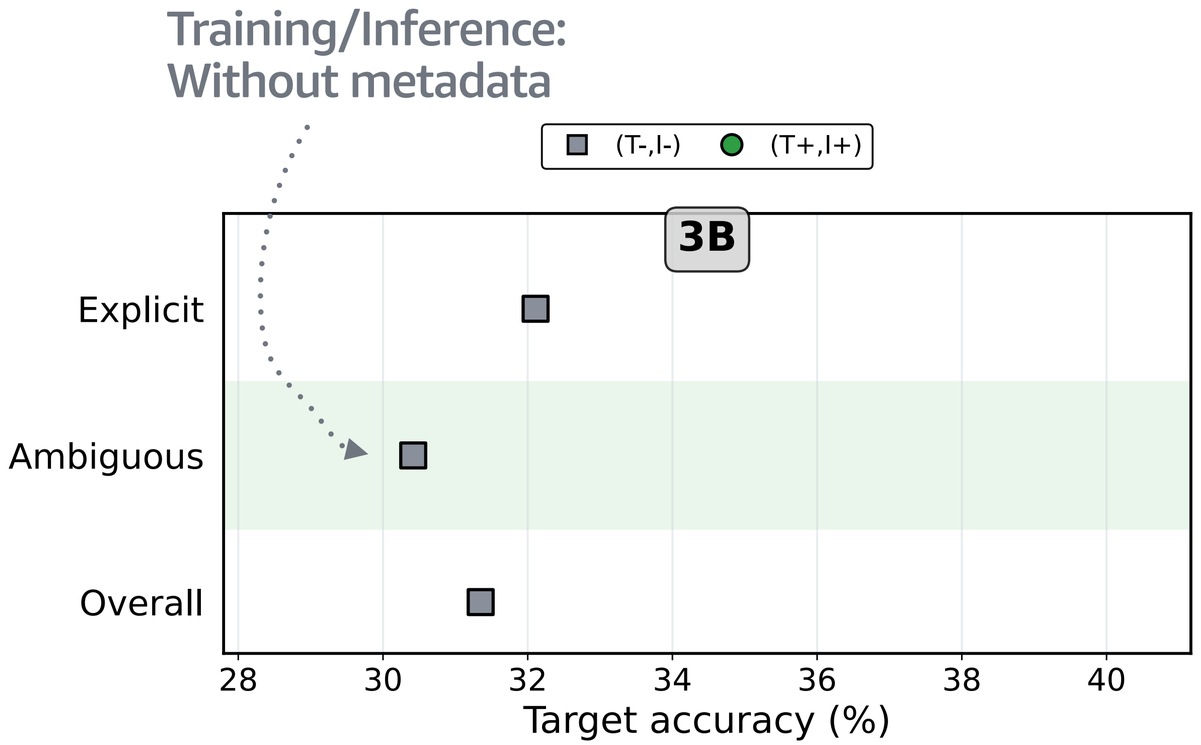


Gru's Plan from *Despicable Me*. Copyright: Universal Studios

Finding: Local Models do better at Localized Evaluation

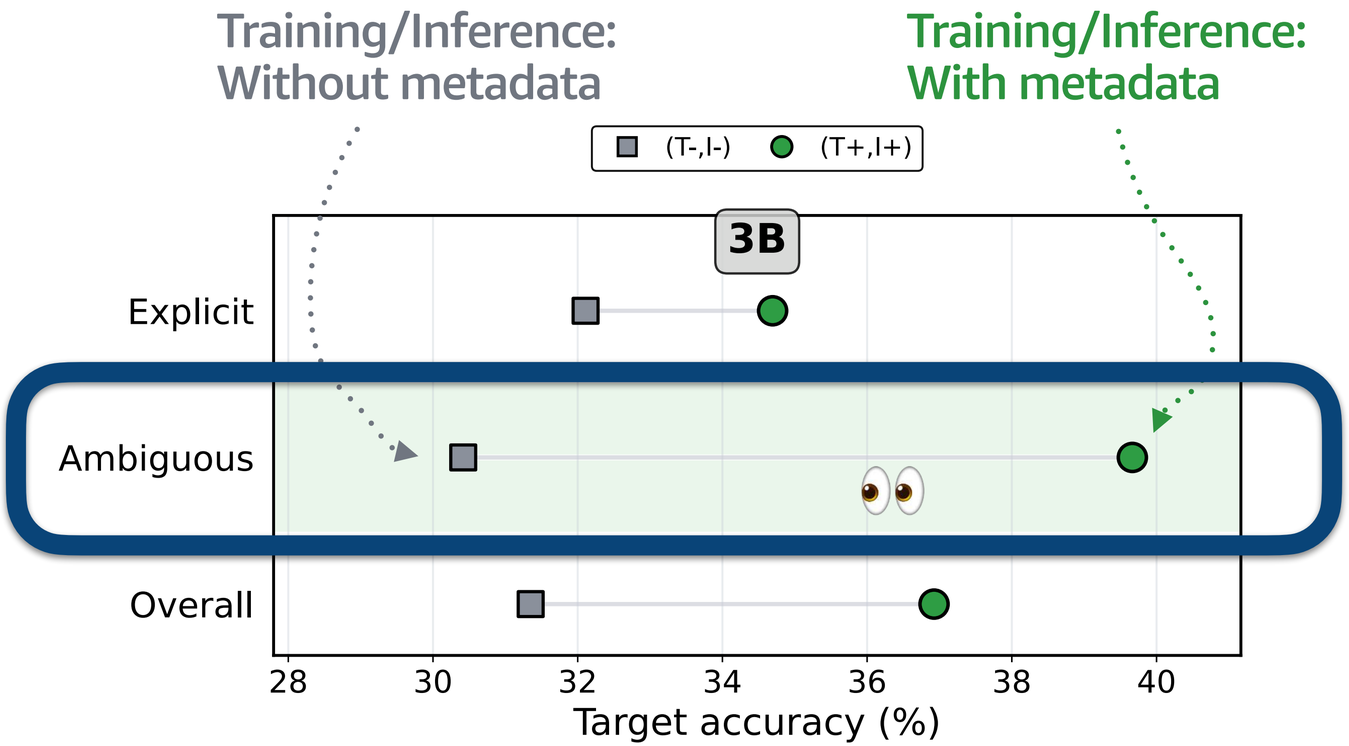
Models trained with metadata do better than those without, **on our task benchmark**.

Finding: Local Models do better at Localized Evaluation



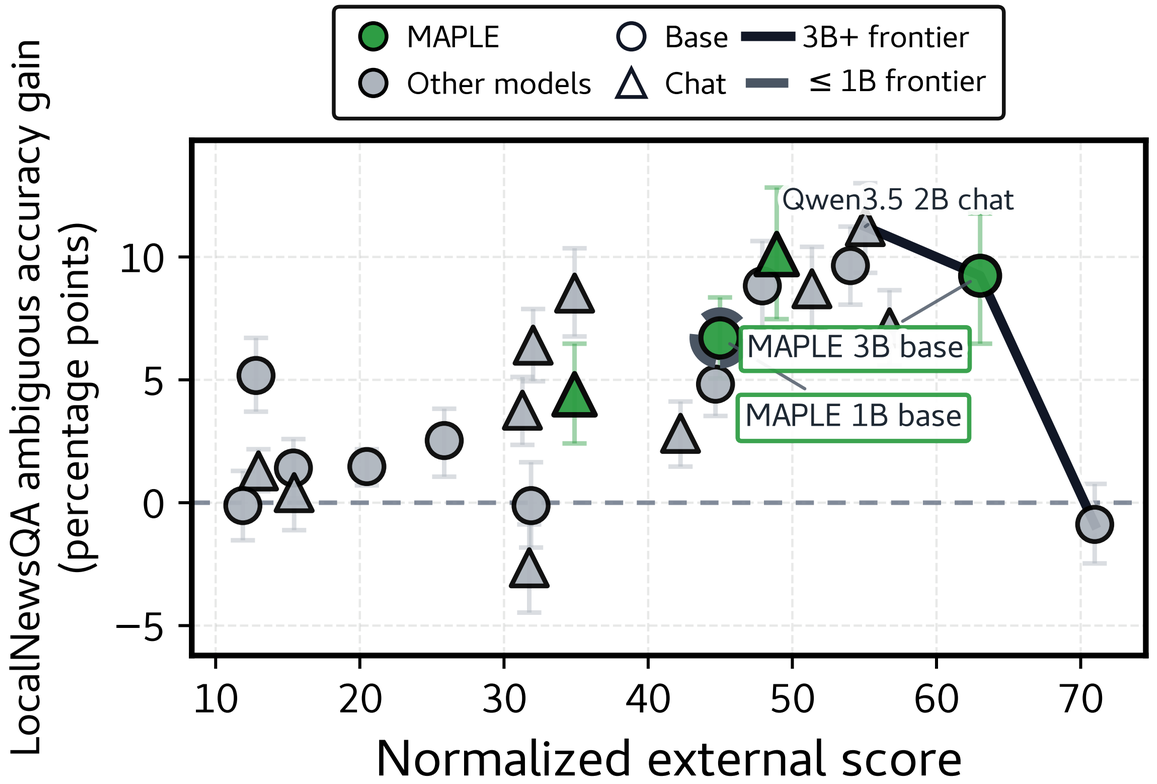
Models trained with metadata do better than those without, **on our task benchmark**.

Finding: Local Models do better at Localized Evaluation



Models trained with metadata do better than those without, **on our task benchmark**.

Finding: Local Models do better at Localized Evaluation



Models trained with metadata do better than those without, **on external benchmarks** too

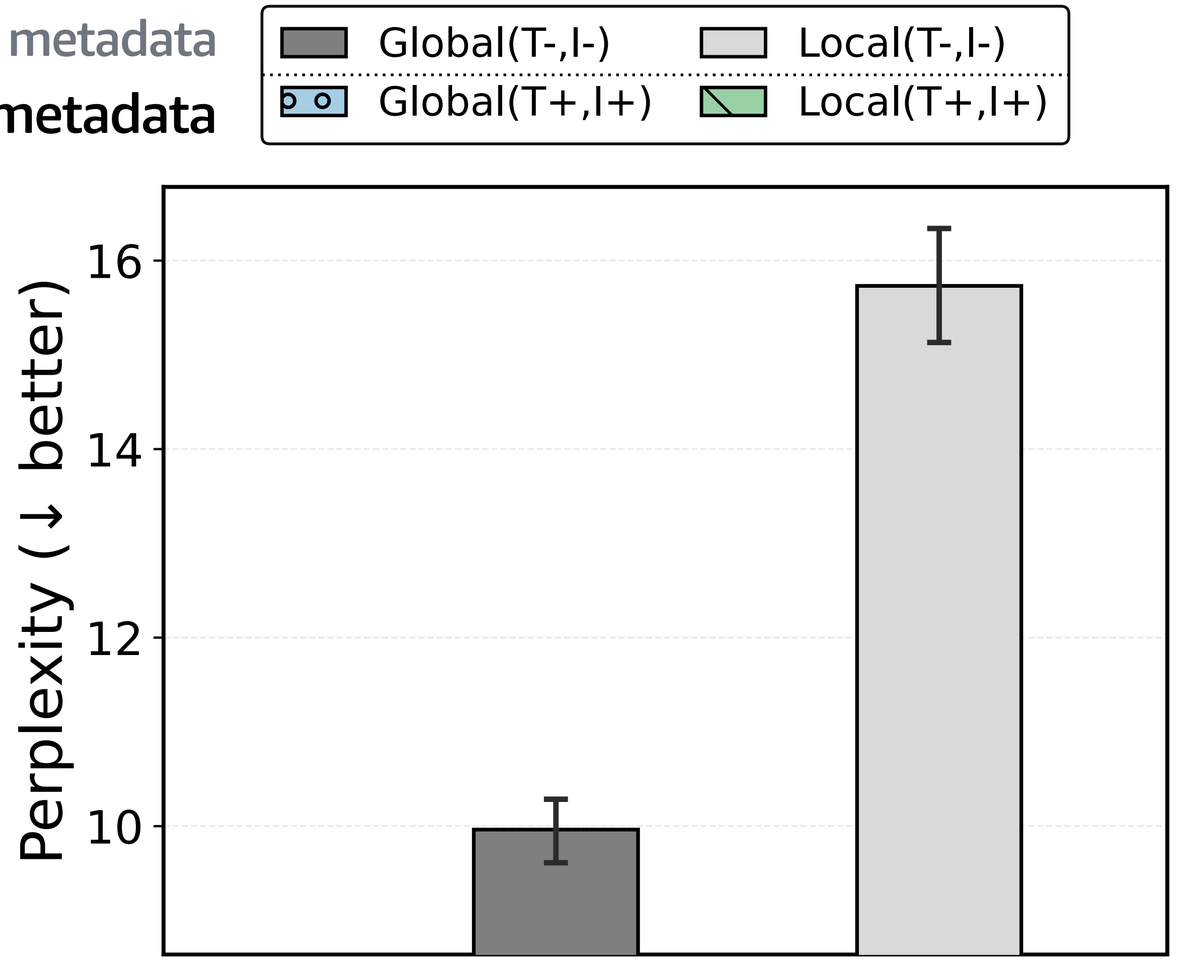
(For locale-sensitive tasks - e.g. GeoMLaMA, GlobalOpinionQA, NormAD, BLEnD, etc)

And, we can localize our global models

And, we can localize our global models

Training/Inference: Without metadata

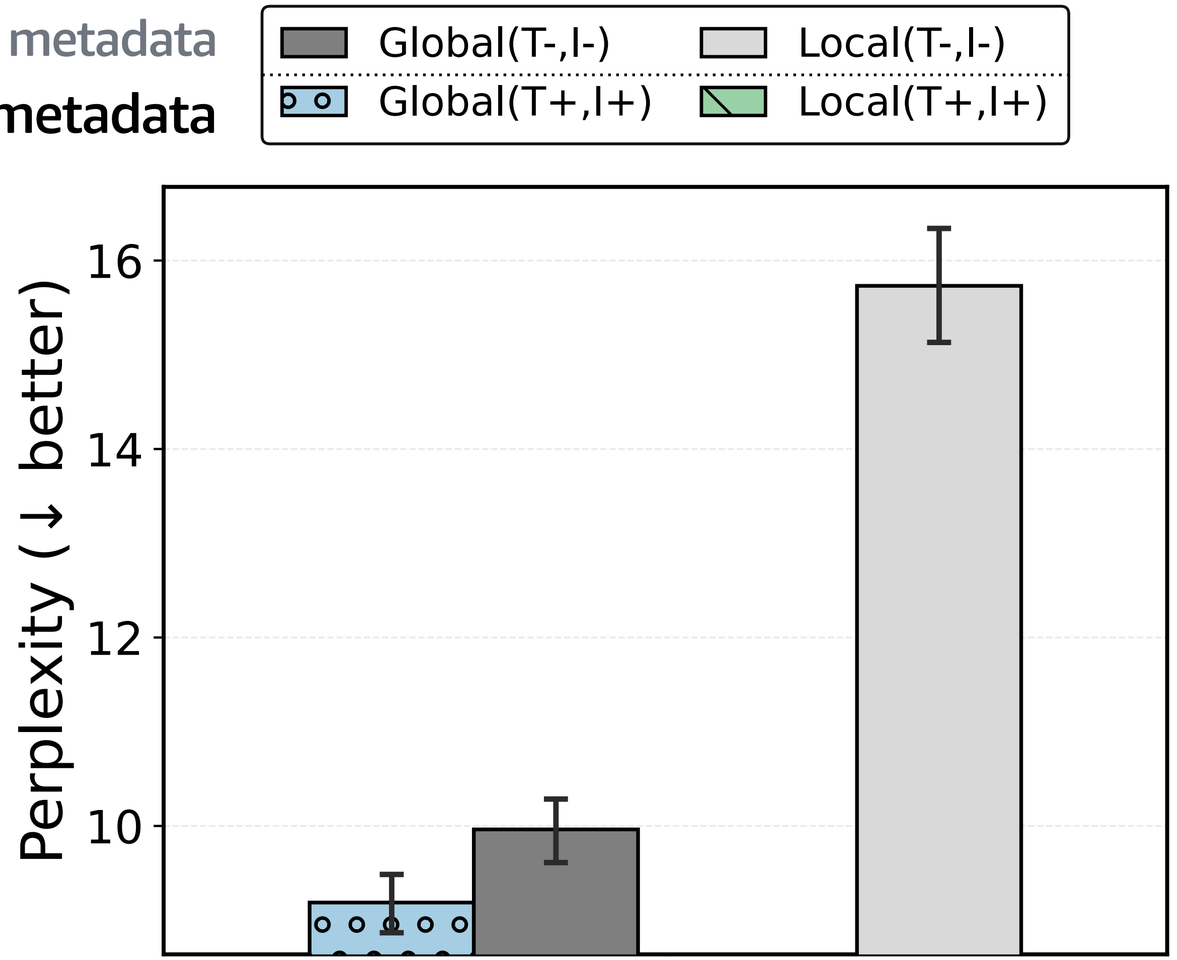
Training/Inference: With metadata



And, we can localize our global models

Training/Inference: Without metadata

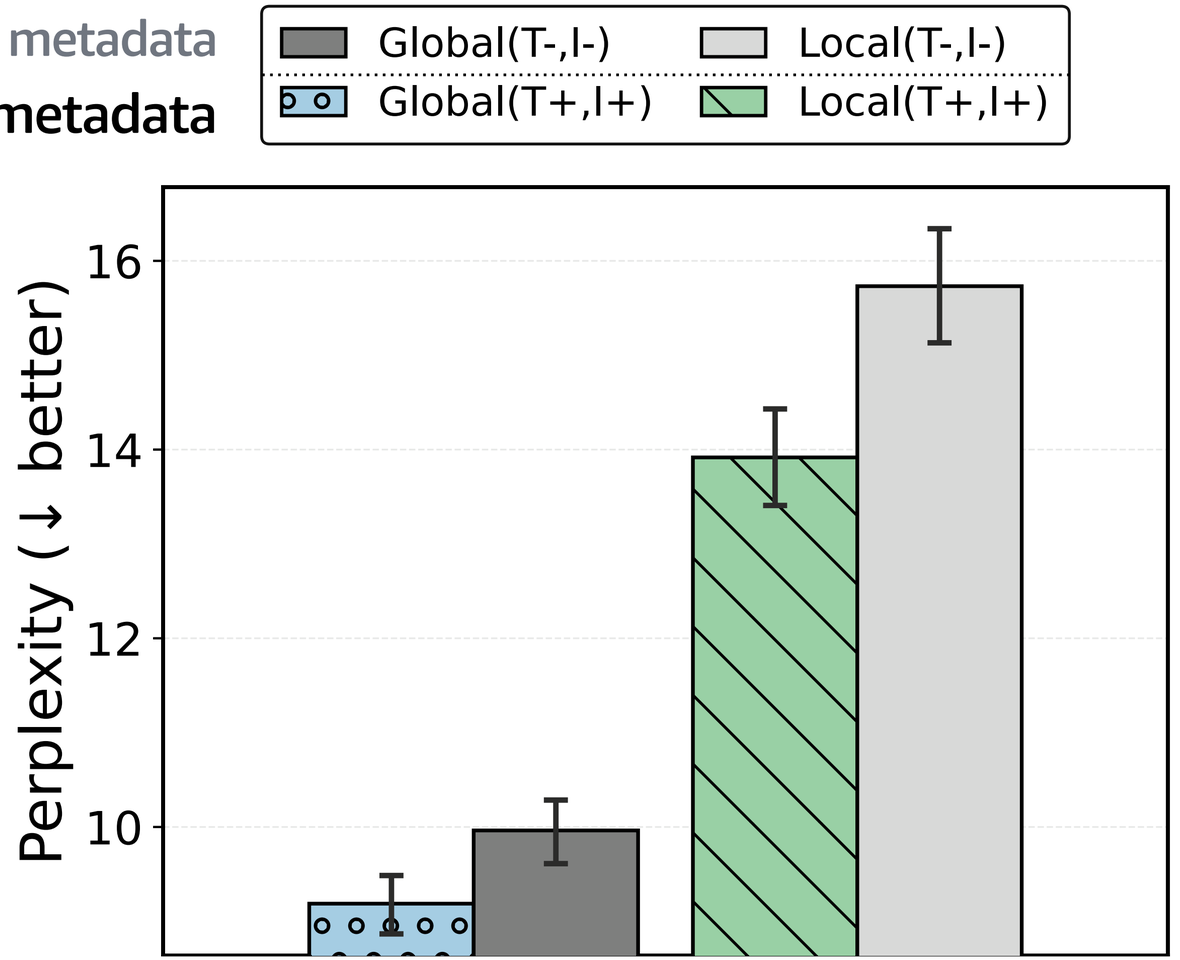
Training/Inference: With metadata



And, we can localize our global models

Training/Inference: Without metadata

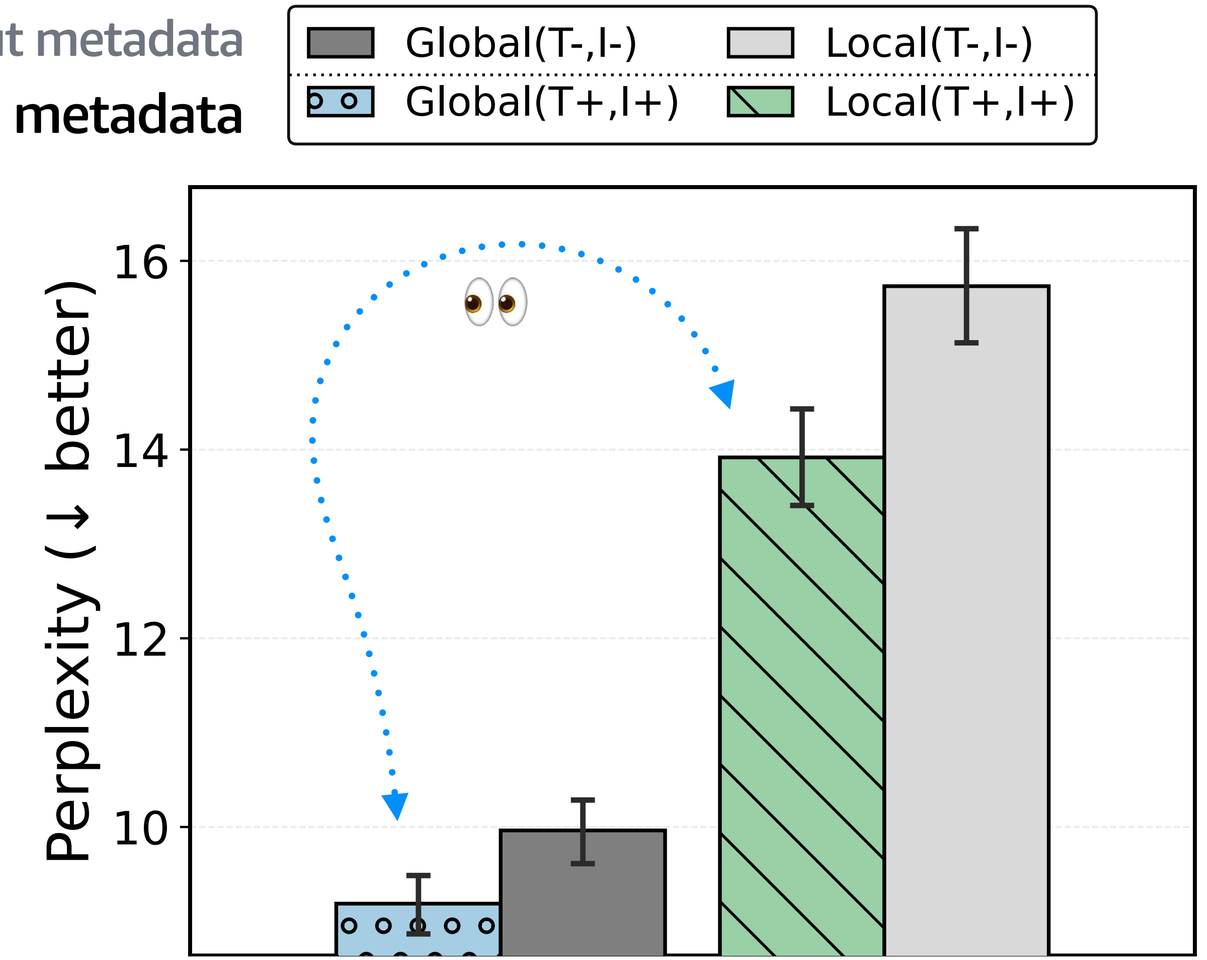
Training/Inference: With metadata



And, we can localize our global models

Training/Inference: Without metadata

Training/Inference: With metadata



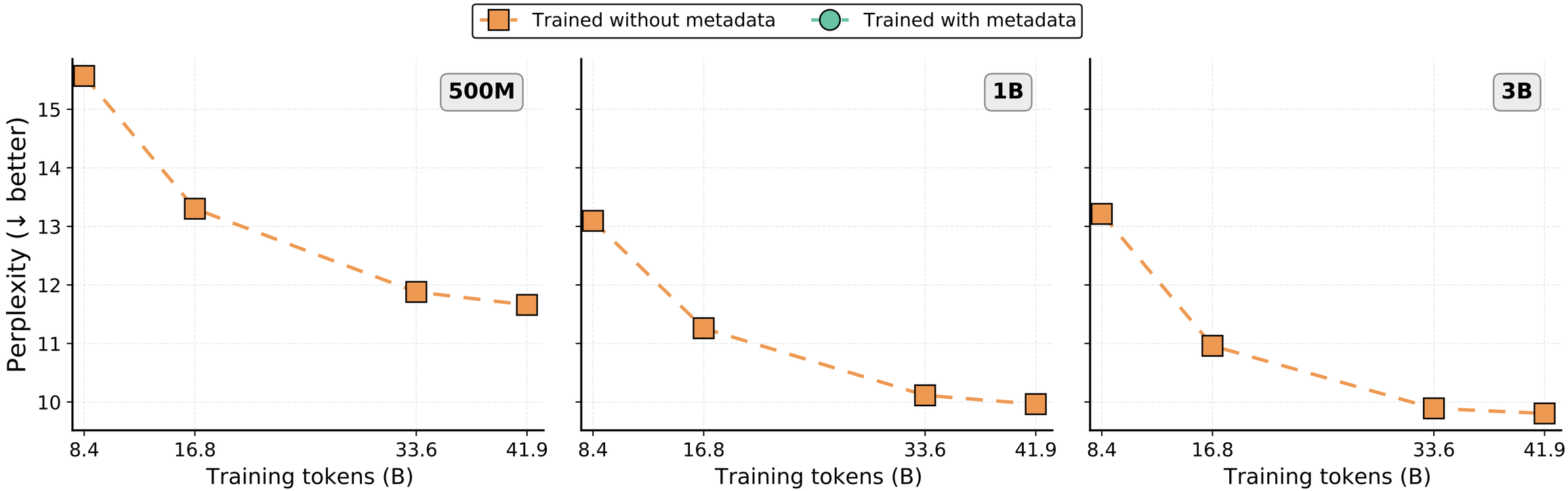
The **global model trained with metadata** has *lower perplexity* on average than **local models trained with metadata** across localized test sets (sentences sampled from within a region only)

Bonus side-effect: Significantly improved **token efficiency**

Bonus side-effect: Significantly improved **token efficiency**

Without metadata | $p(text) = \sum_{\ell} p(\ell)\, p(text \mid \ell)$

mixture across locales



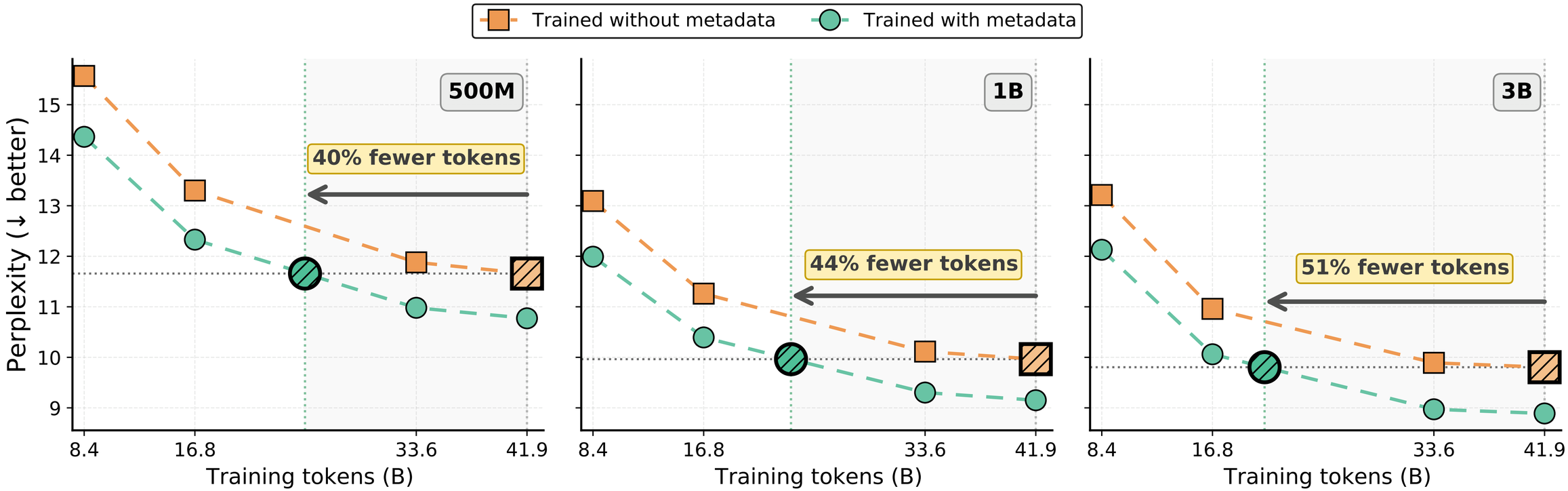
Bonus side-effect: Significantly improved **token efficiency**

Without metadata $p(text) = \sum_{\ell} p(\ell)\, p(text \mid \ell)$

mixture across locales

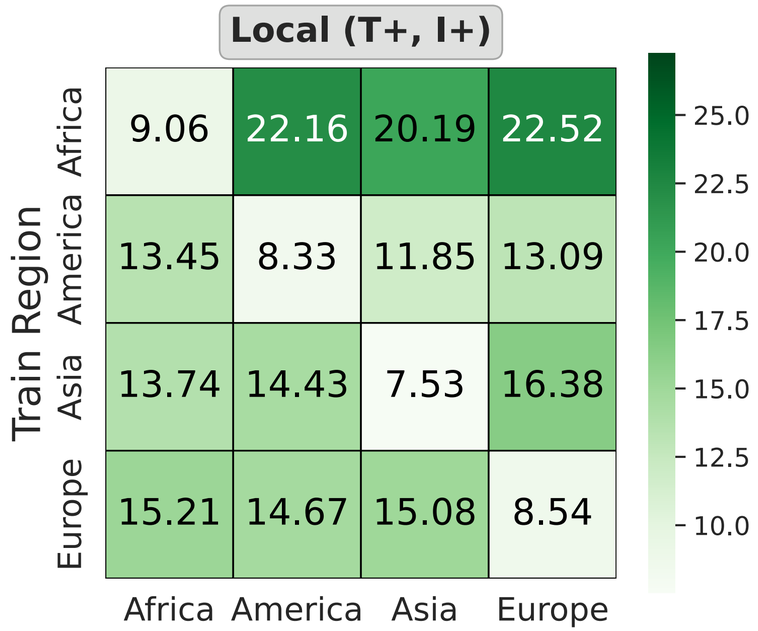
With metadata $p(text \mid m) \approx p(text \mid \ell)$

condition on locale signal



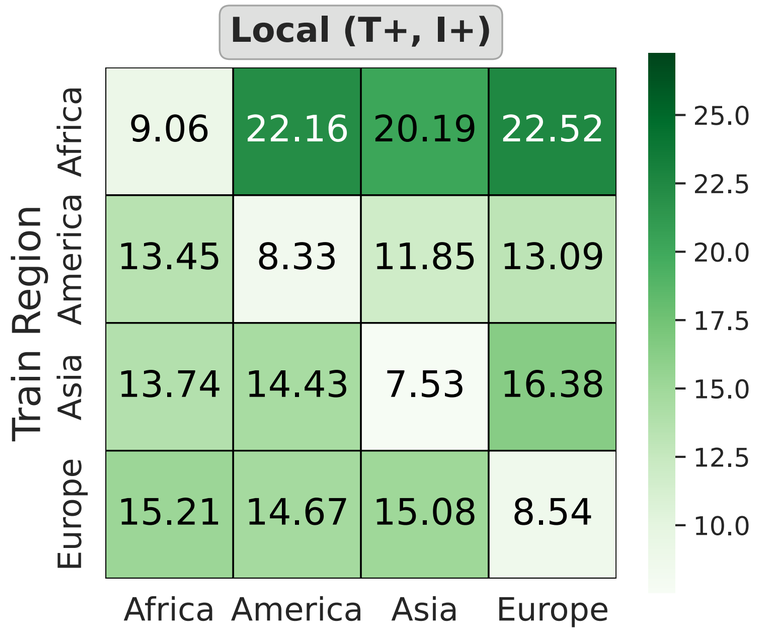
Can training with metadata replace missing training data? **No**

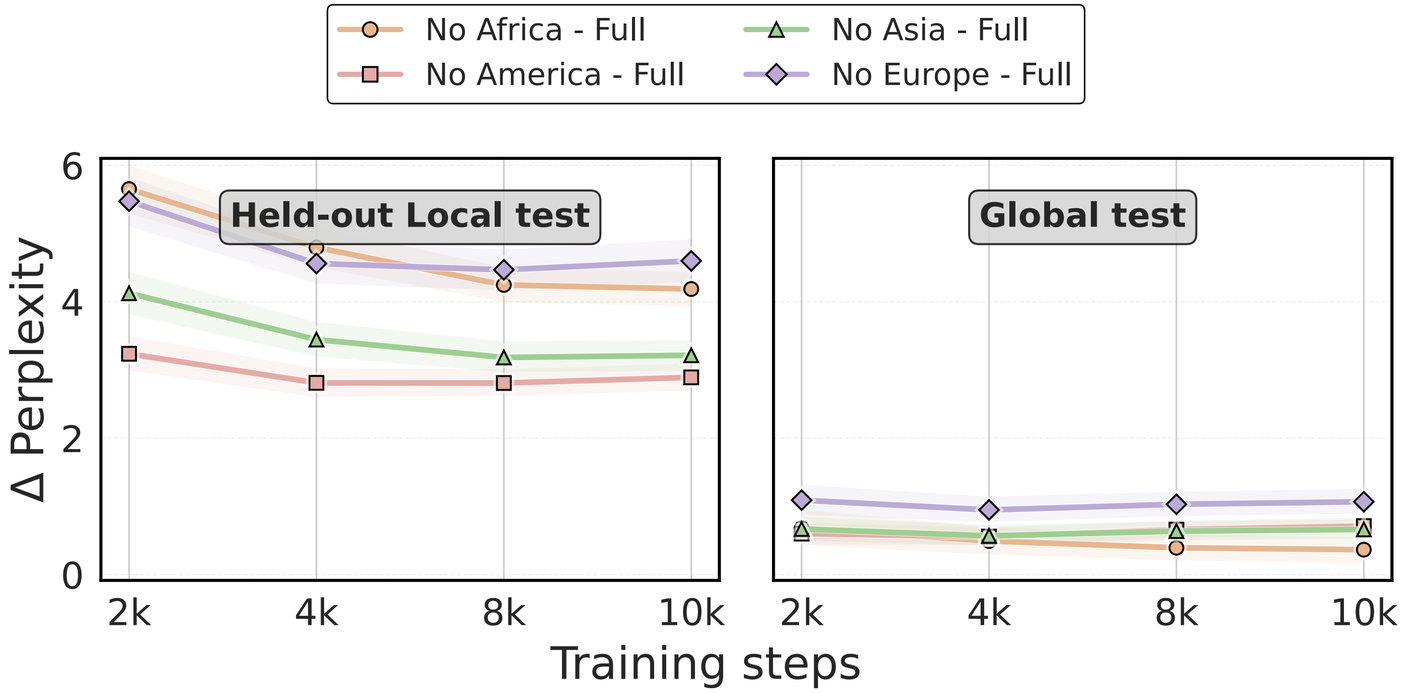
Can training with metadata replace missing training data? **No**



Local models are specialized to their locales

Can training with metadata replace missing training data? **No**

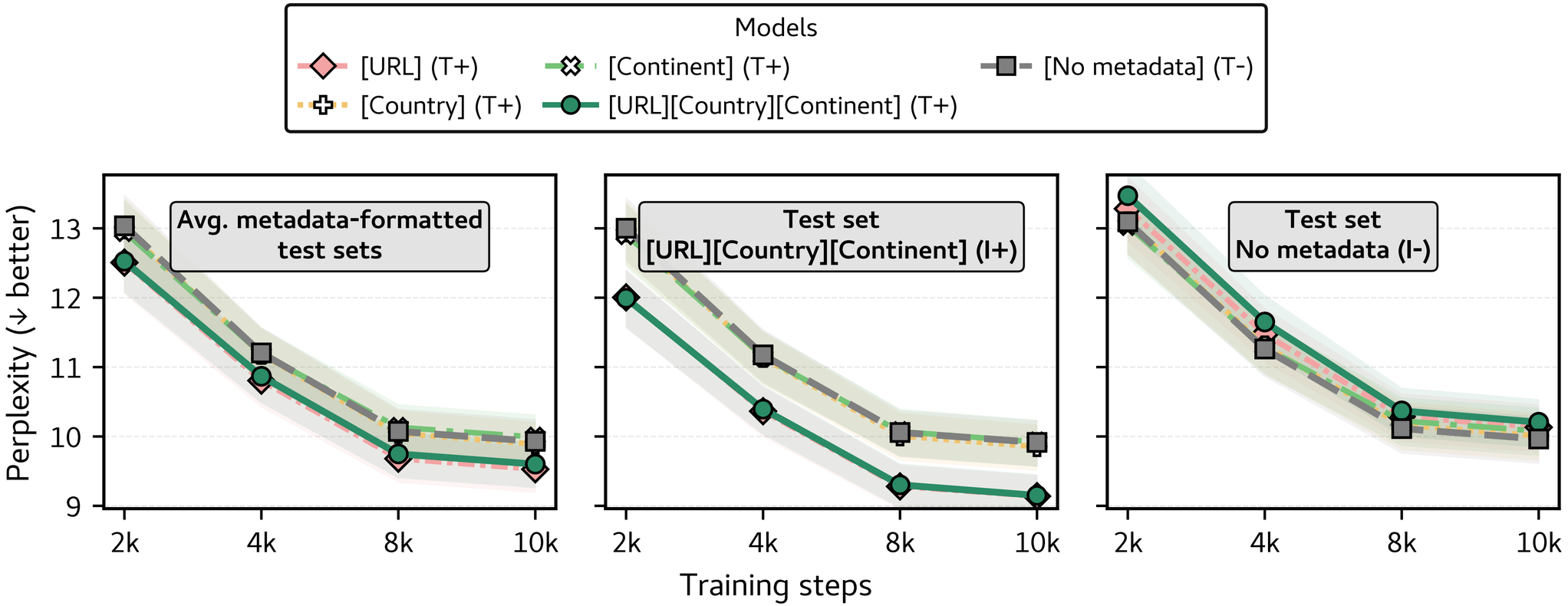




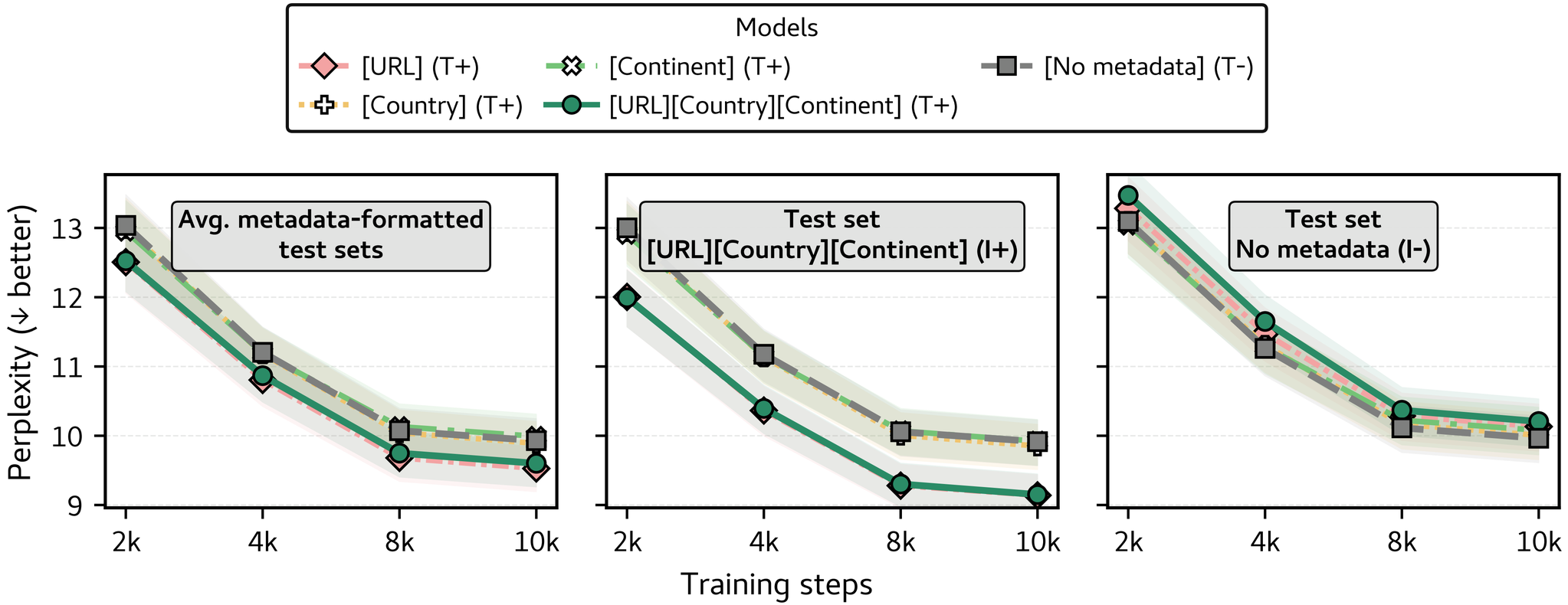
Local models are specialized to their locales

If we remove one region from the training data of global models, they don’t generalize to that region as effectively as they would if that data was used.

Which metadata is most relevant?

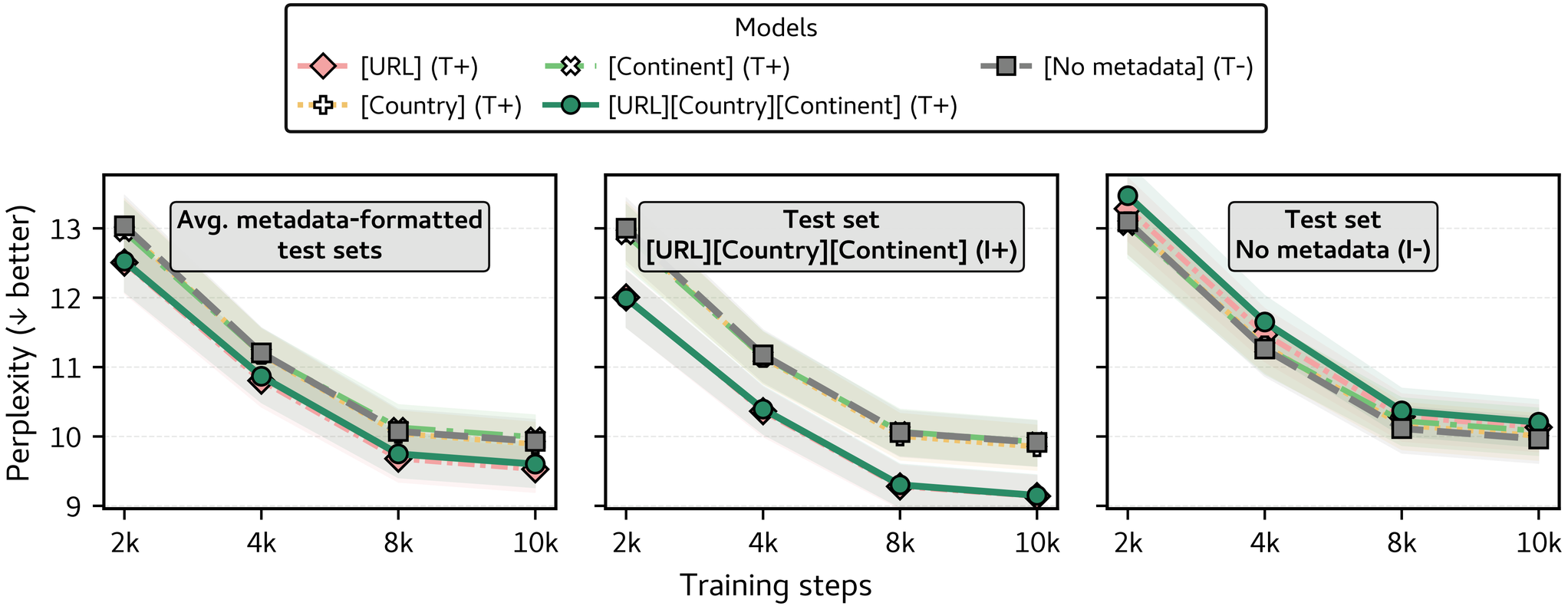


Which metadata is most relevant?



Models trained on URL- only metadata have similar perplexity as models trained with additional country and continent tags.

Which metadata is most relevant?

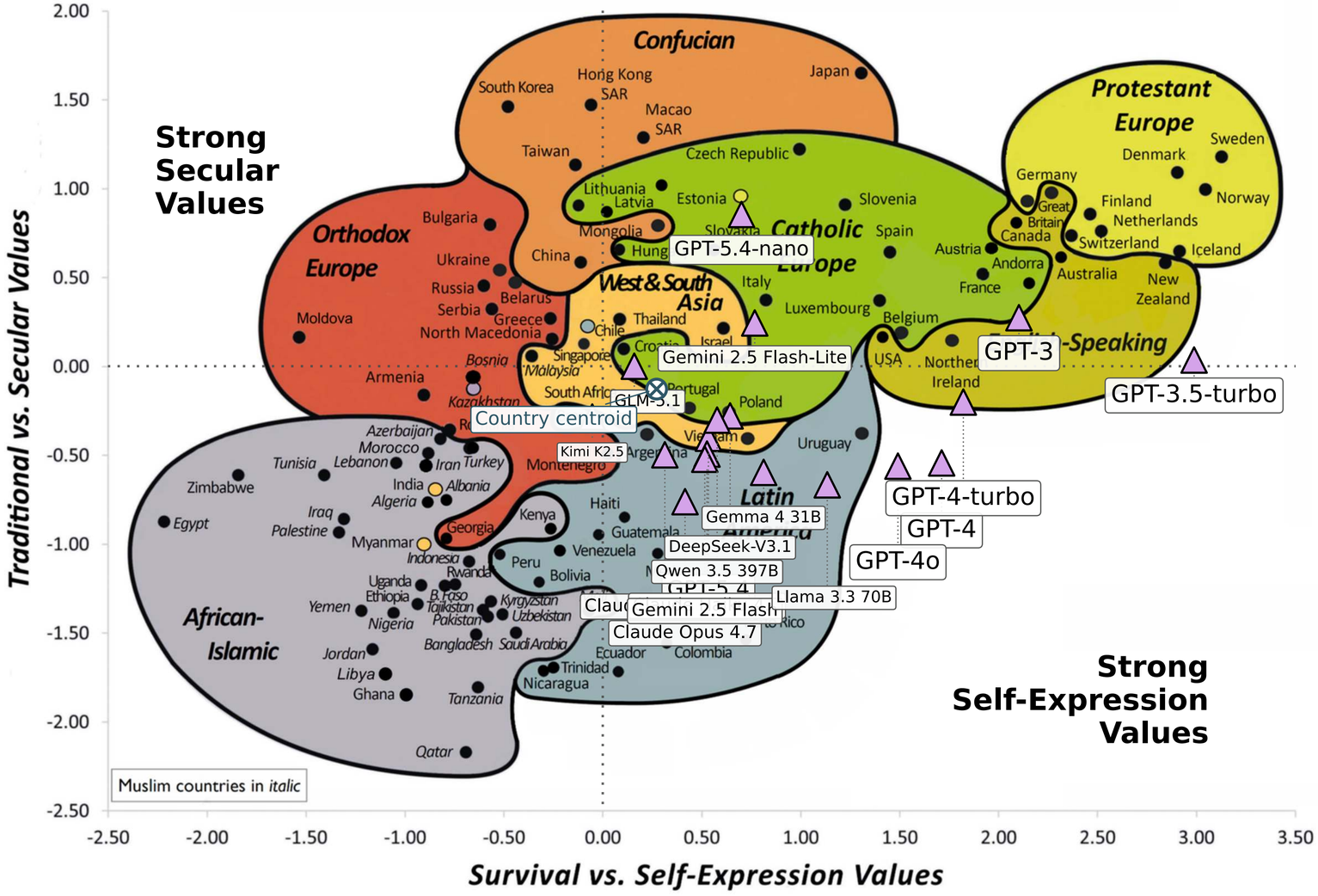


Models trained on URL- only metadata have similar perplexity as models trained with additional country and continent tags.

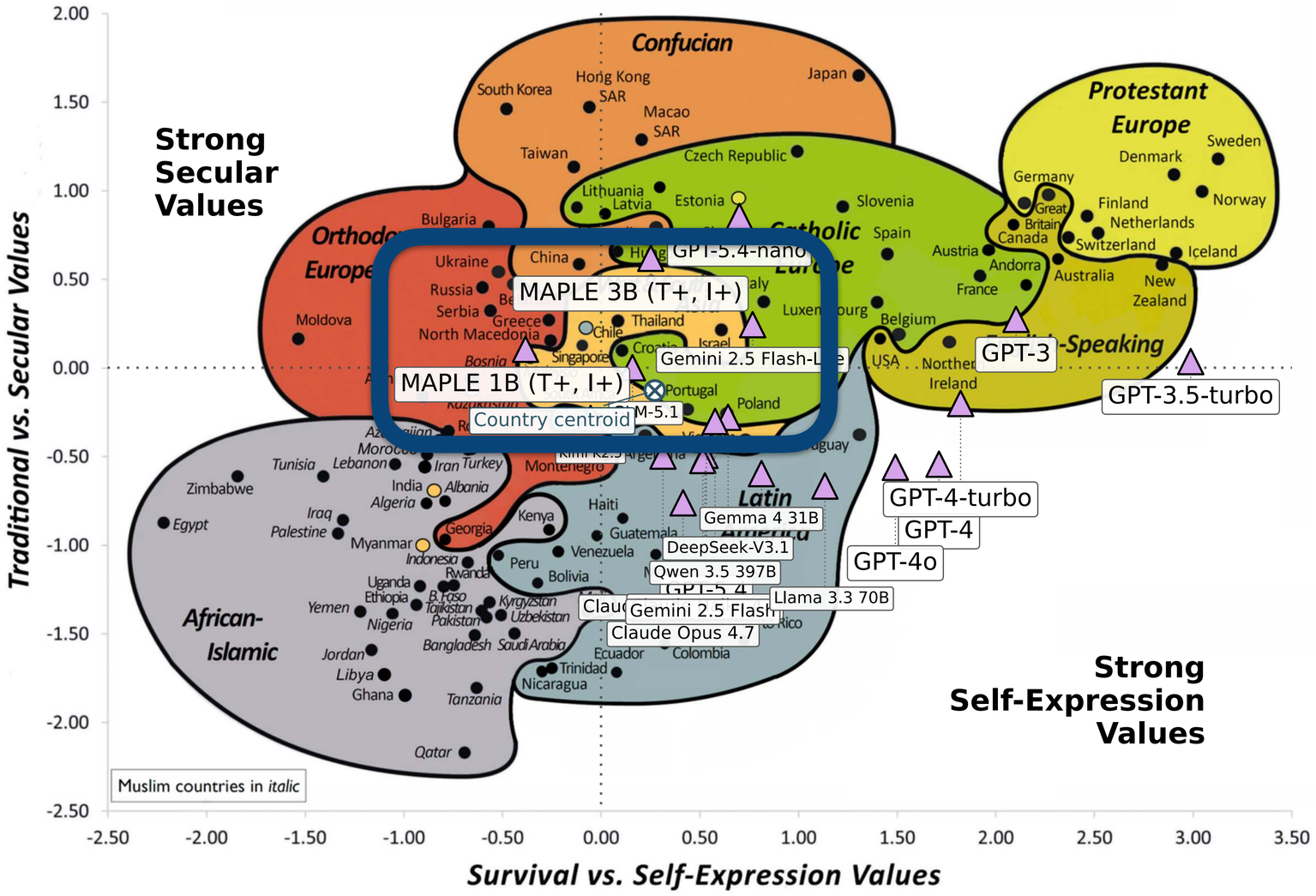
Why? We found a lot of the URLs we use have identifiers that can distinguish countries

Situating the models on the **Cultural Values** Map

Situating the models on the **Cultural Values** Map



Situating the models on the **Cultural Values** Map



MAPLE models are almost as close to country centroids as (much larger) frontier models.

This is good.

They start neutral,

And can be steered towards what's relevant.



https://huggingface.co/collections/iamshnoo/metadata-conditioned-llms



All model checkpoints and data are available on hugging face

https://huggingface.co/collections/iamshnoo/metadata-conditioned-llms

!! Hire Me !!



Anjishnu Mukherjee

Email: *amukher6@gmu.edu*

Website: *iamshnoo.github.io*



Understanding multilingual and multicultural AI

Double Trouble (EMNLP '26)
Tower of Babel (COLM '26)
Global Voices (EMNLP '23)

Adapting AI across languages and cultures

MAPLE (EMNLP '26)
Crossroads (WACV '25)
Global Gallery (NAACL '24)

Evaluating and mitigating social bias in AI generations

KnowBias (ICML '26)
BiasDora (EMNLP '24)
Breaking Bias (AIES '24)

Thank you!