

MAPLE: Metadata Conditioned LLM Pretraining for Locale-Aware Question Answering

Anjishnu Mukherjee Ziwei Zhu Antonios Anastasopoulos

George Mason University

{amukher6, zzhu20, antonis}@gmu.edu

Abstract

Large language models can memorize competing locale-specific facts yet fail to select among them when the locale changes, defaulting instead to a single globally dominant answer. We formalize this as *localized knowledge disambiguation* and introduce LOCALNEWSQA, an 18,700-item English-news benchmark that pairs the same question across two locales and scores whether a model actually switches its answer when the locale changes. We also introduce MAPLE, a controlled family of decoder-only models pretrained with document-level geographic metadata (source URL, country, and continent) already present in the training corpus, and compare it to metadata-free controls trained on identical data with the same token budget, architecture, and optimization. In controlled experiments at 1B and 3B, with inference-time metadata fixed, pretraining with metadata in MAPLE produces measurable switching and improves accuracy on questions whose correct answer depends on locale. Ablations and external-benchmark evaluations further suggest that locale-conditioned prediction benefits from geographic provenance learned during pretraining and that these benefits strengthen at larger model sizes.¹

1 Introduction

Ask a language model “When does the school year start?” and it will answer September, which is correct in the United States and wrong in Australia, India, Japan, and most of the Southern Hemisphere. The model may not be missing knowledge. It has likely memorized competing facts from many countries but has never been taught which fact belongs where.

This pattern is general. Large language models are trained on text from many locales yet optimized

¹Code is available at https://github.com/iamshnoo/metadata_localization

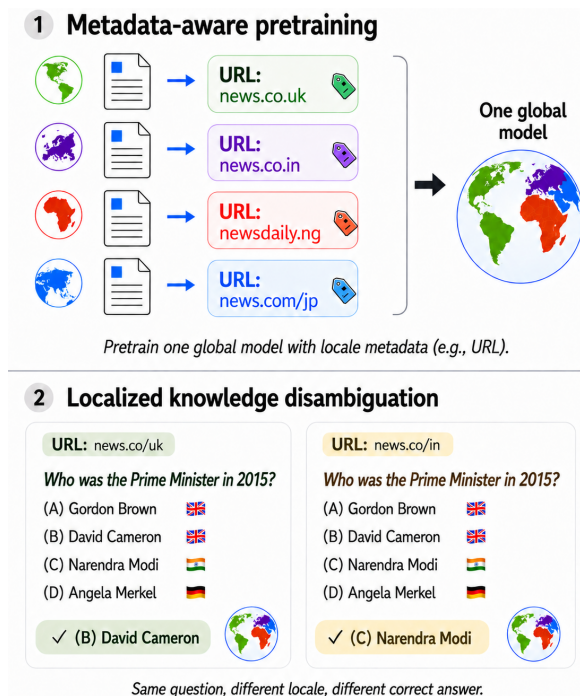


Figure 1. MAPLE turns geographic metadata already present in the training corpus into an inference-time locale signal, allowing one global model to condition factual predictions on source URL, country, and continent.

as a single broad distribution, collapsing locally valid alternatives into geographically homogenized defaults (Weidinger et al., 2022; Ghosh et al., 2024). “What were the highlights of this weekend’s football?” has a different correct answer depending on whether the user is in New York or London, where the word refers to two different sports. “Which agency releases census numbers?” resolves to different institutions by country. “Wrong” answers in such cases are not recall failures. The model likely knows the correct answers but cannot select among them when the locale changes.

We call this problem *localized knowledge disambiguation*, the task of selecting among locale-specific answer distributions rather than producing one unconditioned answer. It draws on under-

specified question answering (Parrish et al., 2022), geographic QA (Yin et al., 2022; Hofmann et al., 2024), cultural evaluation (Adilazuarda et al., 2024; Choenni et al., 2024), and geography-aware adaptation (Cao et al., 2025), but isolates a stricter requirement, namely that the same question frame must resolve differently when the locale changes.

Existing geographic and cultural evaluations score isolated facts, opinions, values, norms, or situated tasks. They are necessary, but they do not directly test whether a model changes its answer under a locale change. Target-locale accuracy alone is insufficient because a model can answer one locale correctly while producing the same answer for a contrast locale. To our knowledge, no existing benchmark scores paired locale switching directly. On the method side, most localization work intervenes through continued pretraining, prompting, adaptation, or evaluation-time context (Gururangan et al., 2020; Adilazuarda et al., 2024; Cao et al., 2025; Hofmann et al., 2024). Whether document provenance available in training corpora can become a learned signal for locale-conditioned prediction, under controlled comparisons, remains untested.

We address both gaps. We introduce LOCALNEWSQA, an English-news benchmark that pairs the same question across two locales with different gold answers, making paired locale switching directly observable. We also introduce MAPLE, a family of decoder-only models trained from scratch with document-level geographic metadata (source URL, country, and continent) from the NOW corpus (Davies, 2016). Because the benchmark questions are generated and validated independently of the training corpus, and the pretraining comparison holds everything except metadata fixed, improvements on LOCALNEWSQA can be attributed to the metadata intervention. Our study asks whether that single intervention changes a specific, directly measurable behavior. Our paper makes four contributions:

- **Task and benchmark.** We define localized knowledge disambiguation and introduce LOCALNEWSQA, an 18,700-item English-news benchmark that pairs the same question across two locales with different gold answers, enabling direct measurement of paired locale switching.
- **Controlled pretraining.** We introduce MAPLE, models equipped to handle document-level geographic metadata (URL, country, con-

tinent), tested while holding data, architecture, and token budget fixed.

- **Efficient localization.** With inference-time metadata held fixed, metadata-conditioned pretraining improves locale switching, with gains increasing from 1B to 3B.
- **Mechanism and transfer.** Ablations, comparisons that cross training-time and inference-time metadata, and external benchmarks suggest that the full effect requires geographic provenance available during both pretraining and inference, and that it transfers to held-out evaluations.

2 Related Work

Localized knowledge disambiguation intersects under-specified question answering (Parrish et al., 2022), locale-sensitive prompting (Cao et al., 2025), culturally grounded evaluation (Adilazuarda et al., 2024; Choenni et al., 2024), and geographic QA (Yin et al., 2022; Hofmann et al., 2024). These works show that models encode uneven geographic and cultural assumptions, but they do not directly test whether a model changes its answer when the locale changes. LOCALNEWSQA targets that stricter question by pairing the same English question frame across two locales and scoring whether the model switches.

XLQA (Roh et al., 2025) and multilingual ODQA benchmarks (Lewis et al., 2020; Artetxe et al., 2020; Clark et al., 2020; Longpre et al., 2021) share the premise that answers need not be locale-invariant, but they vary the language alongside the locale, making it difficult to attribute answer changes to geographic knowledge rather than translation quality or multilingual transfer. We hold language fixed (English throughout), train metadata-conditioned and metadata-free models from scratch, and score whether the model switches its answer when only the locale changes. This design isolates locale-conditioned answer selection from translation quality, multilingual transfer, and evaluation-time prompting.

Metadata conditioning has steered models with control codes (Keskar et al., 2019), temporal signals (Dhingra et al., 2022), language tags (Liu et al., 2020), document structure (Aghajanyan et al., 2022), human-preference labels (Korbak et al., 2023), source attribution (Khalifa et al., 2024; Weller et al., 2024), and social context (Kulkarni et al., 2021; Li et al., 2022). Most relevant to our work, Gao et al. (2025) show that real URL meta-

data accelerates decoder-only pretraining, but they evaluate the effect through perplexity rather than testing whether the model can use provenance to disambiguate competing locale-specific answers at inference time. MAPLE addresses that open question, measuring whether geographic metadata becomes a conditioning signal for locale-sensitive answer selection through paired locale switching rather than perplexity alone.

3 Task and Benchmark

Existing evaluations cannot measure whether a model changes its answer when the locale changes, because they test one locale at a time. We design a task and benchmark that makes paired locale switching directly observable.

3.1 Task Definition

Localized knowledge disambiguation evaluates whether a model can condition a factual answer on the locale where the fact applies. The key difficulty is that the question need not name the locale. Prompts such as “Which agency releases census numbers?”, “Which grades are primary school?”, or “Which football league is the top flight?” are ambiguous unless the locale is supplied, because the relevant statistical office, grade span, or sports code changes by country. The diagnostic examples are therefore not merely country-specific questions, but questions whose answer changes when the locale changes.

LOCALNEWSQA is designed around that requirement. The benchmark contains two evaluation splits. The *explicit* split names the locale in the question and measures localized factual recall; these questions have a single correct answer once the locale is specified. The *ambiguous* split omits the locale from the question text and supplies the locale through the evaluation context; these are locale-sensitive because the correct answer changes with the locale. Each ambiguous item also has a paired contrast locale with a different gold answer. This pairing makes LOCALNEWSQA a direct evaluation of the behavior that accuracy-only benchmarks miss, namely whether the same model changes its answer when the locale changes.

3.2 LOCALNEWSQA Construction and Validation

We construct LOCALNEWSQA in English news because news provides dense local institutions, pub-

lic bodies, sports, education systems, media references, and recent events with traceable source URLs. The benchmark has to satisfy broad country/topic coverage while keeping only rows whose answer can be verified against source evidence. We therefore use generation only to find candidate question frames, and filter to keep only rows that pass split-specific source and formatting checks. The final release contains 18,700 questions from 17 countries across 4 continents, with 17,000 explicit items balanced at 1,000 per country and 1,700 ambiguous items balanced at 100 per country. The benchmark covers 8 topic families, 🏛️ politics/government, 🏢 civic institutions, 📈 economy/business, 🎓 education, 🎭 media/culture, ⚽ sports, 🎉 public life/holidays, and 🗺️ geography/infrastructure. Table 1 summarizes the gold-selection criteria.

Every question in the final benchmark includes source evidence for auditability. In a three-annotator audit of a stratified 100-item gold sample, majority vote accepted 96/100 items, including 50/50 explicit and 46/50 ambiguous, with 98.0% average pairwise agreement and Fleiss’ $\kappa = 0.76$ on the binary accept/reject decision. Appendix A.4 gives the full construction, validation, and human-audit protocol.

3.3 LOCALNEWSQA Evaluation

We report target-locale accuracy and paired locale switching. For an ambiguous pair with frame q , target locale ℓ_t , contrast locale ℓ_c , and answers a_t, a_c , target accuracy asks only whether the model chooses a_t under ℓ_t . The paired metrics are defined as follows.

$$\begin{aligned}
 \text{exact pair} \quad & \underbrace{\hat{a}(q, \ell_t) = a_t}_{\text{correct on target}} \wedge \underbrace{\hat{a}(q, \ell_c) = a_c}_{\text{correct on contrast}}, \\
 \text{margin switch} \quad & \underbrace{s(a_t | q, \ell_t) > s(a_c | q, \ell_t)}_{\text{target ans. preferred under } \ell_t} \\
 & \wedge \underbrace{s(a_c | q, \ell_c) > s(a_t | q, \ell_c)}_{\text{contrast ans. preferred under } \ell_c}.
 \end{aligned}$$

Exact pair is stringent because both sides must be correct; margin switch is softer but still requires the score ordering to move toward the locale-appropriate answer on both sides.

For MAPLE, we score pretrained checkpoints using length-normalized log-probability over each option’s token sequence. Table 3 averages the

Criterion	Explicit split	Ambiguous split
Balance	1,000 items per country; 8 topic families represented	100 items per country; 8 topic families represented
Format	4 options, 1 gold, no duplicate question	4 options, 1 gold, no duplicate question
Evidence	Target answer supported by source	Both target and contrast answers supported
Locale dep.	Question names the locale	Question omits locale; target and contrast answers differ

Table 1. Gold-selection criteria for the final LOCAL-NEWSQA benchmark. Questions that fail any criterion are excluded.

crossed paired-switch comparison over five answer-order seeds; target-accuracy and other analysis fix a shared answer-order seed and use question-level bootstrap intervals. Metadata-formatted inference supplies the locale fields corresponding to the question being evaluated. Target prompts include the target country and continent, and contrast prompts include the contrast country and continent. Figure 9 gives the exact prompts, and Figure 6 reports results on the chat variants obtained from LoRA instruction tuning.

4 MAPLE Controlled Pretraining

We use the NOW corpus (Davies, 2016) throughout. Each document includes a publication year, source URL, and country of origin, which we map to continents and serialize as structured metadata before the article title and content. After filtering, the corpus contains approximately 4M documents from Africa, 10M from America, 7.5M from Asia, and 5.6M from Europe. All local and global models are trained on 41.9B tokens. Local models sample one continent to that budget, while global models mix all four continents before sampling. Appendix A.1 gives the full data formatting and geographic coverage details.

Metadata-formatted training instance

```
URL news.example/jp/article/12345
COUNTRY Japan
CONTINENT Asia
TITLE Election results spark debate ...
CONTENT The recent election results have led to widespread discussions across political parties and the public, with analysts noting ...
```

The pretraining intervention changes only the document serialization. Metadata-formatted documents prepend source URL, country, and continent

# Continents	Train Meta	Eval Meta	Notation
1	✓ T+	✓ I+	LOCAL (T+, I+)
1	✓ T+	✗ I-	LOCAL (T+, I-)
1	✗ T-	✓ I+	LOCAL (T-, I+)
1	✗ T-	✗ I-	LOCAL (T-, I-)
4	✓ T+	✓ I+	GLOBAL (T+, I+)
4	✗ T-	✗ I-	GLOBAL (T-, I-)

Table 2. Model variants used in this paper. “T+/T-” indicate metadata present/absent during training; similarly “I+/I-” indicate metadata available versus not available at inference. (Bordered boxes) also indicate I+ models.

fields before the title and article text, as shown above. Metadata-free documents omit these fields. The comparison is controlled so that changes in behavior can be attributed to the presence of provenance metadata rather than to data volume, architecture, optimizer, or tokenizer changes.

We train 33 models from scratch at two sizes. At 500M, we train 10 models covering 4 continents $\times$ 2 metadata settings plus 2 global settings. At 1B, we train 23 models with the same 10 settings plus 5 provenance-granularity ablations and 8 leave-one-continent controls. We additionally train two 3B global models to test scaling.

Throughout the results, T+/T- indicate metadata present or absent during training, and I+/I- indicate metadata present or absent at inference. The notation separates whether a model has learned from metadata during pretraining from whether metadata is supplied at evaluation time.

5 Results

Each research question builds on the previous one, moving from establishing the basic effect to explaining which signals drive it and whether it generalizes. RQ1 (§5.1) establishes paired locale switching. RQ2 (§5.2) tests whether metadata also improves region-specific language modeling. RQ3 (§5.3) isolates which provenance fields carry the signal. RQ4 (§5.4) checks whether gains transfer to external benchmarks.

5.1 RQ1. Does MAPLE Improve Localized Knowledge Disambiguation?

The first experiment asks whether provenance metadata improves localized disambiguation under a controlled pretraining intervention. If the model learns metadata as a usable conditioning signal, it should help most when the question underspecifies locale, producing larger accuracy gains on ambiguous items than on explicit ones. Paired tar-

get/contrast cases provide the strongest test because the answer must change with the locale, and switching gains should increase at larger model sizes if the metadata signal is being used more effectively.

Results The primary comparison controls inference-time information. Both (T+, I+) and (T-, I+) receive different locale prompts, but only the former learned from metadata during pretraining. Table 3 shows that pretraining raises the paired margin-switch rate from 6.04% to 7.78% at 1B (+1.74 points) and from 21.36% to 24.13% at 3B (+2.77 points). These increments isolate the effect attributable to metadata-conditioned pretraining.

Takeaway 1. With inference-time metadata fixed, metadata-conditioned pretraining improves paired-switching, with increasing gains from 1B to 3B.

The (T-, I-) zero paired-switch result is expected because target and contrast prompts are identical; we include it only as a negative control showing that the evaluation does not leak locale information.

In our additional target-locale accuracy analysis, we compare (T+, I+) against (T-, I-). On the ambiguous split, this comparison yields gains of +6.7 points at 1B and +9.2 points at 3B. The explicit split also improves, but less sharply, because explicit questions already name the locale. These gains are distributed across countries and topics, and their bootstrap intervals exclude zero at both model sizes (Appendix Table 16).

5.2 RQ2. Does Metadata Improve Region-Specific Language Modeling?

If metadata helps the model condition on locale, the benefit should extend beyond the benchmark format and also lower held-out perplexity on text from the same geographic region. We test this at two levels. First, continent-specific local models, where crossed train/evaluation settings separate learned provenance from metadata supplied only at inference. Second, global models trained on all continents, where the question is whether one model can retain multiple local distributions without region-specific parameters.

The four crossed settings are abbreviated as **LOCAL (T+, I+)** (metadata during both training and inference), **LOCAL (T+, I-)** (train-only), **LOCAL (T-, I+)** (inference-only), and **LOCAL (T-, I-)** (no metadata at either stage).

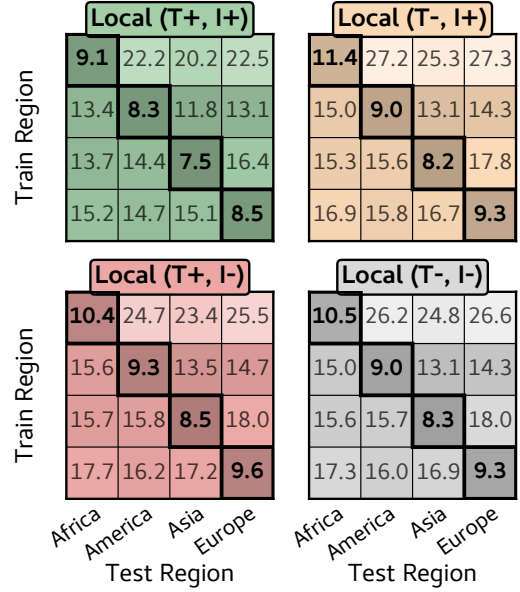


Figure 2. **Cross-continent perplexity for local models.** Diagonal cells are consistently lower than off-diagonal cells, showing locality. Full metadata gives the strongest local performance and improves each train/evaluation setting compared with weaker metadata variants.

Scale	(T-, I-)	(T+, I-)	(T-, I+)	(T+, I+)	Δ_{train}
1B	0.00	0.00	6.04	7.78	+1.74
3B	0.00	0.00	21.36	24.13	+2.77

Table 3. Paired margin-switch rate (%) on the 1,700 ambiguous LOCALNEWSQA pairs, averaged over five answer-order seeds. The primary pretraining contrast, (T+, I+) vs. (T-, I+), holds inference-time locale information fixed; Δ_{train} is their difference in percentage points. Both I- settings use identical target/contrast prompts and therefore yield zero switches by construction; (T-, I-) is the negative control.

Setup For each continent and scale (0.5B and 1B), we train checkpoints with and without metadata and evaluate under all four crossed settings on local and cross-continent test sets (Figure 2). Each continent has a held-out test set of 1,000 documents, and all models are trained on the same total number of tokens. Test perplexity is computed over non-metadata tokens only, so any improvement reflects better content prediction rather than metadata memorization. We additionally train global models on data from all continents at 0.5B, 1B, and 3B, asking whether one model can serve multiple regions without region-specific specialization. The global test set contains 1,000 documents with 250 sampled uniformly from each continent, testing whether a single set of parameters can serve all regions without sacrificing performance on any individual one.

Hypotheses If metadata provides a useful conditioning signal, models trained with it should achieve lower perplexity on their continent’s test set than metadata-free controls, with the crossed settings (metadata at only one stage) falling between the two extremes. Local models should remain local, with lower same-continent than cross-continent perplexities. At the global level, a metadata-conditioned model should approach local-model quality on each region while staying stronger on average on the combined global test set.

Results Figure 2 shows that models trained and evaluated with full metadata achieve the lowest perplexity on every same-continent test set, showing that metadata indeed provide a useful conditioning signal. The crossed settings fall in between; a supporting diagnostic in Appendix Figure 22 suggests that inference-time metadata contributes more than train-time metadata alone. Across all configurations, same-continent perplexity is consistently lower than cross-continent perplexity, showing that local models remain local.

Figure 3 summarizes the global comparison at 1B. A single metadata-conditioned global model lowers perplexity relative to the metadata-free control on every continent-specific test set as well as on the combined global test set. From 500M to 1B, it approaches local-model quality while remaining stronger on the combined global test set. The remaining gap is expected because region-specific models specialize to one distribution, whereas the global model preserves all four under one parameter budget. The useful result is that metadata moves one shared model toward local behavior without sacrificing breadth. This pattern continues at 3B, where the metadata-conditioned global model again outperforms the metadata-free control. The training-token curves in Figure 5a show that metadata-conditioned models reach any given perplexity target using fewer training tokens, an efficiency benefit that persists across model sizes.

Takeaway 2. A single global model can retain multiple regional distributions when trained with provenance metadata. The benefit persists from 500M to 3B and appears as both lower perplexity and faster convergence.

We additionally observe that localized performance is distributed across countries rather than carried by a single dominant country per conti-

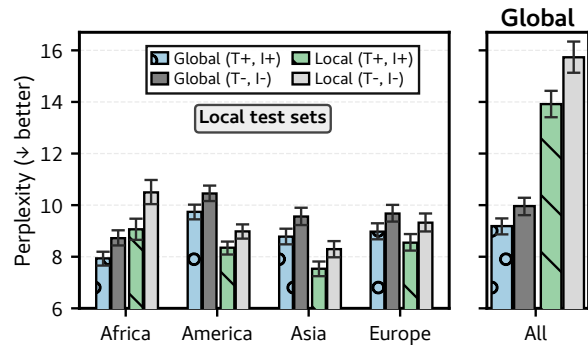


Figure 3. **Global metadata conditioning improves held-out perplexity at 1B.** A single global model with metadata lowers perplexity on every continent-specific test set and the combined global test set.

nent, and that Africa-trained local models transfer less symmetrically to other regions (Appendix Figures 16 and 17).

5.3 RQ3. Which Provenance Signals Matter, and What Can They Not Replace?

This experiment isolates which provenance fields carry the signal and whether metadata can compensate for missing regional coverage.

Setup We evaluate five 1B metadata-granularity variants, `URL`, `URL` `Country`, `URL` `Continent`, `Country`, and `Continent`, alongside full/no metadata and leave-one-continent global controls. Each variant uses the same training data and token budget, differing only in which metadata fields are serialized before the document text. Figure 4 reports the main variants; Appendix Figures 23 and 20 show all checkpoints and the leave-one-continent results.

Leave-one-continent controls separate metadata from data coverage. Each model is trained on three of the four continents and evaluated on both the held-out continent and the combined global test set. This design tests a specific alternative explanation. If metadata merely encoded continent labels, adding it should compensate for the missing region even when that region was excluded from training. If, on the other hand, metadata helps the model organize evidence it has actually seen, held-out-region perplexity should still degrade when the region was absent from training, regardless of what metadata is supplied at evaluation time. The distinction matters for practitioners deciding whether metadata can substitute for broader data collection.

Hypotheses If provenance is the key signal, the URL field alone should recover most of the benefit because URLs implicitly encode geography through country-code top-level domains and

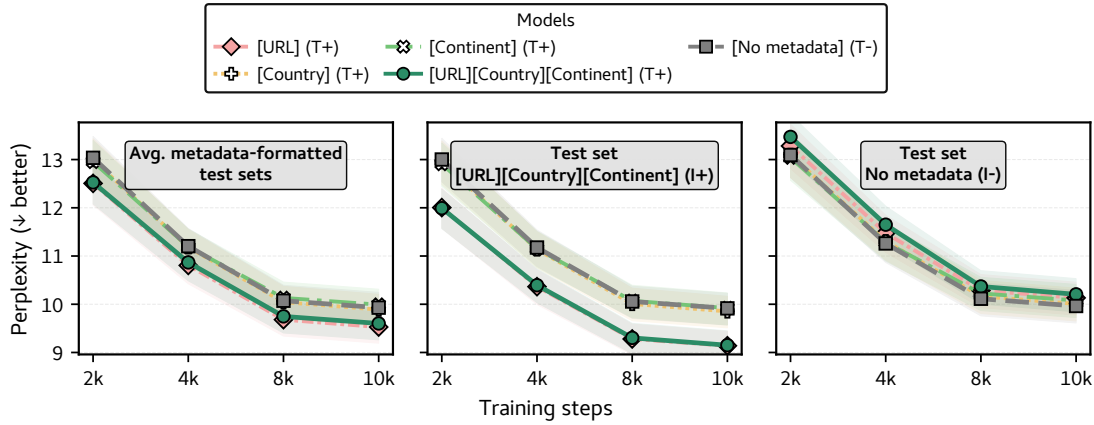


Figure 4. **Metadata ablations across checkpoints.** ●URL-only metadata recovers most of the full-metadata perplexity benefit; ●Country/●Continent tags alone trail ●URL, and adding them to ●URL changes little.

region-specific domain names (Appendix Table 12 quantifies this). Coarser labels such as country or continent names should be less informative in isolation, and metadata of any granularity should not compensate for data the model has never seen during training.

Results The results support all three expectations. Figure 4 shows that ●URL is the strongest individual provenance signal, nearly matching full metadata and substantially outperforming the no-metadata baseline. The URL field alone recovers most of the localization benefit. ●Country and ●Continent trail ●URL and add little when paired with it, suggesting that coarse geographic labels are largely redundant once the URL is present. A supporting leave-one-continent analysis (Appendix Figure 20) shows that excluding any continent raises held-out and global perplexity regardless of metadata, so provenance organizes evidence the model has seen but cannot substitute for regional coverage it lacks. This finding has a practical implication: metadata conditioning and broad geographic data collection serve different purposes and neither replaces the other.

Takeaway 3. The source URL alone carries most of the geographic signal. Adding country or continent labels on top of it changes little, and no amount of metadata can teach the model about regions it has never seen in training.

5.4 RQ4. Does Localization Transfer Beyond LOCALNEWSQA?

Finally, we test whether localized disambiguation gains transfer beyond LOCALNEWSQA. We evaluate six benchmarks covering geographic factual recall (GEOMLAMA; Yin et al., 2022), opinions and

values (GLOBALOPINIONQA; Durmus et al., 2024, WORLDVALUESBENCH; Choenni et al., 2024), norm judgments (NORMAD; Rao et al., 2025), and culturally situated tasks (BLEND; Myung et al., 2024, GLOBAL-MMLU-CS; Singh et al., 2025). Figure 5b plots each model’s LOCALNEWSQA ambiguous gain against its normalized external score (the mean of per-task accuracy scores normalized by the best model on each task). The question is whether training with metadata to improve localized disambiguation hurts or helps performance on broader geographic and cultural evaluations.

Hypotheses A useful localization method should improve controlled disambiguation without sacrificing broad performance on external evaluations. Inference-time metadata alone, without a matching pretraining signal, is not expected to produce the same benefit as reliably.

Results Figure 5b shows that the controlled MAPLE points remain competitive with size-matched models on normalized external score while simultaneously improving LOCALNEWSQA. At 3B, MAPLE shows statistically significant gains on GLOBALOPINIONQA (+8.4), NORMAD (+5.1), and BLEND (+13.5), with positive point estimates on the remaining three tasks (full results in Appendix Table 17). Baseline models with metadata appended only at evaluation time show mixed gains, consistent with the interpretation that locale metadata is more reliable when learned during pre-training than when supplied only at inference.

Takeaway 4. On six external benchmarks, MAPLE remains competitive with size-matched models while improving locale-conditioned behavior.

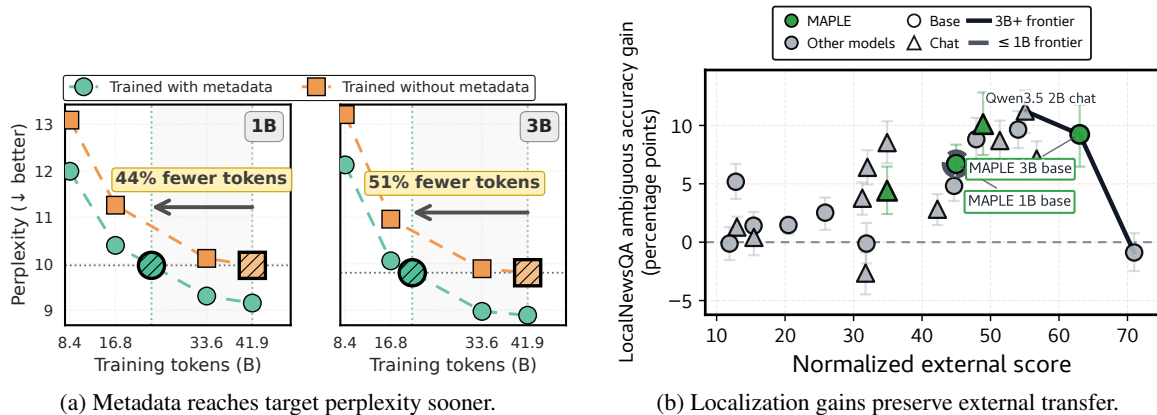


Figure 5. **Scale and transfer.** Metadata lowers global perplexity more quickly from 1B to 3B (lower is better in panel A) while preserving broad external performance. In panel B, each point is a model; the frontier is the set of models not dominated by any other model of similar size on both axes. MAPLE points remain on this frontier. Task-level external results are in Appendix Table 17.

6 Discussion

Taken together, the paired-switching, perplexity, and transfer results suggest that, at least for news-domain text at the scales tested, preserving provenance during pretraining adds a modest but consistent benefit beyond inference-time metadata alone. In practice, this means geographic metadata already present in the training corpus should be preserved rather than stripped during preprocessing. The URL field likely carries most of the signal in part because it encodes country-code TLDs, region-specific domains, and outlet identity (Appendix Table 12), all of which are lost when corpora are reduced to plain text.

Beyond geography The paired-switching protocol is not inherently geographic. Any axis along which valid completions compete could be tested with the same factorial design. Temporal provenance (publication date) would let a model distinguish “Who is the president?” across time periods, and source-type provenance (books vs. web data) could help the model weight formal versus colloquial usage depending on context. The shared requirement is that competing valid answers exist and that an available metadata field can disambiguate among them.

7 Conclusion

Geographic metadata already present in training corpora can help a model bind facts to locales. We show that preserving source URLs, country, and continent fields from the NOW news corpus during pretraining gives a single global model the ability to condition its answers on locale. Holding inference-time locale information fixed,

metadata-conditioned pretrained models are better at switching answers when the locale changes, with progressively increasing strength from 1B to 3B. The source URL alone carries most of the geographic signal without requiring explicit country or continent tags. These gains transfer to three of six external benchmarks with positive point estimates on the remaining three, suggesting that provenance-conditioned pretraining is a promising direction for locale-aware factual prediction.

Limitations

This study is intentionally designed as a controlled isolation of metadata conditioning rather than as an exhaustive survey of every setting where provenance could be useful. We fix the model family, tokenizer, optimizer, token budget, and source domain so that train-time provenance can be compared cleanly against metadata-free controls. The natural next step is broader transfer through the same factorial design across additional model families, multilingual corpora, open-web sources such as Common Crawl, and non-news domains where provenance may encode different kinds of context.

Several extensions would sharpen this picture further. Our from-scratch pretraining experiments cover 500M to 3B parameters; we did not test 70B or larger models because that scale was infeasible under our academic compute budget. Whether the paired-switching gains persist beyond 3B therefore remains untested, and we encourage future large-scale compute collaborations to evaluate this regime. Repeated pretraining runs would also separate evaluation uncertainty from pretraining stochasticity. LOCALNEWSQA is generated to tar-

get a specific behavior, paired locale switching, and is released with split-specific source validation plus a stratified human audit; future benchmark work can combine this controlled template with naturally occurring user questions and richer human annotation. These directions would test generality while preserving the main purpose of the present study, which is to isolate whether geographic metadata already present in training corpora can help a single global model bind facts to the locales where they are valid.

Ethical Considerations

Metadata conditioning could reinforce coarse geographic stereotypes if applied carelessly. Our study uses factual news content with verified provenance and avoids personal data. Geographic metadata should be treated as contextual signals rather than as proxies for individual identity or values. Downstream applications should consider fairness and inclusivity, particularly in high-stakes settings where locale-conditioned outputs could encode or amplify regional biases.

Acknowledgments

We are thankful to the reviewers who provided feedback in earlier versions of this work. This work was generously supported by the US National Science Foundation CAREER award 2439202. This work is also in part supported by NSF grant IIS-2452129. We acknowledge the support by resources provided by the Office of Research Computing at George Mason University (<https://orc.gmu.edu>) and funded in part by grants from the NSF (Award Number 2018631). Any opinions, findings, and conclusions or recommendations expressed in this material are solely those of the authors.

References

- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. [Towards measuring and modeling “culture” in LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.
- Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2022. [HTLM: hyper-text pre-training and prompting of language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. [On the cross-lingual transferability of monolingual representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. [Specializing large language models to simulate survey response distributions for global populations](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3141–3154, Albuquerque, New Mexico. Association for Computational Linguistics.
- Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. [The echoes of multilinguality: Tracing cultural value shifts during language model finetuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15042–15058, Bangkok, Thailand. Association for Computational Linguistics.
- Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. [TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages](#). *Transactions of the Association for Computational Linguistics*, 8:454–470.
- Mark Davies. 2016. Corpus of news on the web (now). <https://www.english-corpora.org/now/>.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. [Time-aware language models as temporal knowledge bases](#). *Transactions of the Association for Computational Linguistics*, 10:257–273.
- Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. [Towards measuring the representation of subjective global opinions in language models](#). In *First Conference on Language Modeling*.

- Tianyu Gao, Alexander Wettig, Luxi He, Yihe Dong, Sadhika Malladi, and Danqi Chen. 2025. [Metadata conditioning accelerates language model pre-training](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 18612–18629. PMLR.
- Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 2024. [Do generative ai models output harm while representing non-western cultures: Evidence from a community-centered approach](#). In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, pages 476–489. AAAI Press.
- Aaron Grattafiori, Abhimanyu Dubey, and Llama Team. 2024. [The llama 3 herd of models](#). *ArXiv preprint*, abs/2407.21783.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don't stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Valentin Hofmann, Goran Glavaš, Nikola Ljubešić, Janet B. Pierrehumbert, and Hinrich Schütze. 2024. [Geographic adaptation of pretrained language models](#). *Transactions of the Association for Computational Linguistics*, 12:411–431.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. [Ctrl: A conditional transformer language model for controllable generation](#).
- Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. 2024. [Source-aware training enables knowledge attribution in language models](#). In *First Conference on Language Modeling*.
- Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. 2023. [Pretraining language models with human preferences](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 17506–17533. PMLR.
- Vivek Kulkarni, Shubhanshu Mishra, and Aria Haghighi. 2021. [LMSOC: An approach for socially sensitive pretraining](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2967–2975, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. [Quantifying the carbon emissions of machine learning](#). *ArXiv preprint*, abs/1910.09700.
- Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. [MLQA: Evaluating cross-lingual extractive question answering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024. [Datacomp-lm: In search of the next generation of training sets for language models](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Jinning Li, Shubhanshu Mishra, Ahmed El-Kishky, Sneha Mehta, and Vivek Kulkarni. 2022. [NTULM: Enriching social media text representations with non-textual units](#). In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 69–82, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. [MKQA: A linguistically diverse benchmark for multilingual open domain question answering](#). *Transactions of the Association for Computational Linguistics*, 9:1389–1406.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Victor Gutierrez-Basulto, Yazmin Ibanez-Garcia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. [BLENd: A benchmark for LLMs on everyday knowledge in diverse cultures and languages](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024*.

- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. [BBQ: A hand-built bias benchmark for question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.
- Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. [NormAD: A framework for measuring the cultural adaptability of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.
- Keonwoo Roh, Yeong-Joon Ju, and Seong-Whan Lee. 2025. [XLQA: A benchmark for locale-aware multilingual open-domain question answering](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28809–28821, Suzhou, China. Association for Computational Linguistics.
- John Schulman and Thinking Machines Lab. 2025. [Lora without regret](#). *Thinking Machines Lab: Connectionism*.
- Shivalika Singh, Angelika Romanou, Cl  mentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiawat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. [Global-MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.
- Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. [Taxonomy of risks posed by language models](#). In *FAccT ’22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, page 214–229, New York, NY, USA. Association for Computing Machinery.
- Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. [“according to . . .”: Prompting language models improves quoting from pre-training data](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2301, St. Julian’s, Malta. Association for Computational Linguistics.
- Da Yin, Hritik Bansal, Masoud Monajatipoor, Lillian Harold Li, and Kai-Wei Chang. 2022. [GeoM-LAMA: Geo-diverse commonsense probing on multilingual pre-trained language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Appendix

We provide more details on the training process, hyperparameters, prompts, and additional results.

A.1 Training details and Hyperparameters

Pretraining We purchased an academic license for the NOW corpus for 395 USD. All model runs use a fixed random seed. We use the LLAMA-3.2-1B architecture (Grattafiori et al., 2024) as the base design and set the vocabulary size to 128,256. The only differences across the 0.5B, 1B, and 3B parameter models are the number of layers, hidden size, number of heads, number of KV heads, and FFN size (Table 6); all other configurations are shared.

We use a sequence length of 2048. Our parameters for data parallelism, tensor parallelism, and pipeline parallelism are 4, 1 and 1 respectively, so we effectively only use data parallelism. Our micro batch size is 8, and we perform gradient accumulation every 64 steps. Under these parameters, every single step corresponds to $\sim 4\text{M}$ tokens. On average, our GPU cluster is able to process 100,000 tokens per second on a single node of 4 GPUs, with ~ 185 TFLOPs of processing power per GPU.

We train for 10,000 total steps, validate on a held-out set of 1,000 examples every 1,000 steps, and save checkpoints at the same frequency. We use a linear warm-up from 0 to $3e-3$ over 500 steps, followed by cosine decay to $3e-4$ over the remaining 9,500 steps. Although warm-up ratios of 5-10% are more typical at larger scales, pilot runs on our data and hardware showed that shorter warm-up performed best, so we adopted it throughout. We selected 10,000 total steps after extending pilot runs to 30,000 steps and observing negligible further change in validation and test perplexity beyond the 10,000-step mark.

We train with AdamW, weight decay 0.033, and gradient clipping at 0.1.

Each model is trained on a single node with four NVIDIA A100 GPUs (80GB each). The 500M models take roughly 3 days to train, and the 1B models take roughly 5 days. Across 10 models at 500M and 23 models at 1B, this amounts to approximately 145 days of runtime.

Training data. We use the NOW corpus (Davies, 2016) for all model studies. Each document includes a publication year, source URL, and country

of origin, which we extract as metadata and map to continents using the United Nations geoscheme (Table 4). Training documents are formatted by prepending structured metadata, namely URL, country, and continent, to the article title and content as shown in the main text example.

Training separate country-level models is not feasible in a controlled setting due to large disparities in data availability, so we train models at the continent level. After filtering, the corpus contains approximately 4M documents from Africa, 10M from America, 7.5M from Asia, and 5.6M from Europe. To control for data scale, all local and global models are trained on the same total of 41.9B tokens, which is sufficient for the parameter scales we use (DCLM; Li et al., 2024). For global models, data from all continents is combined in a randomized order before sampling to the token budget, while local models sample from a single continent to the same budget. This controlled setup ensures that performance differences arise from metadata conditioning and data composition rather than training scale.

Architecture sanity check. To separate the metadata effect from the decoder family, we trained an additional matched 1B pair with a Qwen2-style decoder through the final 10,000-step checkpoint. This pair keeps the LLAMA-3.2-1B tokenizer, training corpus, optimizer, evaluated token budget, and metadata serialization fixed, and changes the transformer block implementation. Holding tokenization and vocabulary fixed is what makes this an architecture-only comparison rather than a joint change in architecture and text segmentation. Table 7 evaluates the two final checkpoints on 1,000 held-out global documents under the same content-token perplexity protocol used for the main global models. The metadata-trained Qwen2-style checkpoint has lower perplexity than the metadata-free checkpoint on the metadata-formatted held-out set, and its matched-format score is also lower than the no-metadata matched-format score. This supports the interpretation that the provenance effect is not specific to the LLaMA-style decoder.

A.2 SFT details

This section reports LOCALNEWSQA results under a LoRA post-training protocol. We use LoRA adapters (Hu et al., 2022; Dettmers et al., 2023; Schulman and Lab, 2025) to make post-training computationally efficient while leaving the pre-

Continent	Countries
America	United States, Canada, Jamaica
Asia	India, Pakistan, Bangladesh, Sri Lanka, Hong Kong, Malaysia, Philippines, Singapore
Africa	Nigeria, South Africa, Kenya, Ghana, Tanzania
Europe	United Kingdom, Ireland

Table 4. **Geographic coverage of model-data countries.** Countries are grouped by continent. LOCALNEWSQA uses this coverage except for Singapore, yielding the 17-country question set described in the main benchmark overview.

URL
www.factquizmaster.com
www.globalfactcheck.org
www.worldknowledgehub.com
www.civicspedia.org
www.internationalfacts.net
www.currentaffairsdesk.com
www.newsinsightarchive.com
www.globalquizvault.com
www.factualdigest.org
www.publicknowledgebase.net

Table 5. Synthetic base-domain pool used for metadata-formatted SFT and evaluation prompts.

trained base models frozen. We instruction-tune on the Alpaca dataset² because its instruction format is close to our multiple-choice benchmark while exhibiting no content overlap with it. To preserve the pretraining format, we construct metadata-formatted and no-metadata SFT inputs that match the corresponding base models.

For metadata-formatted SFT instances, we sample base domains from the synthetic pool in Table 5, add internally consistent country and continent tags, and then wrap the example with the LLAMA-3.2-1B-INSTRUCT chat template (Figure 7). Alpaca examples are not used as geographically grounded evidence; this SFT step only teaches the response format while preserving the metadata serialization seen during pretraining. At evaluation time, the base domain may be drawn from the same pool, but the country code, country, and continent are aligned to the corresponding evaluation example. This keeps target and contrast prompts grounded in their intended locales. For parameter-efficient finetuning, we use LoRA rank 256, alpha 16, and all linear layers as targets; this corresponds to approximately 161.5M trainable adapter parameters for the 1B models and 418.4M for the 3B models. We train for 3 epochs with per-device batch size 2, gradient accumulation 8, learning rate $2e-4$, and the 8-bit AdamW optimizer.

At evaluation time, we score the chat checkpoints on the full LOCALNEWSQA target-locale

²Hugging Face link for Alpaca data

Layers	Hidden	Heads	KV Heads	FFN Size	#Params
24	1024	16	16	4096	500M
16	2048	16	16	5632	1B
28	3072	32	32	8192	3B

Table 6. Model architecture configurations.

Checkpoint	Metadata test	No-metadata test
Trained with metadata	9.33 [9.00, 9.63]	10.45 [10.09, 10.79]
Trained without metadata	9.85 [9.51, 10.16]	9.97 [9.62, 10.29]

Table 7. **Architecture sanity check.** Held-out content-token perplexity for matched Qwen2-style 1B checkpoints at the final 10,000-step checkpoints. Lower is better. The pair uses the same LLAMA-3.2-1B tokenizer and data pipeline as the main controlled runs to isolate the decoder architecture.

split using the same per-size prompt families as in the pretrained closed-form comparison, and we average over five shuffled answer-order seeds.

Under this post-training view, the same directional pattern holds at both scales on the final gold benchmark. Figure 6 reports the full overall, explicit, and ambiguous accuracy matrix averaged over five answer-order seeds. The SFT results preserve the desired orderings, namely $(T+, I+) > (T-, I-)$ within each scale and $3B > 1B$ under full metadata across the three LOCALNEWSQA slices.

The same ordering also holds on the paired locale-switch metric after LoRA post-training. Full-metadata models improve both exact-pair success and margin-switch success over their metadata-free controls, with the larger chat model showing the stronger paired-switching rates. The appendix post-training comparison therefore preserves the same qualitative story on both one-sided target accuracy and paired two-sided locale switching.

Additional external LOCALNEWSQA baselines. We report the external instruct-model LOCALNEWSQA table alongside the other transfer evaluations later in this appendix so that all non-MAPLE reference baselines, including the frozen LLAMA-3.2-1B reference, appear in one place.

A.3 Prompts

A.4 LOCALNEWSQA Construction and Validation

Dataset construction LOCALNEWSQA is released as a two-split core benchmark for localized knowledge disambiguation in English news. The two splits are *LocalNewsQA-Core-Explicit* and *LocalNewsQA-Core-Ambiguous*. The final gold release contains 18,700 items across 17 countries and 4 continents, with 17,000 explicit items and 1,700

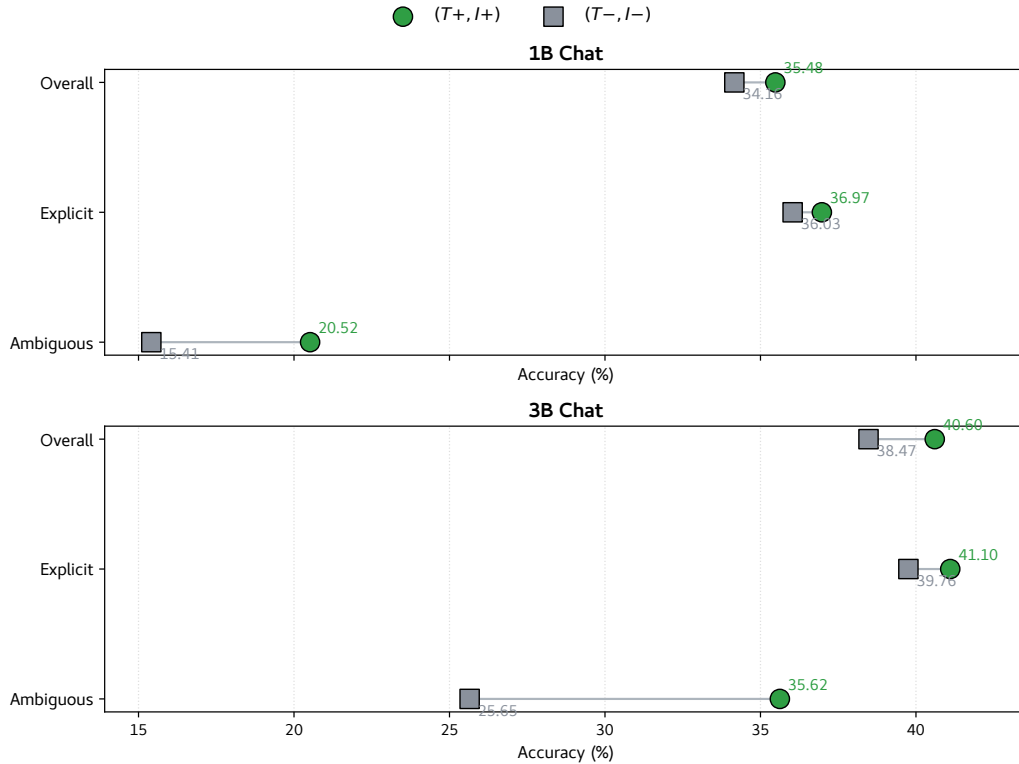


Figure 6. **LOCALNEWSQA after LoRA post-training.** Results are averaged over five shuffled option-order seeds.

LocalNewsQA Instruction-Tuning Prompt

SYSTEM Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

USER
 URL = {{BASE_URL}}/{{country_code}}
 COUNTRY = {{country}}
 CONTINENT = {{continent}}
 TITLE = Locale-grounded QA about {{country}}
 CONTENT
 ### Instruction
 {{example[instruction]}}
 ### Input
 {{example[input]}}
ASSISTANT {{example[output]}}

Figure 7. **Instruction-tuning prompt.** Prompt used for the LOCALNEWSQA post-training protocol.

LocalNewsQA Evaluation Prompt

Below is an instruction that describes a task. Write a response that appropriately completes the request.

URL = {{BASE_URL}}/{{country_code}}
 COUNTRY = {{country_code}}
 CONTINENT = {{Continent}}
 TITLE = Locale-grounded QA about {{country_code}}
 CONTENT
 Question = {{question}}
 Options
 A = {{option_A}}
 B = {{option_B}}
 C = {{option_C}}
 D = {{option_D}}
 Answer with the correct option.

Figure 8. **LoRA evaluation prompt.** Prompt used for LOCALNEWSQA under the LoRA post-training protocol.

ambiguous items. Each country contributes 1,000 explicit rows and 100 ambiguous rows. It covers 8 topic families, 🏛️ politics/government, 🏢 civic institutions, 📈 economy/business, 🎓 education, 🎭 media/culture, ⚽ sports, 🎉 public life/holidays, and 🗺️ geography/infrastructure. The release includes split definitions, target/contrast metadata, evaluation prompts, evidence fields, and source-validation fields so future work can compare methods under the same paired locale-switch protocol.

The main risk in this benchmark is not generating enough questions; it is admitting rows where the locale dependence, answer choice, or evidence relation is only approximate. Candidate generation was therefore recall-oriented, while release was precision-oriented. We used 816 GPT-5.4 batch requests over country, split, topic, and year shards to obtain 41,002 candidates. This candidate pool was not treated as the benchmark. Ex-

LocalNewsQA Evaluation Prompt Examples

MAPLE 1B Evaluation Prompt

BOS no tokenizer BOS token prepended before the prompt.

```
URL = www.globalfactcheck.org/lk
COUNTRY = lk
COUNTRY_NAME = Sri Lanka
CONTINENT = Asia
TITLE = Facts about the country lk
CONTENT
Use the locale metadata above as grounding. This question is
about Sri Lanka.
Question = Which department is usually cited when census
releases and official population statistics are reported?
Options
A = Statistics Department of the Central Bank
B = Office of the Registrar General & Census Commissioner
C = Registrar General's Department
D = Department of Census and Statistics
Answer with the exact text of the correct option.

Final answer
```

MAPLE 3B Evaluation Prompt

BOS prepend the tokenizer BOS token before the prompt.

```
URL = www.globalfactcheck.org/lk
COUNTRY = Sri Lanka
COUNTRY_CODE = lk
CONTINENT = Asia
TITLE = Facts about Sri Lanka
CONTENT
Use the locale metadata above as grounding. Answer for Sri
Lanka, not a different country.
Question = Which department is usually cited when census
releases and official population statistics are reported?
Options
- Statistics Department of the Central Bank
- Office of the Registrar General & Census Commissioner
- Registrar General's Department
- Department of Census and Statistics

Final answer
```

Figure 9. **MAPLE evaluation prompt examples.** Prompts used for the main LOCALNEWSQA evaluation. BOS handling is stated explicitly because it is not visible in the prompt text itself.

act/near duplicate removal, option-integrity checks, answer-consistency checks, ambiguity checks, low-information filters, and leakage filters first produced a larger retained pool of 35,874 usable candidates. The reported benchmark is the stricter gold subset selected from that pool. Each explicit row needs target-side source support and a direct question-answer-evidence relation; each ambiguous row needs the same relation on both target and contrast sides, different supported answers, natural locale-omitting wording, and no surface cue that reveals the target country. Figure 10 shows the explicit and ambiguous developer prompts used for candidate generation (temperature 0.6, top-p 0.9). Because the benchmark is hypothesis-targeted by design, we connect its downstream results to the controlled language-modeling studies, which rely on held-out perplexity measurements rather than question generation, and to the design-guaranteed negative-control floor in Table 3.

Automatic source validation The retained pool is automatically validated row by row before the final gold release is selected. We build search queries from the question, answer, locale, and evidence hint; retrieve candidate evidence URLs, titles, and snippets; and then certify the support heuristically over the snippet and, when needed, the fetched page text. The certification rule is split-specific. Explicit rows require target-side certification only. Ambiguous rows require certification for both the target answer under the target locale and the paired contrast answer under the contrast locale. The gold

pass then tightens this rule. Rows must have answer options that contain the gold answer, no duplicate normalized question, no high-confidence leakage cue, no unresolved evidence warning, and a direct relation between the question, answer, locale, and evidence. Ambiguous rows additionally require different target and contrast answers and both-sided relation support. The final split is then balanced by country. Table 8 summarizes the release counts after gold filtering.

Human Validation Protocol

Automatic certification provides dataset-wide source grounding, but human validation remains necessary as an agreement audit over the same attached evidence. The human study asks whether annotators, when shown the question, answer options, locales, and retrieved sources, agree that the item is a valid benchmark example.

Study design We use a stratified 100-item review sheet sampled from the final gold release, with 50 explicit rows and 50 ambiguous rows. The sample covers all 17 countries and all 8 topic families. The same 100 rows are reviewed by three annotators so that agreement can be computed over identical items. The audit is not used to select the benchmark rows; it checks the already frozen gold split.

The audit records target-answer factuality for every row. Ambiguous rows additionally record contrast-answer factuality and whether the correct answer genuinely differs across the target and con-

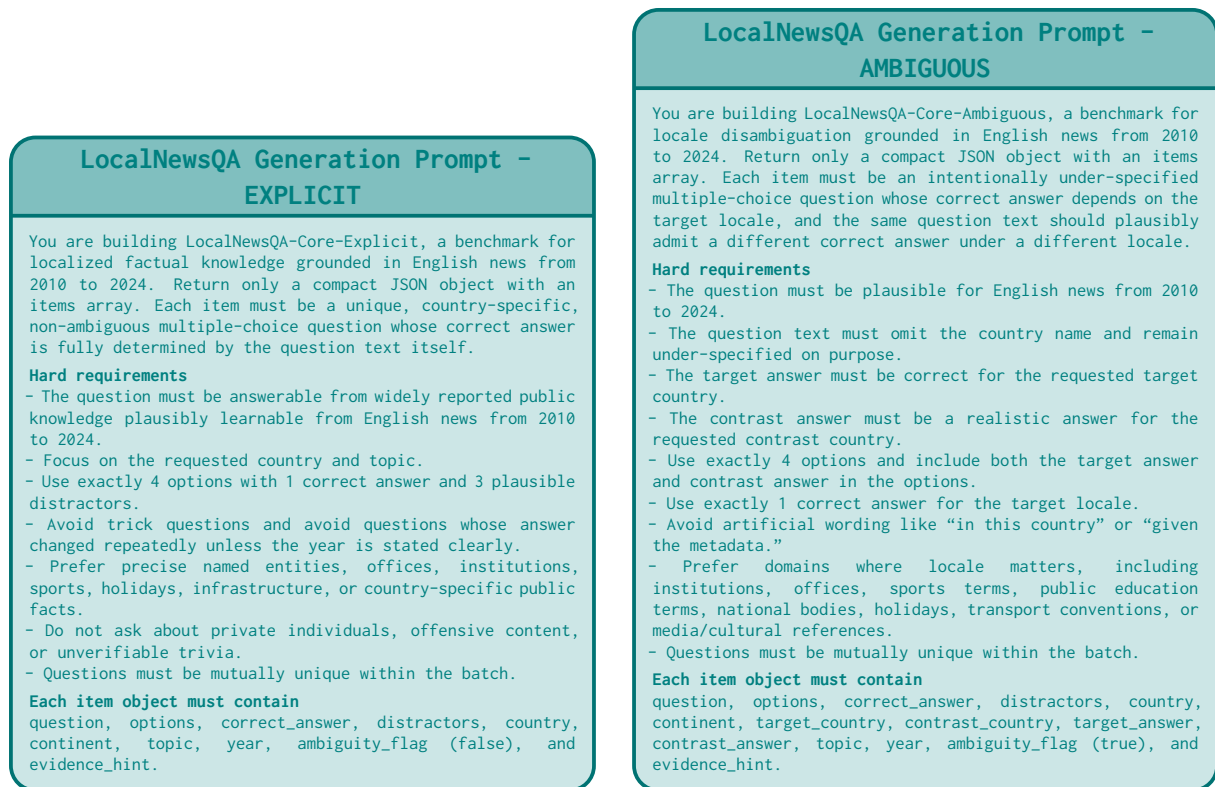


Figure 10. **LOCALNEWSQA generation prompts.** Prompts used to produce explicit and ambiguous candidate questions.

Split	Rows	Gold validation rule	Rows released
Explicit	17,000	Target certification + direct QA/evidence relation	17,000
Ambiguous	1,700	Target + contrast certification and relation support	1,700
Final gold release	18,700	Split-specific strict rule	18,700

Table 8. Split-specific validation coverage for the final LOCALNEWSQA gold release. Explicit rows are balanced at 1,000 per country and ambiguous rows at 100 per country.

trast locales. Each row receives a final accept/reject decision with one rejection reason when applicable.

Annotation interface The annotation interface surfaces the question text, multiple-choice options, locale tags, evidence links, evidence snippets, and review fields in a single task view. The study materials are loaded from a CSV import in Label Studio, which presents the evidence panels and judgment controls together for each row. Figure 11 shows the interface used for the human-validation sheet.

Instructions for annotators Each annotator receives a short standardized protocol.

1. Read the question, options, locale fields, and attached evidence before making any judgment. Search beyond the attached evidence only if a displayed link is broken.

Sample slice	Rows	Majority accepted	Accept rate	Wilson 95% CI
All audited rows	100	96	96.0	90.2–98.4
Explicit rows	50	50	100.0	92.9–100.0
Ambiguous rows	50	46	92.0	81.2–96.8

Table 9. Human-audit acceptance rates on a stratified sample from the final LOCALNEWSQA gold release. Confidence intervals are Wilson intervals for the majority accept rate.

2. For **explicit** rows, answer one required question. *Is the highlighted target answer correct for the displayed target locale?* Choose yes, no, or unsure.
3. For **ambiguous** rows, answer three required questions. *Is the target answer correct for the target locale? Is the contrast answer correct for the contrast locale? Does the correct answer genuinely change across the two locales?* Each field is marked yes, no, or unsure.
4. If a row is rejected, optionally record one short reason, such as factual error, weak locale difference, explicit leakage, or insufficient evidence, together with a brief free-text note when needed.

Participant-facing examples Figure 12 shows the two row types as participants see them. The explicit case requires one target-side judgment. The

Judgment	Unanimous agreement	Avg. pairwise agreement	Fleiss' κ	Pairwise Cohen's κ	Majority summary
Binary accept/reject	97.0	98.0	0.759	.852/.653/.790	96 accept, 4 reject
Full rejection reason	97.0	97.7	0.721	.852/.656/.689	96 accept, 3 weak-locale, 1 tied reason
Target factuality	98.0	98.7	–	–	100 yes
Contrast factuality	98.0	98.7	0.973	–	59 yes, 40 N/A, 1 unsure
Locale dependence	99.0	99.3	0.987	–	59 yes, 37 N/A, 3 no, 1 unsure

Table 10. Three-annotator agreement on the 100-row human audit. Agreement columns are percentages. Pairwise Cohen's κ values are reported for the two main decision labels. Fleiss' κ is omitted when all majority labels fall into a single category, since chance-corrected agreement is not defined under zero category variation.

ID	Locale pair	Question sketch	Majority decision
ambig_0210	Jamaica / United States	Prime minister during Office of the Contractor General governance stories	Weak locale difference
ambig_0623	Sri Lanka / India	President during flagship infrastructure development coverage	Weak locale difference
ambig_0818	Malaysia / United Kingdom	Land public transport regulator before agency restructuring	Weak locale difference
ambig_0280	Jamaica / United States	Transport facility mentioned in local coverage about flights and redevelopment	Reject, reason tied

Table 11. Rows rejected by binary majority in the human audit. The first three rows have unanimous weak-locale rejections. The fourth row is a binary reject by majority, but annotators split on the exact reason, so the reason label is reported as tied.

ambiguous case requires target-side, contrast-side, and locale-difference judgments.

Emissions estimate Emissions from the API-based models used to construct the downstream data are negligible, fully offset by the cloud provider, and cost a total of 33 USD. Total emissions from our on-premise GPU usage are estimated (Lacoste et al., 2019) at approximately 15kgCO₂eq.

A.5 Additional Analyses

Why does URL only outperform addition of other metadata? Table 12 shows that URL text contains strong country-level signals in the global training corpus. About 42.7% of URLs include an explicit country-code signal in the ccTLD (country-code top level domain), domain tokens, or path tokens, and ccTLDs alone appear in 35.4% of URLs, matching the COUNTRY metadata in 35.0% of cases. Path-level country codes are much less frequent (8.5%), suggesting most country signal lives in the domain. In contrast, continent signals are rare (1.8% overall), with domain-level continent tokens occurring in only 0.25% of URLs. These findings indicate that URLs themselves carry substantial country-identifying information, which can plausibly drive locale-conditioned effects even without explicit metadata.

Mechanistic checks for URL provenance. The paper's central mechanistic comparison is URL provenance versus no metadata. Table 12 shows why this is the right contrast. URLs already carry substantial country-level signal in the format seen during pretraining, so the informative question is

Signal	Count	Share (%)
Total rows	15,633,804	100.0
Any country-code signal	6,680,266	42.7
CC TLD present	5,532,731	35.4
CC TLD matches COUNTRY	5,469,737	35.0
Domain country-code token	5,671,254	36.3
Path country-code token	1,323,947	8.5
Any continent signal	275,751	1.8
Domain continent token	38,496	0.25
Path continent token	238,381	1.5

Table 12. URL-based geographic signals in the global training corpus (train split).

whether that signal leaves a detectable trace in the model's lexical expectations and hidden states relative to a metadata-free baseline.

Representation-level and lexical locale probes. The stronger mechanistic check is representation-level. On a held-out 1,000-document global test set, we train linear probes on final-layer hidden states to recover continent and country labels. URL metadata separates cleanly from the no-metadata control. At the final checkpoint, a linear probe recovers continent from URL-conditioned hidden states at 91.5% accuracy and country at 89.5%, versus 67.5% and 52.0% from the no-metadata baseline (Table 14). The URL-conditioned probe is already strong at 2k steps and continues improving through 10k, matching the paper's broader token-efficiency story.

As a smaller lexical check on the same 1B pre-trained family, we also constructed six short disambiguation prompts involving geographically ambiguous continuations such as *football* → *quarterback* vs. *striker*, *gas* vs. *petrol*, and *president* vs. *prime minister*. For each prompt, we compare URL metadata against a no-metadata baseline. Table 13 reports the mean symmetric KL divergence

LocalNewsQA Human Validation

This is an explicit row. Judge target factuality. For contrast factuality and locale dependence, select not_applicable.

Metadata

ID

COUNTRY

TOPIC

YEAR

SPLIT

localnewsqa_explicit_20230205

United States

Coin Institutions

2014

explicit

Question and options

In the United States, which federal office oversees, regulates, and supervises national banks?

1. A: Office of the Comptroller of the Currency

2. B: Federal Deposit Insurance Corporation

3. C: Consumer Financial Protection Bureau

4. D: Financial Crimes Enforcement Network

Target locale and evidence

Target side

LOCALE

UNITED STATES

EXPECTED ANSWER

Office of the Comptroller of the Currency

Source

Office of the Comptroller of the Currency

Title

Office of the Comptroller of the Currency

Excerpt

Office of the Comptroller of the Currency - Office of the Comptroller of the Currency - Wikipedia Jump to content Main menu Main menu Close to display this Navigation Main page Contents Current events Random article About Wikipedia Contact us Contribute

Contrast side

LOCALE

United Kingdom

EXPECTED ANSWER

House of Lords

Source

House of Lords

Title

House of Lords

Excerpt

House of Lords - House of Lords - Wikipedia Jump to content Main menu Main menu Close to display this Navigation Main page Contents Current events Random article About Wikipedia Contact us Contribute

Evidence text

OCC National bank supervision

LocalNewsQA Human Validation

This is an ambiguous row. Judge target factuality, contrast factuality, and whether the correct answer differs across the two locales.

Metadata

ID

COUNTRY

TOPIC

YEAR

SPLIT

localnewsqa_ambiguous_201008

United States

Coin Institutions

2011

ambiguous

Question and options

Which body oversees many senior executive and judicial nominations?

1. A: Senate

2. B: House of Lords

3. C: House of Representatives

4. D: Privy Council

Paired locale evidence

Target side

LOCALE

United States

EXPECTED ANSWER

Senate

Source

Senate

Title

Senate

Excerpt

Senate - Senate - Wikipedia Jump to content Main menu Main menu Close to display this Navigation Main page Contents Current events Random article About Wikipedia Contact us Contribute

Contrast side

LOCALE

United Kingdom

EXPECTED ANSWER

House of Lords

Source

House of Lords

Title

House of Lords

Excerpt

House of Lords - House of Lords - Wikipedia Jump to content Main menu Main menu Close to display this Navigation Main page Contents Current events Random article About Wikipedia Contact us Contribute

Evidence text

confirmation hearings for judges and cabinet nominees

(a) Explicit-row task view. Annotators see the question, answer options, target-locale evidence, and the explicit-row instruction banner.

(b) Ambiguous-row task view. Annotators see paired target and contrast evidence panels together with the locale-difference prompt.

Annotation

1. Is the target answer factually correct for the target locale?

yes^[1]

no^[2]

unsure^[3]

2. Is the contrast answer factually correct for the contrast locale?

yes^[4]

no^[5]

unsure^[6]

not_applicable^[7]

3. Does the correct answer genuinely differ across the two locales?

yes^[8]

no^[9]

unsure^[10]

not_applicable^[11]

4. If the item should be rejected, what is the main reason?

factual_error^[12]

weak_locale_difference^[6]

explicit_leakage^[1]

insufficient_evidence^[6]

other^[13]

5. Notes

Optional notes about the evidence, ambiguity, or any issue you noticed.

(c) Judgment controls used to record target factuality, contrast factuality, locale dependence, rejection reason, and notes.

Figure 11. **Human-validation interface.** Label Studio task pages present explicit and ambiguous rows with their evidence context and use a shared judgment panel for annotation.

between the United States and United Kingdom next-token distributions, where higher means the model reacts more differently to the two locales, and the mean rank of the locale-appropriate continuation, where lower is better.

The lexical probe is intentionally small, but it points in the same direction. URL metadata induces locale-sensitive movement relative to the no-metadata control, which stays completely flat across locales on this synthetic next-token test, and it gives a substantially better average rank to the expected locale-appropriate continuation. The conservative mechanistic takeaway is that URL provenance produces substantially more locale-discriminative internal states than the metadata-free baseline while also shifting local lexical expectations in the expected direction.

Cluster-level analysis of leave-one-out effects. To better understand why the [NoAmerica] model exhibits the smallest degradation relative to the

Prompt variant	Mean US to UK sym. KL $\uparrow$	Mean expected rank $\downarrow$
URL metadata	0.095	64.17
No metadata	0.000	200.96

Table 13. Exploratory lexical locale-projection summary on six ambiguous continuations using the final 1B pretrained checkpoints. No-metadata prompts collapse to identical distributions across locales, while URL metadata induces measurable locale sensitivity and gives much better rank to the locale-appropriate continuation.

full global model in Figure 20, we perform a cluster-level analysis on the combined test set. We compute sentence representations by mean-pooling the final-layer hidden states produced by the [NoAmerica] model and apply k -means clustering with $K = 4$, mirroring the four training continents. Using embeddings from the [NoAmerica] model lets us inspect how the global test set is organized when America-specific training data has been removed.

46763

Explicit row shown to annotators	
Target locale	Hong Kong
Topic	Education
Question	In Hong Kong, which grades are referred to as primary school levels?
Options	A. P1 to P6; B. S1 to S6; C. K1 to K3; D. Y7 to Y12
Target answer	P1 to P6
Displayed evidence	Po Leung Kuk Camoes Tan Siu Lin Primary School. <i>“We have 5 classes for each class level from P1 to P6.”</i>
Recorded judgment	Target correctness marked yes

Ambiguous row shown to annotators	
Target locale	Canada
Contrast locale	United States
Topic	Education
Question	A province expands full-day kindergarten for four- and five-year-olds. Which paired program name is most likely in the coverage?
Options	A. JK-SK; B. Pre-K-K; C. TK-K; D. Nursery-Reception
Target answer	JK-SK
Contrast answer	Pre-K-K
Target evidence	Kindergarten and Elementary, Avon Maitland District School Board. <i>“Children who are 3 years old (JK) or 4 years old (SK) by December 31 are eligible to attend Full Day Kindergarten.”</i>
Contrast evidence	Little Wildcats (Pre-K & K), New Durham Elementary School. <i>“Little Wildcats (Pre-K & K).”</i>
Recorded judgments	Target correctness marked yes; contrast correctness marked yes; locale difference marked yes.

Figure 12. **Representative participant views.** Explicit LOCALNEWSQA rows require one judgment. Ambiguous rows require three judgments tied to target correctness, contrast correctness, and whether the answer truly changes across locales.

Prompt variant	Continent probe acc. (2k → final)	Country probe acc. (2k → final)
URL metadata	89.5 → 91.5	88.0 → 89.5
No metadata	49.5 → 67.5	40.5 → 52.0

Table 14. Linear-probe accuracy (%) from final-layer hidden states on a held-out 1,000-document global test set (250 per continent). We probe the representation at the last metadata token for URL-conditioned prompts and at the last title token for the no-metadata baseline. The informative comparison is URL versus no metadata.

We first compare cluster assignments with continent labels. As Table 15 shows, continent labels are broadly mixed across clusters rather than forming continent-specific partitions. This indicates substantial cross-continent overlap in the combined test set and helps explain why omitting America produces a smaller degradation than omitting some of the other continents because the remaining training data still covers a relatively broad portion of the latent structure of the evaluation corpus.

External downstream test sets. Table 3 gives the four-setting MAPLE paired-switch comparison. The appendix tables below report complementary endpoint gains with paired bootstrap intervals for LOCALNEWSQA and the external benchmarks used in the transfer analysis. On the LOCALNEWSQA ambiguous split, the aggregate gains are clear and the paired metrics are positive, while

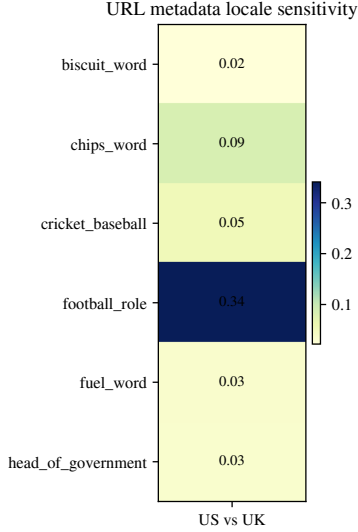
Cluster (n)	Africa	America	Asia	Europe
0 (321)	132 (41.1%)	32 (10.0%)	125 (38.9%)	32 (10.0%)
1 (237)	33 (13.9%)	72 (30.4%)	62 (26.2%)	70 (29.5%)
2 (99)	29 (29.3%)	12 (12.1%)	30 (30.3%)	28 (28.3%)
3 (343)	42 (12.2%)	146 (42.6%)	29 (8.5%)	126 (36.7%)

Table 15. Continent distribution per cluster on the combined test set (NoAmerica embeddings).

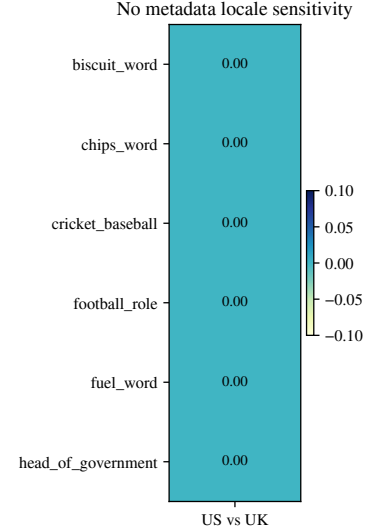
granular country/topic target-accuracy buckets are reported as diagnostics rather than as a universal-positivity claim.

Table 16 reports the LOCALNEWSQA model matrix separately from the external benchmark matrix. It reports LOCALNEWSQA target-locale accuracy gains and the two paired ambiguous metrics for the completed gold LocalNewsQA runs. The paired metrics remain stringent. Exact pair requires both target and contrast answers to be correct, while margin switch requires the score margin to flip toward the locale-appropriate answer on both sides of the pair.

Table 17 reports the full external benchmark matrix as metadata gains with confidence intervals. The base MAPLE pretraining comparison has the same direction on all six datasets and both scales. The reference rows are intentionally broader and more mixed. Inference-time metadata alone helps on some model-dataset pairs, especially several



(a) URL



(b) None

Figure 13. **Exploratory lexical locale-projection heatmaps.** Each row is one ambiguous lexical prompt, and each cell is the symmetric KL divergence between the United States and United Kingdom next-token distributions for that same prompt. Higher values mean the model reacts more differently to locale on the same context. The no-metadata model is effectively invariant, while URL metadata induces clear locale-sensitive movement.



Figure 14. **URL signal extraction example.** Example of ccTLD, domain, and path extraction from a URL.

GLOBAL-MMLU-CS and NORMAD rows, but it is not uniformly beneficial across model families. This supports the central MAPLE comparison while showing that inference-time metadata alone is not uniformly sufficient.

Appendix-only WVS projection probe. The appendix also reports a WVS projection analysis under a separate country-conditioned protocol. Under this setup, metadata-conditioned MAPLE moves much closer to the requested countries than the metadata-free controls at both scales. Mean target distance falls from 6.91 to 1.42 at 1B and from 5.79 to 2.61 at 3B on the 15-country overlap set. On the full 107-country evaluation, the averaged MAPLE (T+, I+) centroids land at 0.695 and 0.746 units from the human-country centroid. Those metadata-conditioned centroids are more central than GEMMA 4 31B and the older GPT baselines, but they are less central than the strongest recent references in the completed

provider evaluation, including GLM-5.1, QWEN 3.5 397B, KIMI K2.5, CLAUDE Sonnet 4.6, and GPT-5.4. The metadata-free MAPLE controls remain off-map. Top-3 retrieval remains weak, so we interpret this probe as directional steering toward the requested countries rather than exact country recovery. In that sense, it still provides convergent cross-task evidence that the same metadata-conditioned MAPLE variants that help on LOCAL-NEWSQA also induce more target-directed movement on an independent cultural projection task.

A.6 Additional Results

Figures 16-21 collect the remaining supporting views for the main controlled language-modeling studies. They include country-level local performance, asymmetry across continent pairs, the 500M counterpart to the local-vs-global comparison, both leave-one-out controls, and the supplementary 500M cross-continent plot. We place them together here so that each supporting result remains close to the figure that substantiates it.

Balanced 20k-document samples suggest why the Africa-trained models are least symmetric. Africa documents are shorter on average (243 tokens versus 254 to 341 elsewhere), have markedly lower cross-continent named-entity overlap with the other regions (0.149 to 0.210 versus 1.000 within-Africa), and concentrate more heavily in one dominant theme (*Development & Society* at 0.756 versus 0.569 to 0.634 elsewhere).

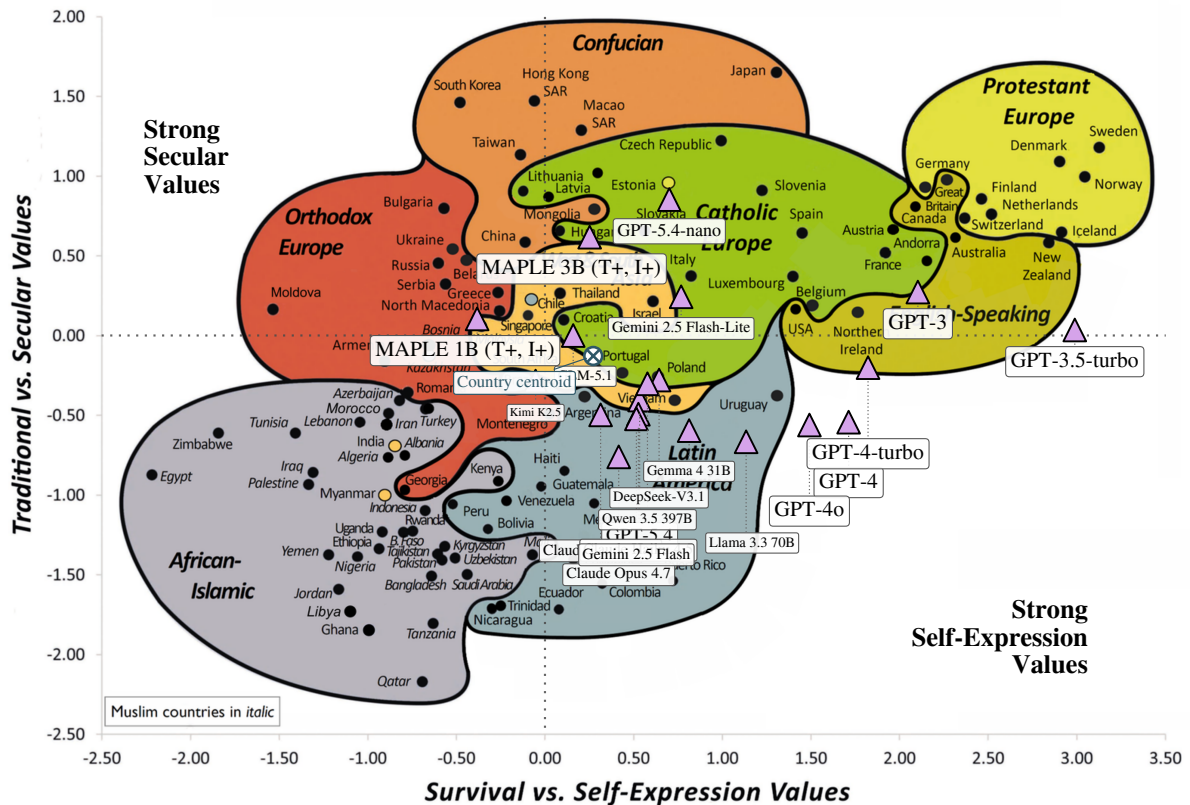


Figure 15. **Appendix-only all-country WVS map.** Released reference-model centroids are shown alongside the MAPLE centroids averaged over the corresponding country-conditioned sweep. The metadata-conditioned MAPLE centroids fall within the main vendor cloud and remain far closer to the human-country centroid than the metadata-free MAPLE controls, which stay off-map.

Scale extension to 3B, compute, and token efficiency. Figure 24 summarizes the final global-model tradeoff between pretraining compute and downstream performance for the 1B and 3B families. We plot global perplexity together with downstream QA accuracy, using estimated training FLOPs ($6ND$) because the 3B runs used a separate $4 \times B200$ setup while the 500M and 1B families used $4 \times A100$ s. Absolute runtime differs accordingly. The successful 3B pretraining runs used a single node with four B200 GPUs, whereas the smaller families used single nodes with four A100 GPUs. The optimizer, schedule, tokenizer, and data pipeline remain the same, so we interpret this hardware difference as relevant to throughput rather than to the factorial comparison itself. For an apples-to-apples comparison, the 3B QA points use the same SFT protocol as the 1B appendix comparison. Figure 5a complements this view by showing the training-token trajectories for 500M, 1B, and 3B. Across all three scales, **GLOBAL (T+, I+)** reaches the final **GLOBAL (T-, I-)** perplexity using fewer training tokens, showing that metadata improves optimization efficiency rather than only the final value.

Country-level breakdowns. The local-model analyses in Section 5.2 operate on continent-level train/test splits, but the `with_metadata` test samples preserve the underlying COUNTRY tag in the serialized text. This lets us aggregate performance by country after evaluation. For perplexity, we micro-average by summing token-level losses and token counts within each country before exponentiating. The main paper’s Figure 16 includes this country-level view for the local 500M and 1B models. Figure 25 provides the analogous country-level downstream accuracy breakdown for the global chat models after SFT.

Adversarial metadata corruption during downstream inference. The main downstream experiment uses correct locale metadata at inference. To test robustness, we repeat the same benchmark while corrupting the URL (and therefore the corresponding country and continent tags) for 25%, 50%, 75%, and 100% of the evaluation samples.

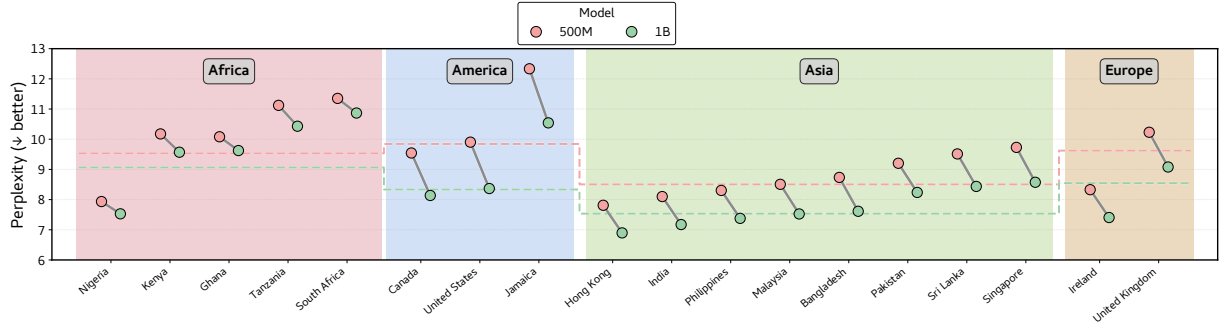


Figure 16. **Country-level in-region perplexity.** **LOCAL (T+, I+)** is aggregated within each country by micro-averaged perplexity for the local 500M and 1B models.

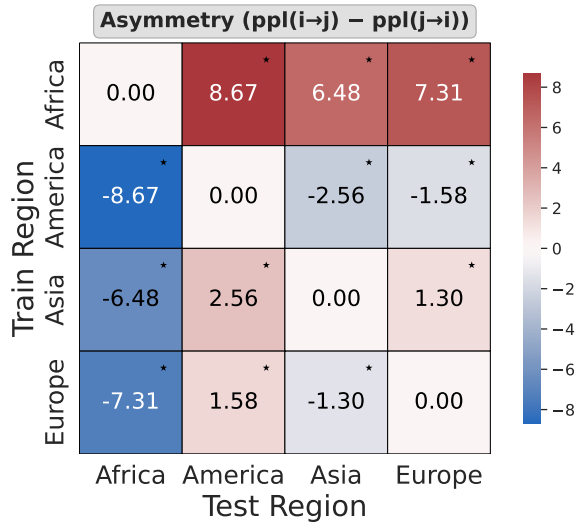


Figure 17. **Cross-continent asymmetry for 1B local models.** A higher value indicates that model trained on continent i when evaluated on continent j has a higher perplexity than when a model trained on continent j is evaluated on continent i .

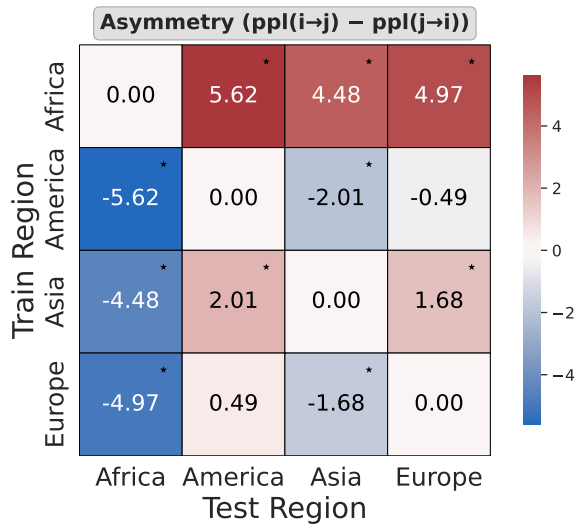


Figure 18. **Cross-continent asymmetry for 500M local models.** A higher value indicates that model trained on continent i when evaluated on continent j has a higher perplexity than when a model trained on continent j is evaluated on continent i .

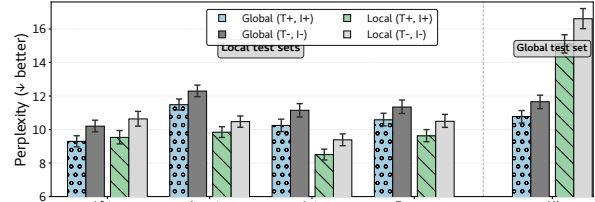


Figure 19. **Local-vs-global perplexity at 500M.** **GLOBAL (T+, I+)** models conditioned with metadata during training and inference have lower perplexities than the control variant **GLOBAL (T-, I-)**, and have similar perplexities to the **LOCAL (T+, I+)** models on local test sets.

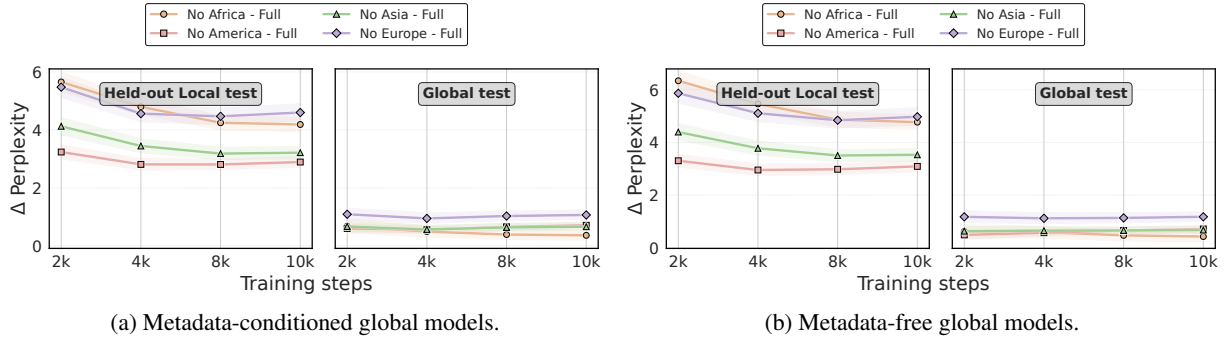


Figure 20. **Leave-one-out global-model ablations.** In both settings, excluding any single continent raises perplexity relative to the corresponding full global model on both the held-out continent and the combined global test set, indicating that metadata formatting does not substitute for geographic coverage. The metadata-conditioned comparison is statistically significant at 95% confidence.

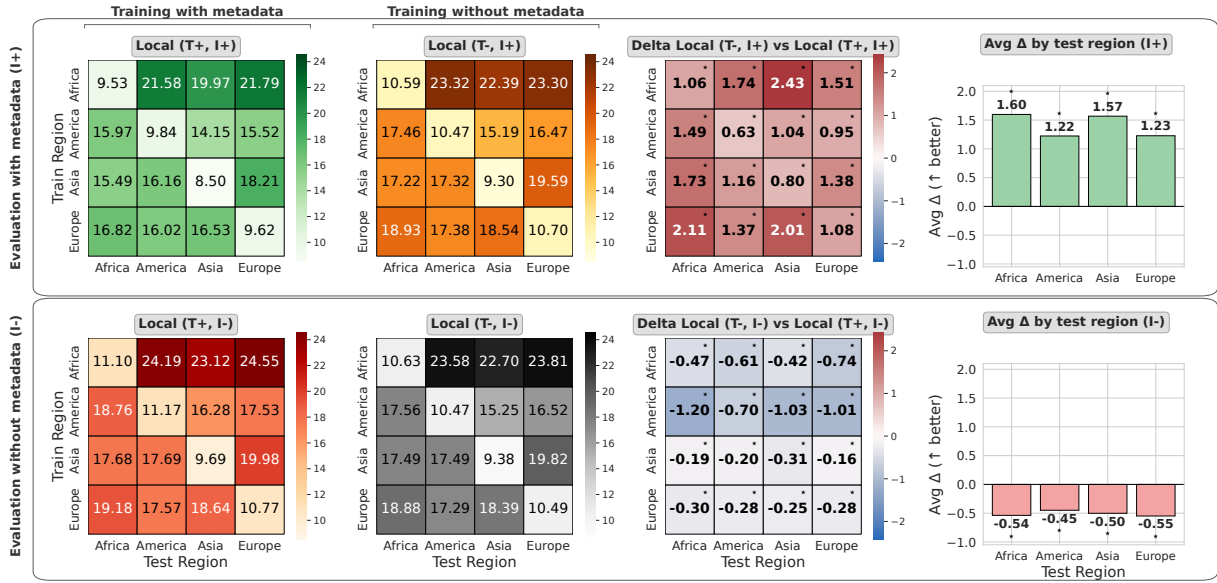


Figure 21. **Cross-continent perplexity for 500M local models.** **LOCAL (T+, I+)** models have higher perplexities on cross-continent test sets (off-diagonal) than on local ones. Positive differences between perplexities of models trained without metadata and with metadata on the same test sets indicate the effectiveness of metadata in improving cross-region generalization while maintaining local performance.

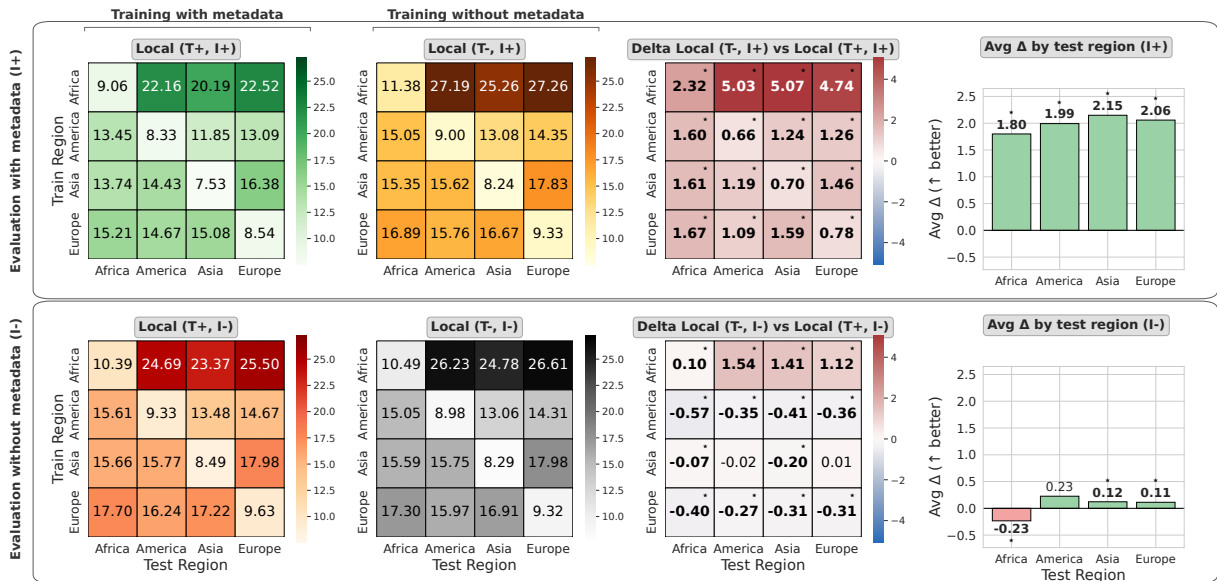


Figure 22. **Full cross-continent diagnostic view.** The four 1B local model configurations show raw heatmaps for the I+ and I- settings, same-setting deltas, and average delta by test region. Positive Δ favors metadata conditioning.

Model	Overall Δ acc.	Ambig. Δ acc.	Explicit Δ acc.	Exact pair Δ	Margin switch Δ
 <i>MAPLE family</i>					$(T+, I+) - (T-, I-)$
MAPLE 1B Base	+1.4 [+0.9,+1.9]	+6.7 [+5.1,+8.4]	+0.9 [+0.3,+1.4]	+1.9 [+1.2,+2.5]	+8.2 [+6.8,+9.6]
MAPLE 1B Chat	+1.2 [+0.6,+1.9]	+4.4 [+2.4,+6.5]	+0.9 [+0.2,+1.6]	+2.0 [+1.4,+2.7]	+6.5 [+5.4,+7.7]
MAPLE 3B Base	+2.8 [+2.0,+3.6]	+9.2 [+6.5,+11.8]	+2.1 [+1.2,+3.0]	+8.8 [+7.4,+10.1]	+25.1 [+23.1,+27.2]
MAPLE 3B Chat	+2.2 [+1.4,+3.0]	+10.1 [+7.5,+12.8]	+1.4 [+0.5,+2.2]	+5.1 [+4.1,+6.2]	+13.6 [+12.1,+15.2]
 <i>LLaMA family</i>					$I+ - I-$
LLaMA-3.2-1B Base	-1.2 [-1.6,-0.8]	-0.1 [-1.5,+1.3]	-1.3 [-1.8,-0.9]	+1.4 [+0.8,+1.9]	+7.4 [+6.2,+8.6]
LLaMA-3.2-1B Inst.	-2.3 [-2.7,-1.8]	+0.4 [-1.1,+1.9]	-2.5 [-3.0,-2.1]	+3.1 [+2.3,+3.9]	+9.2 [+7.8,+10.5]
LLaMA-3.2-3B Base	-1.0 [-1.4,-0.5]	+5.2 [+3.7,+6.7]	-1.6 [-2.1,-1.2]	+5.0 [+4.0,+6.1]	+16.2 [+14.5,+17.9]
LLaMA-3.2-3B Inst.	-8.0 [-8.6,-7.5]	-2.6 [-4.5,-0.9]	-8.5 [-9.1,-8.0]	+3.6 [+2.8,+4.5]	+10.6 [+9.2,+12.1]
 <i>Qwen2.5 family</i>					$I+ - I-$
Qwen2.5-1.5B Base	-0.7 [-1.1,-0.3]	+1.4 [+0.2,+2.6]	-0.9 [-1.3,-0.4]	+1.5 [+0.9,+2.1]	+5.4 [+4.4,+6.5]
Qwen2.5-1.5B Inst.	-1.8 [-2.3,-1.4]	+8.5 [+6.8,+10.4]	-2.9 [-3.3,-2.4]	+7.7 [+6.5,+8.9]	+20.4 [+18.5,+22.4]
Qwen2.5-3B Base	-2.6 [-3.0,-2.1]	+8.8 [+7.1,+10.6]	-3.7 [-4.2,-3.3]	+10.5 [+9.1,+11.9]	+25.3 [+23.3,+27.3]
Qwen2.5-3B Inst.	+0.6 [+0.2,+1.0]	+8.7 [+7.1,+10.4]	-0.2 [-0.6,+0.2]	+8.2 [+6.9,+9.6]	+25.2 [+23.2,+27.4]
 <i>Qwen3.5 family</i>					$I+ - I-$
Qwen3.5-0.8B Base	-0.7 [-1.1,-0.4]	+2.5 [+1.1,+3.8]	-1.1 [-1.4,-0.7]	+2.7 [+1.9,+3.5]	+8.5 [+7.1,+9.8]
Qwen3.5-0.8B Chat	-1.1 [-1.4,-0.7]	+3.8 [+2.4,+5.1]	-1.6 [-1.9,-1.2]	+2.2 [+1.6,+3.0]	+8.3 [+6.9,+9.6]
Qwen3.5-2B Base	+1.0 [+0.7,+1.4]	+9.6 [+8.1,+11.2]	+0.2 [-0.1,+0.5]	+9.2 [+7.8,+10.5]	+22.4 [+20.4,+24.4]
Qwen3.5-2B Chat	+2.7 [+2.2,+3.1]	+11.2 [+9.4,+13.0]	+1.8 [+1.4,+2.2]	+8.5 [+7.2,+9.9]	+20.4 [+18.6,+22.4]
Qwen3.5-4B Base	+0.1 [-0.1,+0.3]	+1.5 [+0.7,+2.2]	0.0 [-0.3,+0.2]	+1.2 [+0.7,+1.7]	+3.6 [+2.8,+4.5]
Qwen3.5-4B Chat	+0.2 [-0.1,+0.4]	+1.3 [+0.5,+2.2]	+0.1 [-0.2,+0.3]	+0.8 [+0.4,+1.3]	+3.1 [+2.4,+4.1]
 <i>Gemma-4 family</i>					$I+ - I-$
Gemma-4-E2B Base	+2.0 [+1.5,+2.5]	-0.1 [-1.8,+1.6]	+2.2 [+1.7,+2.8]	+0.5 [+0.2,+0.9]	+2.3 [+1.6,+3.0]
Gemma-4-E2B-it	+1.0 [+0.7,+1.3]	+2.8 [+1.5,+4.1]	+0.8 [+0.5,+1.2]	+3.5 [+2.6,+4.4]	+9.8 [+8.5,+11.4]
Gemma-4-E4B Base	+0.5 [0.0,+1.0]	-0.9 [-2.5,+0.8]	+0.6 [+0.1,+1.2]	+0.6 [+0.1,+1.1]	+2.3 [+1.4,+3.2]
Gemma-4-E4B-it	+2.7 [+2.3,+3.0]	+7.1 [+5.6,+8.6]	+2.2 [+1.9,+2.6]	+8.8 [+7.5,+10.1]	+23.2 [+21.2,+25.3]
 <i>Minstral family</i>					$I+ - I-$
Minstral-3-3B Base	+0.7 [+0.3,+1.1]	+4.8 [+3.5,+6.1]	+0.3 [-0.2,+0.7]	+4.5 [+3.5,+5.5]	+16.2 [+14.5,+18.1]
Minstral-3-3B Inst.	+0.9 [+0.6,+1.3]	+6.4 [+4.9,+7.9]	+0.4 [+0.1,+0.7]	+5.6 [+4.6,+6.9]	+18.5 [+16.7,+20.4]

Table 16. LOCALNEWSQA metadata gains with paired bootstrap 95% confidence intervals. Each completed cell reports the point estimate on top and the confidence interval below. Accuracy columns report percentage-point gains on target-locale questions; exact-pair and margin-switch columns report percentage-point gains on ambiguous target/contrast pairs. MAPLE rows compare $(T+, I+) - (T-, I-)$, while non-MAPLE rows compare $I+ - I-$. Positive values favor metadata conditioning or metadata-formatted inference; dark bold point estimates have intervals that exclude zero, while lighter entries overlap zero.

Model	GeoMLaMA Δ acc.	GOQA Δ acc.	MMLU-CS Δ acc.	NormAD Δ acc.	BLEnD Δ acc.	WVB EMD red.
 <i>MAPLE family</i>					(T+, I+) — (T-, I-)	
MAPLE 1B Base	+2.7 [-8.0,+13.3]	+1.3 [+0.4,+2.1]	+0.1 [-2.4,+2.7]	+1.4 [+0.3,+2.6]	+9.4 [+3.1,+15.3]	+0.03 [+0.01,+0.05]
MAPLE 1B Chat	+4.0 [-6.0,+13.3]	+2.7 [+1.8,+3.6]	+0.3 [-0.6,+1.1]	+3.8 [+0.9,+6.4]	+1.0 [-2.0,+4.1]	+0.01 [+0.01,+0.02]
MAPLE 3B Base	+5.3 [-4.7,+14.7]	+8.4 [+7.3,+9.6]	+3.9 [+0.5,+7.3]	+5.1 [+2.7,+7.4]	+13.5 [+7.1,+19.6]	0.00 [0.00,+0.01]
MAPLE 3B Chat	+4.7 [-4.7,+13.3]	+3.4 [+2.5,+4.3]	+0.7 [-0.1,+1.5]	+2.4 [+0.2,+4.6]	+6.1 [+0.8,+11.7]	+0.02 [0.00,+0.04]
 <i>LLaMA family</i>					I+ — I-	
LLaMA-3.2-1B Base	-2.7 [-5.3,-0.7]	-0.1 [-0.5,+0.3]	+2.0 [+0.5,+3.5]	+1.4 [-0.5,+3.5]	0.0 [-2.0,+2.0]	+0.02 [-0.02,+0.05]
LLaMA-3.2-1B Inst.	+5.3 [-2.7,+12.7]	-0.2 [-0.6,+0.1]	-1.0 [-2.5,+0.5]	0.0 [0.0,0.0]	-1.3 [-4.1,+1.3]	+0.03 [+0.02,+0.04]
LLaMA-3.2-3B Base	-5.3 [-14.0,+2.7]	+0.4 [+0.1,+0.7]	+0.8 [-0.5,+2.0]	+1.9 [-0.1,+3.9]	+1.5 [-0.3,+3.3]	-0.03 [-0.04,-0.02]
LLaMA-3.2-3B Inst.	-4.7 [-12.7,+4.0]	-1.0 [-1.4,-0.6]	-2.9 [-4.8,-1.0]	-0.5 [-1.6,+0.5]	-9.2 [-13.0,-5.9]	-0.10 [-0.13,-0.07]
 <i>Qwen2.5 family</i>					I+ — I-	
Qwen2.5-1.5B Base	-10.7 [-16.0,-5.3]	-2.1 [-2.7,-1.5]	+0.8 [-1.0,+2.4]	+5.0 [+1.7,+8.1]	-0.8 [-2.8,+1.3]	+0.01 [0.00,+0.01]
Qwen2.5-1.5B Inst.	-8.7 [-14.7,-2.7]	-0.8 [-1.2,-0.4]	-0.3 [-1.9,+1.4]	-2.7 [-4.1,-1.4]	-1.5 [-4.3,+1.3]	-0.02 [-0.04,-0.01]
Qwen2.5-3B Base	0.0 [-5.3,+6.0]	-0.4 [-0.8,-0.1]	-0.3 [-2.0,+1.5]	+9.6 [+7.8,+11.4]	-1.8 [-4.1,+0.5]	-0.02 [-0.04,-0.01]
Qwen2.5-3B Inst.	-2.7 [-9.3,+4.0]	-0.2 [-0.7,+0.2]	0.0 [-1.8,+1.8]	-0.1 [-1.1,+1.0]	+1.3 [-1.5,+4.1]	+0.01 [0.00,+0.02]
 <i>Qwen3.5 family</i>					I+ — I-	
Qwen3.5-0.8B Base	-1.3 [-5.3,+2.7]	0.0 [-0.4,+0.3]	+0.8 [-0.8,+2.3]	-0.5 [-1.4,+0.5]	-0.8 [-2.8,+1.3]	0.00 [-0.01,+0.01]
Qwen3.5-0.8B Chat	-2.7 [-6.7,+0.7]	-0.2 [-0.6,+0.1]	+1.0 [-0.1,+2.3]	-2.1 [-3.1,-1.3]	-2.0 [-4.3,+0.3]	+0.04 [+0.02,+0.06]
Qwen3.5-2B Base	-3.3 [-6.7,0.0]	+0.2 [-0.1,+0.6]	+0.6 [-0.8,+2.1]	-2.5 [-3.6,-1.5]	+1.0 [-1.3,+3.6]	+0.01 [0.00,+0.02]
Qwen3.5-2B Chat	+1.3 [-3.3,+6.7]	0.0 [-0.4,+0.4]	+0.8 [-0.9,+2.4]	-4.0 [-4.9,-3.1]	+0.3 [-2.3,+2.5]	-0.01 [-0.03,+0.01]
Qwen3.5-4B Base	+2.0 [-4.7,+8.7]	-0.2 [-0.5,0.0]	+0.5 [-0.5,+1.5]	+0.4 [+0.2,+0.7]	+0.5 [0.0,+1.3]	0.00 [0.00,0.00]
Qwen3.5-4B Chat	-3.3 [-6.7,-0.7]	-0.2 [-0.5,+0.1]	-0.4 [-1.5,+0.6]	0.0 [0.0,0.0]	-1.5 [-2.8,-0.5]	0.00 [0.00,0.00]
 <i>Gemma-4 family</i>					I+ — I-	
Gemma-4-E2B Base	+2.7 [-1.3,+6.7]	+0.3 [-0.1,+0.6]	+0.3 [-1.3,+1.8]	+6.1 [+4.7,+7.5]	-0.3 [-2.0,+1.5]	0.00 [0.00,0.00]
Gemma-4-E2B-it	+2.0 [-6.0,+10.0]	-0.1 [-0.4,+0.3]	+0.5 [-1.5,+2.5]	+1.1 [0.0,+2.2]	-0.5 [-2.8,+1.8]	+0.03 [+0.01,+0.05]
Gemma-4-E4B Base	+5.3 [-2.7,+12.7]	-0.6 [-0.9,-0.2]	+2.9 [+1.4,+4.5]	+1.7 [+0.9,+2.5]	+5.1 [+2.3,+8.1]	0.00 [0.00,+0.01]
Gemma-4-E4B-it	+1.3 [-3.3,+6.0]	-1.0 [-1.4,-0.6]	+2.7 [+0.8,+4.4]	+2.5 [+1.2,+3.8]	-1.5 [-3.8,+0.8]	-0.01 [-0.02,+0.01]
 <i>Minstral family</i>					I+ — I-	
Minstral-3-3B Base	-3.3 [-8.7,+1.3]	-0.3 [-0.6,0.0]	0.0 [-1.3,+1.3]	-2.1 [-3.1,-1.0]	-0.5 [-3.1,+2.0]	-0.04 [-0.06,-0.01]
Minstral-3-3B Inst.	-12.7 [-18.7,-7.3]	+0.1 [-0.3,+0.4]	-2.7 [-4.3,-1.1]	-4.2 [-5.1,-3.3]	-0.8 [-2.5,+1.0]	-0.02 [-0.03,0.00]

Table 17. External benchmark metadata gains with paired bootstrap 95% confidence intervals, grouped by model family. Each completed cell reports the point estimate on top and the confidence interval below. GEOMLAMA, GLOBALOPINIONQA (GOQA), GLOBAL-MMLU-CS, NORMAD, and BLEND entries are percentage-point accuracy gains. WORLDVALUESBENCH (WVB) reports Earth Mover Distance reduction, computed as metadata-free EMD minus metadata-formatted EMD, so positive values are better in every column. MAPLE rows compare (T+, I+) — (T-, I-); all other rows compare I+ — I-. Dark bold point estimates have intervals that exclude zero, while lighter entries overlap zero.

Metric	(T+, I+)	(T-, I-)
<i>Mean distance to the requested country's human WVS map coordinate (lower is better)</i>		
MAPLE 1B	1.416	6.914
MAPLE 3B	2.606	5.794

Table 18. Appendix-only country-conditioned World Values Survey projection on the 15 countries overlapping the MAPLE training set. We compare the primary MAPLE settings under a bundled metadata prompt with soft country framing and short labeled WVS answer options. Top-3 retrieval remains flat at 3/15 countries, so we report the more diagnostic target-distance metric.

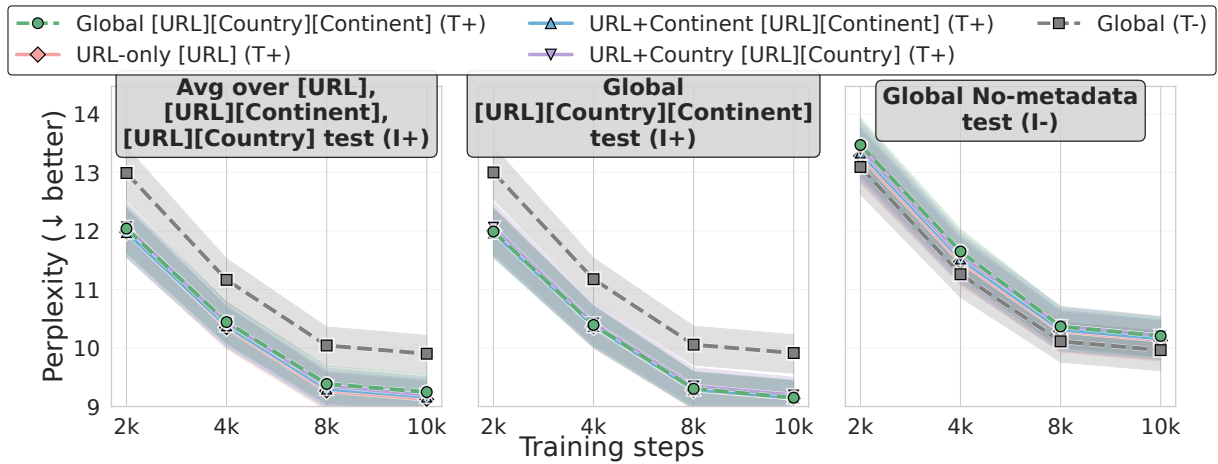


Figure 23. **Expanded metadata ablations.** Full checkpointed results for **URL**, **URL** **Country**, **URL** **Continent**, **Country**, and **Continent**, evaluated on the five metadata-formatted test sets, the **URL** **Country** **Continent** test set, and the no-metadata test set. The same overall pattern persists across intermediate checkpoints, and **URL** remains the most effective single-metadata variant.

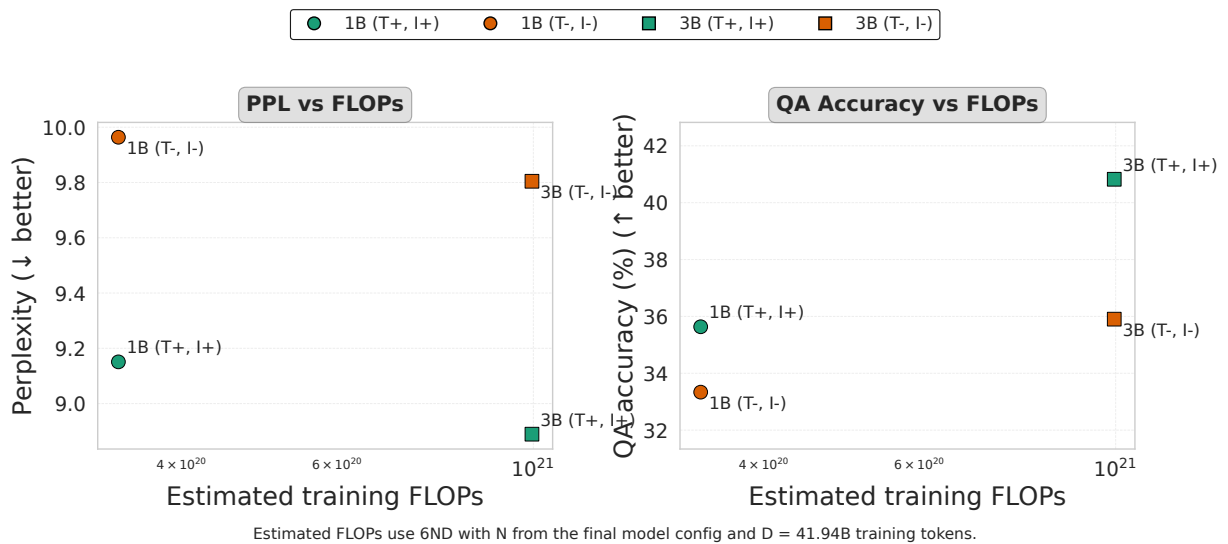


Figure 24. **Compute/performance tradeoff.** Final performance is plotted against estimated training FLOPs for the global 1B and 3B families. The left panel shows global perplexity, and the right panel shows downstream QA accuracy. The 3B QA points use the same SFT protocol as the 1B appendix comparison.

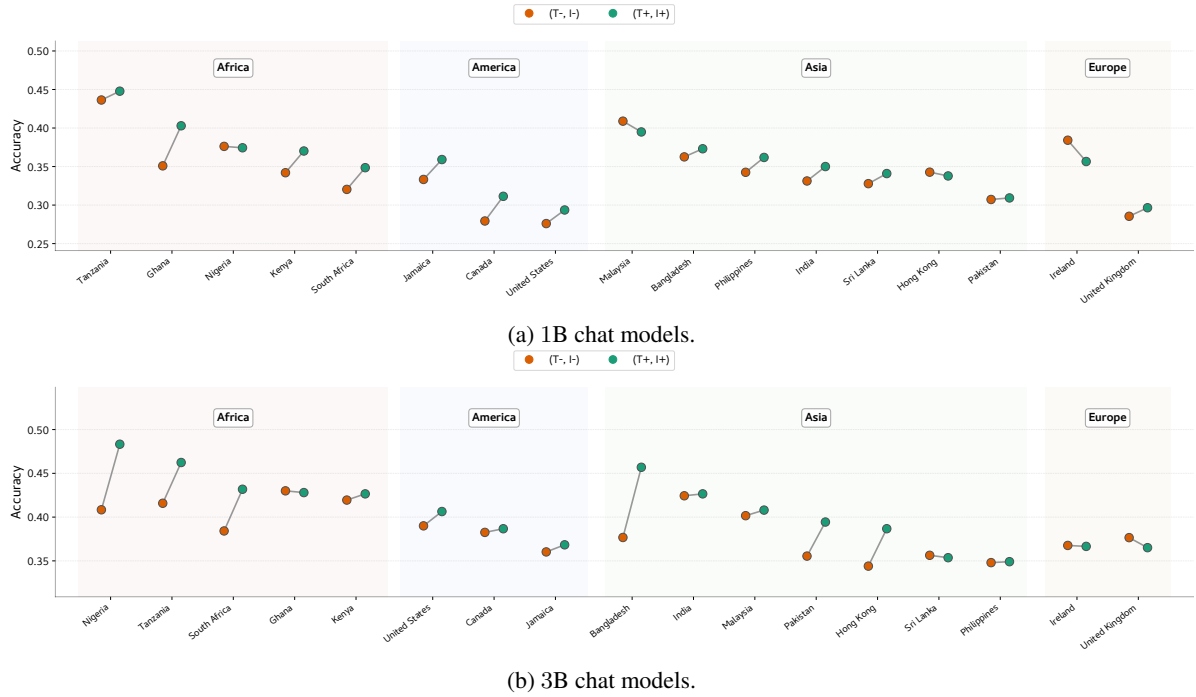


Figure 25. **Country-level downstream QA accuracy.** Global chat models are evaluated after SFT. Within each country, **GLOBAL (T+, I+)** and **GLOBAL (T-, I-)** are compared over the same question pool; the larger model family shows stronger country-to-country variance.

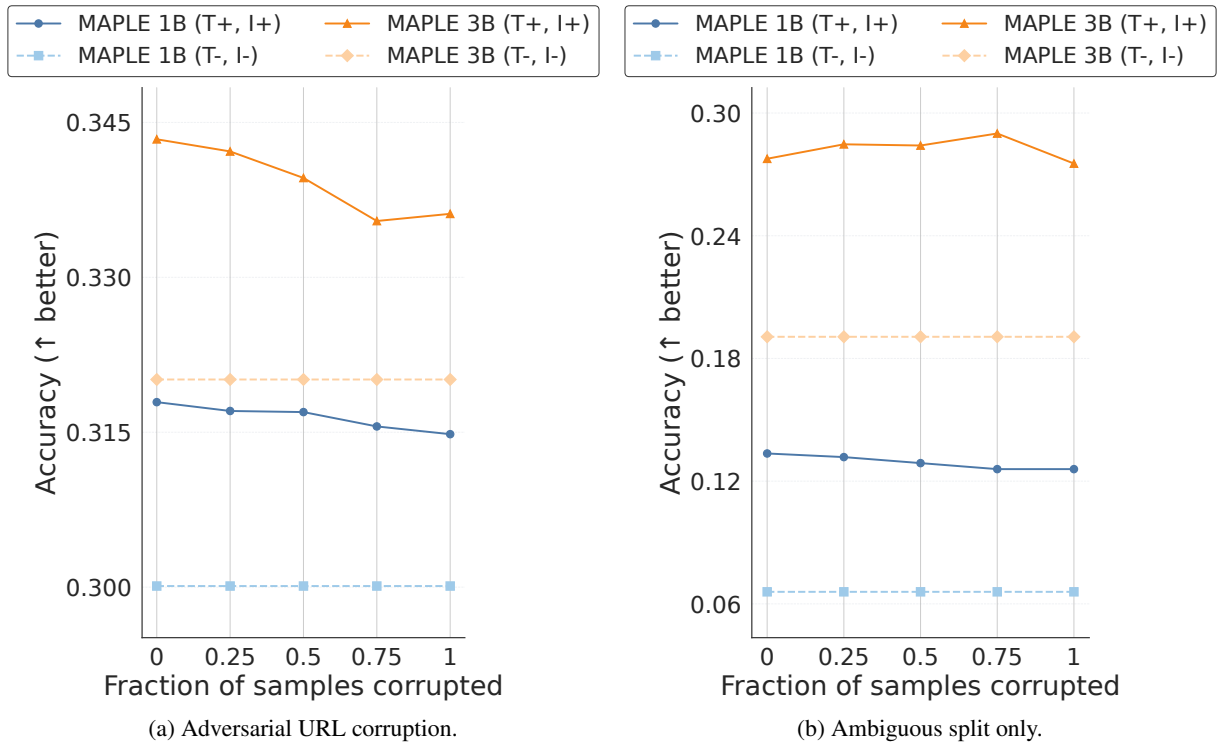


Figure 26. **Adversarial downstream evaluation.** Pretrained closed-form scoring is evaluated under metadata corruption on the final gold LOCALNEWSQA split. Figure a reports full-set target accuracy, and Figure b reports the ambiguous split only. Increasing metadata corruption reduces metadata-conditioned inference, while metadata-free inference is largely unchanged. The narrower URL-country-mask variant remains closer to the clean run than to the metadata-free control, indicating that country-code cues matter but broader URL structure also contributes.