Meta-Pretraining for Zero-Shot Cross-Lingual Named Entity Recognition in Low-Resource Philippine Languages

David Demitri Africa* **Suchir Salhan** **Yuval Weiss**
Paula Buttery **Richard Diehl Martinez**
University of Cambridge

Abstract

Named-entity recognition (NER) in low-resource languages is usually tackled by fine-tuning very large multilingual LMs, an option that is often infeasible in memory- or latency-constrained settings. We ask whether small decoder LMs can be pretrained so that they adapt quickly and transfer zero-shot to languages unseen during pretraining. To this end we replace part of the autoregressive objective with first-order model-agnostic meta-learning (MAML). Tagalog and Cebuano are typologically similar yet structurally different in their actor/non-actor voice systems, and hence serve as a challenging test-bed. Across four model sizes (11 M – 570 M) MAML lifts zero-shot micro-F_1 by 2–6 pp under head-only tuning and 1–3 pp after full tuning, while cutting convergence time by up to 8%. Gains are largest for single-token person entities that co-occur with Tagalog case particles *si*/*ni*, highlighting the importance of surface anchors.


davidafrica/pico-maml


DavidDemitriAfrica/pico-maml-train

1 Introduction

Named-entity recognition (NER) locates and categorises Persons (PER), Organisations (ORG) and Locations (LOC) in unstructured text (Chinchor and Robinson, 1997). It is used in a variety of important domains such as healthcare (Kundeti et al., 2016; Polignano et al., 2021; Shafqat et al., 2022) and law (Leitner et al., 2019; Au et al., 2022; Naik et al., 2023), yet progress remains concentrated in a handful of well-resourced languages. Cross-lingual named-entity recognition is therefore important to better serve underserved communities, yet recent advancements remain unevenly distributed since NER performance in many languages remains poor due to limited training resources.

A key challenge is that entity boundaries and categories are not universal: languages differ in their morphosyntactic cues, word order, and orthographic conventions. Models trained primarily on Indo-European data thus fail to generalize reliably to underrepresented settings. In this paper, we address this problem through **meta-pretraining**: shaping language model initializations to adapt rapidly to new linguistic conditions. Unlike standard pretraining, which minimizes average loss over a static corpus, episodic meta-pretraining (e.g. via MAML; Finn et al. 2017) explicitly optimizes for fast transfer. For low-resource NER, this offers two potential benefits: (i) rapid adaptation to languages with typologically distinct cues (e.g. case particles, voice systems, code-switching), and (ii) stronger zero-shot prototypes for common entity types, even without in-language exposure. While meta-learning has been explored for classification tasks in English or cross-lingually at BERT scale (Wu et al., 2020; Li et al., 2020; de Lichy et al., 2021), its efficacy for small decoder LMs and morphologically rich languages is underexplored.

As a case study, we focus on NER in Tagalog and Cebuano, the two most widely spoken Philippine languages (Miranda, 2023). Typologically, both languages combine Austronesian features such as voice alternations, case particles, and reduplication with pervasive borrowing and code-switching (Figure 8; Table 1). These languages stress-test whether meta-pretraining can yield more adaptable NER representations than vanilla pretraining alone. We ask the following research questions:

RQ1 Efficacy. How much does first-order MAML improve zero-shot NER on Tagalog and Cebuano relative to vanilla autoregressive pretraining?

RQ2 What transfers? Which entity classes, mor-

*Corresponding Author:
david.demitri.africa@gmail.com

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 106–127
November 8-9, 2025

Typological Feature	Tagalog	Cebuano
Voice system	✓ Four-way	✓ Reduced two-way
Case marking	✓ Obligatory	✗ Often dropped
Borrowing / code-switch	✓ High density	✗ More conservative
Morphological richness	✓ Productive affixation	✓ Regular affixation
Word order flexibility	✓	✓
Pronominal systems	✓ Rich clitic pronouns	✓ Similar
Reduplication	✓ Common	✓ Widespread
Orthography variation	✓ Multiple conventions	✗ Multiple conventions
Pivot marking	✓ Consistently overt	✓ Overt but less consistent

Table 1: A selection of Typological Features of Tagalog and Cebuano relevant for NER. ✓ indicates strong presence, ✗ indicates reduced/less overt presence in each language. We highlight **high divergence** features, **moderate divergence** and **similar** features compared to Indo-European Languages, motivating these languages as a case-study for low-resourced NER. We provide a more detailed comparison along with an illustrative gloss in Appendix A.

phological cues, and lexical patterns (especially those tied to Tagalog/Cebuano typology) explain the observed gains or failures?

We answer these questions by systematically comparing first-order MAML and vanilla pretraining on LLaMa-style Pico Decoders across scales, analyzing both downstream performance and representation dynamics (Diehl Martinez, 2025; Martinez et al., 2025). This allows us to investigate:

RQ3 How does the effect of meta-pretraining vary with model size? Are benefits stronger at small scales, or do they persist as capacity increases?

1.1 Contributions.

We provide the following contributions:

- A systematic evaluation of meta-pretrained small decoder LMs for zero-shot NER in Tagalog and Cebuano, comparing against strong vanilla pretraining baselines across four model scales.
- Quantitative and qualitative evidence that MAML-based meta-pretraining produces sharper single-token entity prototypes, improving zero-shot NER, especially for person entities and Tagalog's particle-rich syntax.
- An analysis of failure modes and learning dynamics, showing the capacity-dependent nature of meta-learning gains and the tradeoff between prototype sharpening and contextual generalization.

2 Method

2.1 Motivation

Why these two languages? Tagalog and Cebuano are used every day by well over 100 million people. However, they occupy only a small fraction of the web text that current language models are pretrained on, which makes them both socially important and under-served by existing NLP tools (Miranda, 2023). Linguistically, these languages also offer complementary typological challenges for NER, which we summarise in Figure 1. Tagalog and Cebuano combine Austronesian voice systems, case particles, reduplication, and discourse-driven topic marking in ways that are rare in widely studied NLP benchmarks. In particular, Tagalog offers more overt morphosyntactic cues than Cebuano: it retains a four-way actor/non-actor voice paradigm, while Cebuano reduces this to two (Tanangkingsing, 2011) and marks syntactic roles with case particles (*si/ni/ang/ng/sa*). These languages offer a test bed for multilingual NER models that must generalize beyond Indo-European NER cues – where entities are typically identifiable through fixed word order and stable orthography– to handle the interaction of morphological marking, argument interaction and code-switching. Tagalog contains more Spanish loans and code-switching into English, while Cebuano maintains a more conservative Austronesian lexicon (Bautista, 2004; Baklanova, 2019). We provide a more detailed comparison of Tagalog and Cebuano typological features in Table 3.

Why Meta-learning? Being underrepresented in natural language processing (NLP) corpora (Cajote et al., 2024; Quakenbush, 2005; Dita et al., 2009; Bandarkar et al., 2024), Philippine language datasets suffer from size and quality issues. In low-resource settings, where pretraining data is scarce or absent, it is important to ask the question: will a given checkpoint finetune or transfer rapidly when exposed to a novel language (such as in deployment)?

Meta-learning addresses this by shaping initializations for quick adaptation. Model-Agnostic Meta-Learning (MAML) optimizes an LM backbone so that a few gradient steps yield high performance on a new task (Finn et al., 2017). We ask whether such an initialization, learned entirely without Tagalog/Cebuano exposure, can transfer to these languages' distinct morphological and lexical

cues for NER. Our working hypothesis is that a pretraining routine that is itself optimized for rapid adaptation will induce representations that generalize more readily across languages. Prior NLP studies have tested this mostly on English or on "BERT-scale" encoder models (Wu et al., 2020; Ma et al., 2022; Li et al., 2020; de Lichy et al., 2021); we explore whether episodic meta-pretraining of small decoder LMs, without any exposure to Tagalog or Cebuano, can still yield zero-shot gains for NER. We do not evaluate a multilingual language-model baseline, as our objective is to isolate the effect of episodic meta-pretraining under a matched corpus and schedule; training a competitive multilingual baseline would require different data and budgets, confounding a like-for-like comparison.

Our working hypothesis is that a pretraining routine that is itself optimized for rapid adaptation will induce representations that generalize more readily across languages, so that a model exposed only to high-resource sources can still zero-shot transfer to typologically distant, low-resource targets.

2.2 Architecture

We build upon the PICO decoder stack (Diehl Martinez, 2025), a LLaMa-style causal Transformer implemented in PyTorch. Four capacity tiers (**tiny** (11 M), **small** (65 M), **medium** (181 M) and **large** (570 M)) share all hyper-parameters except hidden width $d \in \{96, 384, 768, 1536\}$. Each model comprises L=12 RMS-normalised decoder blocks (Zhang and Sennrich, 2019) with grouped-query self-attention (Ainslie et al., 2023), RoPE positions (Su et al., 2024) and SwiGLU feed-forwards (Shazeer, 2020) that expand to $4d$.

2.3 Hybrid pretraining objective

Training alternates between two outer-loop updates:

1. **Autoregressive LM step.** Standard next-token prediction on a pre-tokenized version of Dolma (Soldaini et al., 2024) released by the Pico library (Diehl Martinez, 2025).

2. **First-order MAML episode.** A 32-way, 4-shot Subset-Masked LM Task (SMLMT; Bansal et al., 2020) is sampled, where the model predicts a masked token from the corpus on the fly. The inner loop finetunes a lightweight MLP head for ten SGD steps ($\alpha = 10^{-3}$) and the outer loop back-propagates the query loss through the frozen backbone.

The branch decision is a Bernoulli draw with probability $\rho = 0.5$, synchronised across four A100-80 GB GPUs. The pseudocode for both can be found in Appendix C.

2.4 Optimisation and monitoring

We run 6,000 outer updates with AdamW ($\eta_{\text{peak}} = 3{\times}10^{-4}$, 2.5 k warm-up, cosine decay), accumulating eight micro-batches of 256 sequences to reach an effective batch of 2048 sequences (1024 for **tiny**). Every 100 steps we log: Paloma perplexity (Magnusson et al., 2024), singular-value spectra of three attention and three feed-forward weight matrices, from which we compute proportional effective rank (PER; Diehl Martinez et al., 2024), and support and query accuracy within MAML episodes.

2.5 Finetuning on High-Resourced Languages

We deliberately choose high-resource languages as the finetuning sources because, in realistic deployments, these are the languages for which sizable, high-quality NER data already exists. They therefore form the most natural setting for cross-lingual transfer into low-resource settings.

After pretraining we attach an untrained linear conditional random field head (Lafferty et al., 2001), which is a well-known method used often for NER (Bundschus et al., 2008; Ma and Hovy, 2016). We finetune on a high-resource language (Danish, English, Croatian, Portuguese, Slovak, Serbian, Swedish, Chinese, Chinese-Simplified, and a mixture of all languages) before zero-shot evaluation on Tagalog (`tl_trg`, `tl_ugnayan`) and Cebuano (`ceb_gja`) from Universal NER v1 (Mayhew et al., 2024). Results are later broken down by finetuning language. Further, two finetuning regimes are compared: head-only, where the transformer is frozen and only the classifier learns, and full, where all parameters are freed to update.

Finetuning uses AdamW ($3{\times}10^{-5}$) for up to ten epochs with early stopping on development F_1. We report micro-F_1, with full details in Appendix D.

2.6 Baselines

For each capacity tier we also evaluate a "vanilla" Pico model (no MAML, pure autoregressive loss) under identical data, schedule and compute. Pretraining results can be found in Appendix E with model configuration details in Appendix F. A more detailed discussion of pretraining results and overall methodology can be found in Africa et al. (2025).

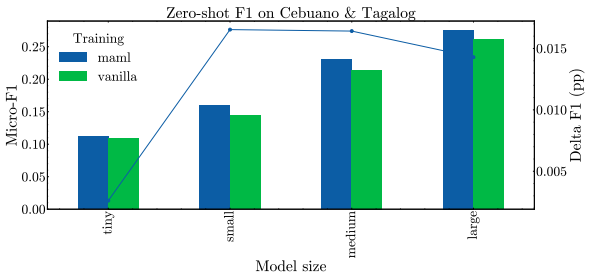


Figure 1: **Scale curve.** Zero-shot Micro-F_1 on Cebuano & Tagalog versus parameter count. Bars compare PICO-MAML (blue) to vanilla pretraining (green); the overlaid line shows the relative gain of MAML (Delta F1, right axis). Meta-pretraining helps at every scale, but the relative lift shrinks from +38 % (11 M) to +6 % (570 M), revealing a capacity threshold below which the inner loop cannot extract reusable features.

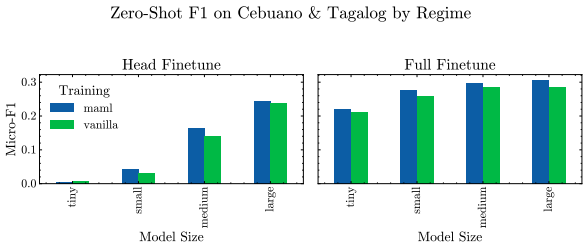


Figure 2: **Impact of finetuning regime.** Head-only tuning (left) magnifies the meta-learning advantage up to +2.5 pp at 570 M, likely because the backbone must already encode entity cues. Full tuning (right) reduces but does not erase the gap, suggesting that MAML primarily accelerates convergence rather than acting as a regulariser.

3 Zero-Shot Transfer Results

Zero-shot evaluation. Unless stated otherwise, all scores are obtained without seeing any Tagalog/Cebuano data during finetune, relying solely on the UNER test sets (§ 2.4).

Figure 1 shows that PICO-MAML improves Cebuano/Tagalog micro-F_1 at every parameter budget. The relative lift is largest for moderate sizes and tapers with scale (+6% at 570M). These results indicate that adding a single outer-loop meta-update per batch yields a cross-lingual prior not captured by vanilla pretraining under our setup.

Comparison of head-only tuning and full tuning. Decomposing by finetuning regime (Fig. 2), MAML yields 1–2 pp gains when only the CRF head is trained, implying that the frozen weights already embeds better entity cues. Full tuning narrows the gap to 0.5–1.3 pp, indicating that the lift persists even when the optimiser is free to overwrite the initialisation.

Further, results indicate that the benefit provided by the meta-objective is scale-dependent. For the 11 M (**tiny**) model, MAML moves the overall score by < 1 pp and yields no gain under head-only tuning. From 65 M parameters upward the benefit becomes clearer with larger head-only lifts, suggesting a threshold at which meta-gradients can provide reusable entity features without crowding out the LM signal.

Sensitivity to finetuning language. Figure 3 profiles performance after adapting on nine high-resource languages. Eight of nine languages exhibit positive deltas; the largest relative lifts occur for Slovak (+18 %) and Croatian (+13 %). Gain in Slovak might be due to fixed case endings that consistently bracket entity names, providing a clear surface boundary signal for the model (similar in function to Tagalog's case particles but realised morphologically rather than syntactically.) The sole regression (–2 pp on Simplified Chinese) is most likely due to a known issue in poor cross-script transfer to Chinese, but it may also be due to subword sparsity in the shared vocabulary rather than a failure of the meta-objective. (Mayhew et al., 2024).

Overall, MAML appears to teach the model to exploit shallow lexical anchors (particles, affixes) that generalise well across Indo-European languages while still transferring to more typologically distant Austronesian targets. To better understand the mechanisms underlying these gains, we conduct a focused qualitative analysis on a representative configuration.

4 Analysis of MAML Pretrained Models

In order to analyze the learning process, rather than just the last checkpoint, we focus our qualitative study on a MEDIUM-sized model (181 M parameters) finetuned in a head-only regime on Slovak (sk_snk), finetuning on all 61 checkpoints from step 0 of pretraining to step 6000. We restrict our analysis to this slice because while finetuning 9760 (2 pretraining regimes x 2 finetuning regimes x 4 model sizes x 10 finetuning languages x 61 checkpoints) models would be prohibitively expensive, this configuration at least offers a reasonable signal-to-cost trade-off. This is for a few reasons: (i) the medium tier is the smallest model that still exhibits a clear 2–3 pp head-only lift (Figure 1) yet is three-

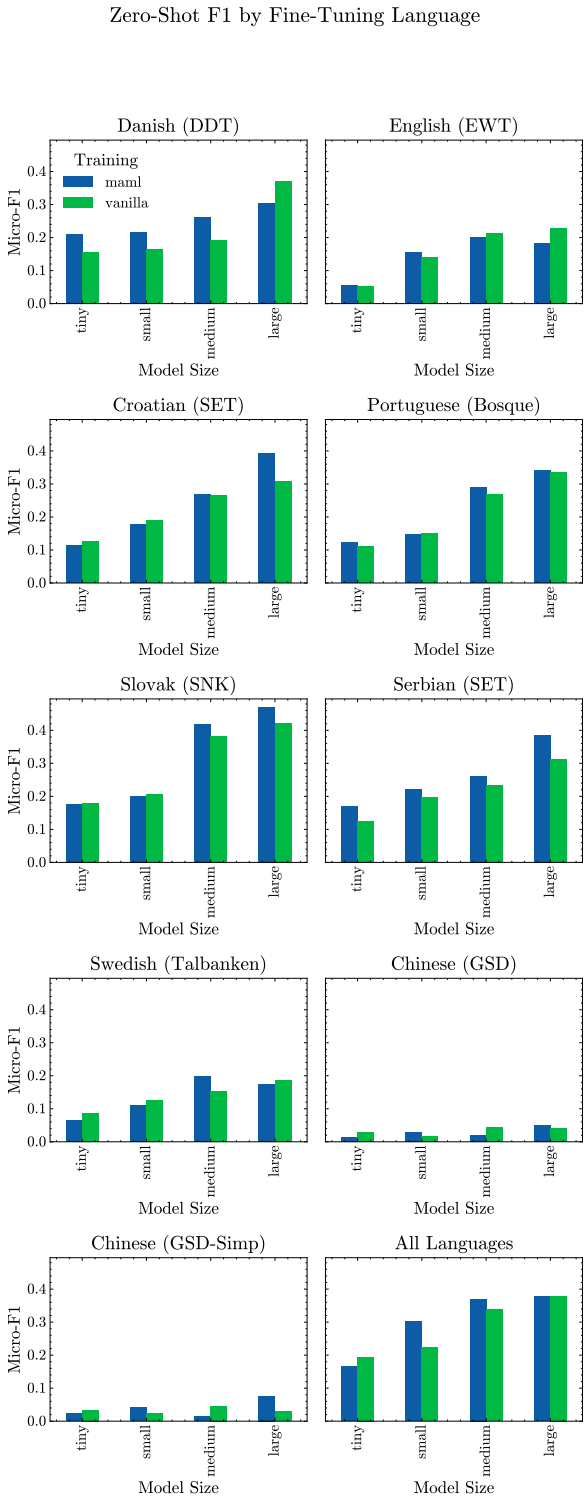


Figure 3: **Sensitivity to finetuning language.** Grid of zero-shot F_1 curves after adapting on nine high-resource languages plus an *All*-languages mixture. Eight of nine languages show positive deltas; the largest relative gains occur for Slovak and Croatian, while Simplified Chinese is the lone outlier (–2 pp). This pattern indicates that the meta-objective encourages reliance on surface affixes and particles that generalise well across Indo-European sources yet still transfer to Austronesian targets.

times cheaper to run than the 570 M variant, (ii) Slovak delivers one of the largest relative gains without vocabulary sparsity issues and, as a Slavic language, should produce transfer errors that differ sharply from those in Tagalog and Cebuano, and (iii) freezing the backbone during head-only finetuning ensures that any performance delta must stem from representations learned during meta-pretraining rather than from subsequent weight updates. In the next subsection, we inspect how pretraining affects finetuning performance across checkpoints.

4.1 Checkpoint Analysis

Does the head-only learner actually learn? Figure 4 overlays the complete finetuning trajectories for every Slovak head-only run (61 checkpoints, `maml_s0000`–`maml_s6000`). Viridis traces show the individual runs (getting darker the later the model checkpoint was taken), while the bold line and ribbon denote the median and inter-quartile range (IQR). The train-loss fan collapses to its asymptote within the first ≈ 800 steps and stays flat thereafter; in parallel the evaluation F_1 rises smoothly to 0.14 and plateaus with a narrow ± 0.01 IQR. Crucially, no run diverges or oscillates, confirming that freezing the backbone and training only a linear chain CRF head is both stable and something is learned. This satisfies the prerequisite for using the configuration as a clean test-bed: any downstream difference between MAML and vanilla is likely to stem from the initial representations, not from optimisation quirks or training instabilities.

Does meta-pretraining yield transfer-relevant representations? The checkpoint sweep in Figure 5 confirms the other prerequisite for this qualitative analysis: that meta-pretraining produces representations which become increasingly helpful for zero-shot transfer. First, the top panel shows that, regardless of which MAML snapshot we freeze, the linear chain CRF head always converges to essentially the same narrow band of train loss (0.10-0.15); optimisation is therefore stable and predictable, satisfying our first prerequisite. More importantly, the bottom panel reveals a very different story for cross-lingual evaluation: while Slovak dev F_1 plateaus early (by around step 1k), Tagalog and Cebuano F_1 continue to climb for another four thousand meta-updates, ending 0.15 and 0.12 points higher than at the initial checkpoint. In

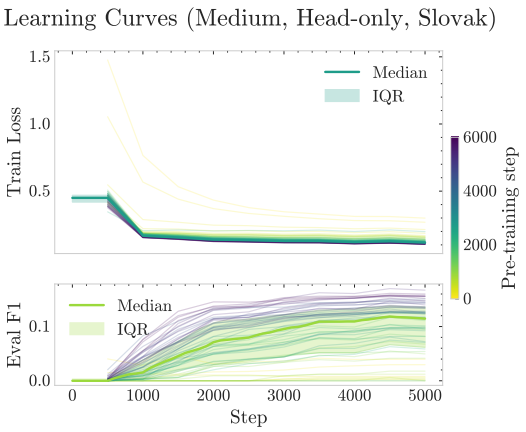


Figure 4: **Learning curves for the Slovak head-only setting.** Top: train loss; bottom: eval micro-F_1. Faint green lines = all individual checkpoints; bold line = median; shaded band = 25–75 % IQR. Both metrics converge monotonically and remain tightly bunched, indicating a stable optimisation surface for the linear head.

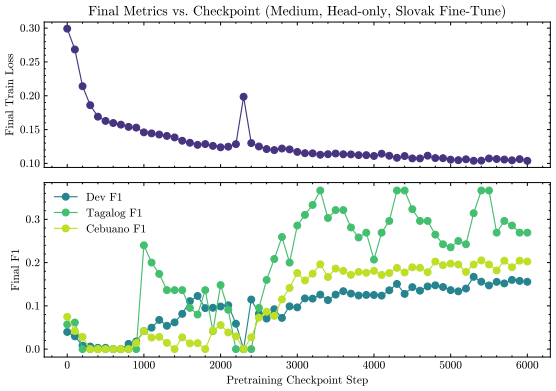


Figure 5: **Final metrics vs. pretraining checkpoint** for the MEDIUM MAML backbone frozen during head-only finetuning on Slovak. Top: final train loss of the CRF head, every run converges to the same narrow range. Bottom: final micro-F_1 on Slovak dev (blue), Tagalog (green) and Cebuano (yellow). Although in-language performance saturates early, cross-lingual F_1 keeps improving up to step 6000, indicating that later meta-updates learn representations useful specifically for zero-shot transfer.

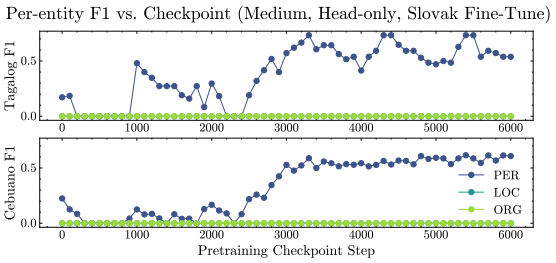


Figure 6: **Per-entity F_1 across MAML checkpoints.** PER (dark viridis) improves steadily with more meta-steps; LOC and ORG curves remain at chance level, indicating that the frozen backbone provides transferable features for single-token personal names but little for multi-token locations or organisations. Tagalog benefits earlier than Cebuano, consistent with its obligatory case particles.

other words, additional MAML steps learn features that are invisible to the in-language dev set yet directly benefit unseen Austronesian targets. Tagalog improves earlier and peaks higher than Cebuano, hinting that the meta-objective is capturing surface cues (e.g. case particles) that are more diagnostic in Tagalog. Taken together with the "fan" plot of learning curves, the sweep demonstrates that meta-pretraining yields encoder states that are both optimisation-friendly and transfer-relevant, justifying the focus on this snapshot for deeper qualitative inspection. As such, we deepen the analysis in the next subsection by inspecting the behavior of our models on the level of the NER tags predicted.

4.2 Tag-level Analysis

Per-tag behaviour. Figure 6 reports per-entity F_1 obtained after head-only finetuning the Slovak CRF head on each MAML checkpoint. PER climbs to 0.6-0.7 while LOC and ORG remain at zero. This is not a case of the classifier "over-fitting" in the usual sense—i.e. collapsing to always predicting a single label. A linear-chain CRF is free to emit any BIO tag at any position; if it were truly degenerate we would see train loss stagnate near the log-uniform baseline and the PER curve itself would also be flat. Instead, train loss converges to the same narrow band for every checkpoint (Fig.4) and PER performance tracks the amount of meta-pretraining, so the head is learning a genuine decision boundary. It simply has informative features for people but none for locations or organisations.

Observed imbalance and potential causes. First, the Slovak finetune set is intrinsically person-heavy. As Table 4 shows, PER spans outnumber LOC by roughly 8:1 and ORG by 15:1. Under head-only training, every gradient step passes through the frozen encoder unchanged and the CRF receives thousands of positive updates for persons but only a few hundred for the other classes. This likely leads to only the PER decision boundary sharpening. Second, 87.6% of Slovak person mentions are single tokens compared with 75.1 % for locations and 56.9% for organisations. A single-

token span can be captured by one weight vector, whereas multi-word spans require the head to model boundaries and label transitions—a capacity it simply does not have when the encoder cannot adapt. Third, Tagalog still offers a comparatively reliable surface cue. The case particles *si* and *ni* precede roughly 11% of gold PER spans, almost double the 5–6 % rate observed in Cebuano (Table 5). The earlier lift and higher ceiling of the Tagalog PER curve are therefore consistent with the backbone having learned to map the pattern "particle + token" to the PER label, a cue that is informative in Tagalog but is sparser in Cebuano. Finally, cross-lingual lexical overlap is likely higher for personal names, many of which (e.g. *Obama*, *Manuel*) appear verbatim in English corpora used during pretraining; locations and organisations, by contrast, are often translated or abbreviated. All four factors act in the same direction, favouring PER. Disentangling their individual contributions would require targeted ablations (particle masking, balanced resampling, controlled name substitution, etc.) which we leave for future work. In the next subsection, we assess behaviors on the level of words and tokens to relate NER performance to the low-resource languages being transferred to.

4.3 Word-level Analysis

Figures 7a–7d visualise the checkpoint-by-checkpoint evolution of token-level confidence (p(correct tag)) for the ten most frequent surface words in each evaluation set. Entities and non-entities are split so the dynamic range is not drowned out by O tokens. Two qualitative patterns emerge.

Fast confidence in frequent tokens. Non-entity function words such as *ng, ang, sa* in Tagalog and the Cebuano clitic *-ng* start with high confidence and barely budge after the first 200 meta-updates (Fig. 7b, 7d). As these tokens dominate the language-model loss, autoregressive training achieves a high confidence in them early and MAML has little head-room to improve over checkpoints.

Monotonic gains for high-overlap proper names. In the Tagalog set, international names (*City, Maynila, Maria*) and locations transliterated from English (*Pasay*) become steadily brighter (lower loss) until about step 3000 (Fig. 7a). Similar behaviour appears for *Maria, Cebu, Mary* in Cebuano (Fig. 7c). These words either appear verbatim in the English Dolma corpus or share sub-tokens (`Ma_`, `Ceb_`) with it, so the meta-objective can reuse prototypes that happen to be used by the Austronesian targets. The timing matches the checkpoint-sweep (Fig. 5): cross-lingual F_1 continues to climb long after Slovak dev has saturated likely because the back-bone is still lowering loss on these anchor words. We illustrate these mechanisms further in two case studies in Appendix B.

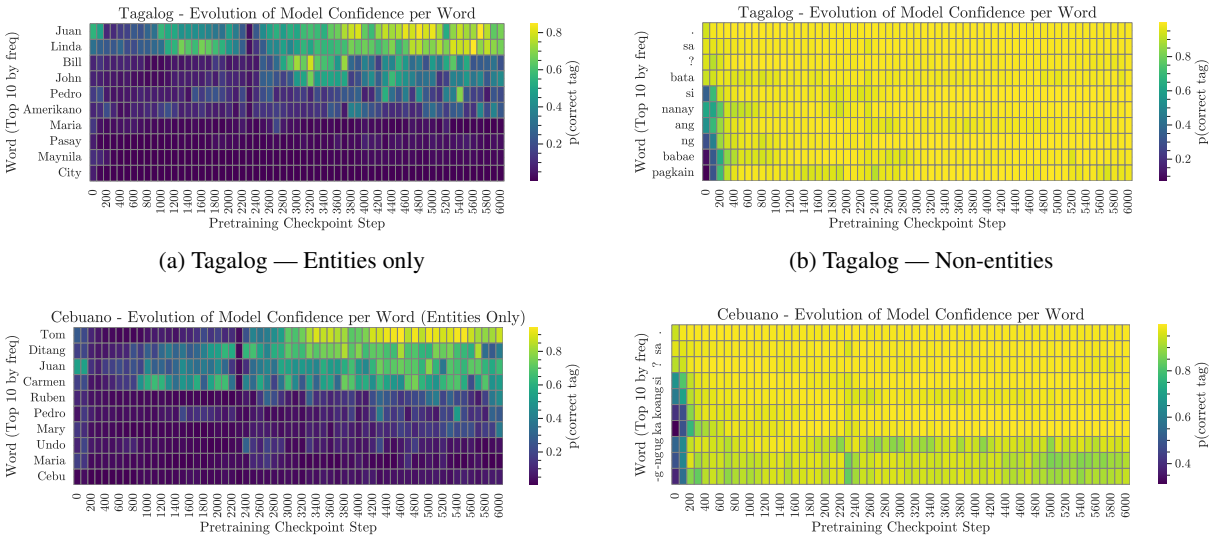
Size	Regime	Δt_{90}	ΔAUC	Δslope
`large`	full	-111.1	-0.004	0.0e-05
	head	-55.6	-0.012	1.0e-05
`medium`	full	0.0	-0.005	0.0e-05
	head	55.6	-0.011	0.0e-05
`small`	full	0.0	0.003	0.0e-05
	head	-55.6	0.003	-0.0e-05
`tiny`	full	-111.1	0.004	-0.0e-05
	head	55.6	-0.023	5.0e-05

Table 2: Finetuning convergence speed metrics Δ (MAML-Vanilla) averaged over nine in-language tasks. The largest and smallest models enjoy the most pronounced speed-ups from full MAML meta-initialization, while medium and tiny models show negligible Δt_{90} under full-model tuning. Under head-only tuning, large and small decoders still benefit modestly, whereas medium and tiny decoders actually slow down. Across all settings, slope remains near zero, indicating that meta-training primarily accelerates mid-to-late convergence rather than the very first gradient steps.

5 Finetuning Speed of Meta-Pretraining

Finally, we assess finetuning speed using convergence time (measuring time to achieve 90% of final loss t_{90}), normalized area under the loss curve (measuring aggregate convergence behavior over the curve), and initial slope (measuring the initial speed of learning in the first few steps), as seen in Table 2. Across nine in-language tasks, full-model finetuning shows the clearest acceleration for the largest and smallest models: MAML cuts t_{90} by roughly 8% ($\approx$ 111 steps) and modestly reduces loss AUC. Medium and small models show negligible or inconsistent speed-ups under full tuning, suggesting that the effect depends strongly on model capacity. In head-only tuning, large and small models again benefit slightly, while medium and tiny models slow down, likely due to underpowered or collapsed meta-dynamics.

Initial slopes remain effectively unchanged across all settings, indicating that MAML does not alter the very first gradient steps but instead reorga-



(a) Tagalog — Entities only

(b) Tagalog — Non-entities

(c) Cebuano — Entities only

(d) Cebuano — Non-entities

Figure 7: **Evolution of token-level confidence** (p(correct tag)) across pretraining checkpoints. Top row: Tagalog; bottom row: Cebuano. Left: entities only. Right: non-entities.

nizes the loss landscape to make mid- to late-stage convergence more efficient. These results align with earlier findings that MAML's main benefit lies in providing sharper, more reusable token-level features for high-capacity backbones, with limited or negative effects when capacity is insufficient to retain both language modeling and episodic priors.

6 Related Work

NER in Filipino, Tagalog, and Cebuano. NER for Philippine languages remains underexplored, with most work focusing on resource construction rather than cross-lingual modeling. Recent corpora include TLUnified-NER (Miranda, 2023), TF-NERD (Ramos and Vergara, 2023), CebuaNER (Pilar et al., 2023), and UniversalNER (Mayhew et al., 2024). Modeling efforts in this area primarily use NER-specific systems (Sagum and Sagum, 2025; Eboña et al., 2013; Dela Cruz et al., 2018) incorporating a simpler backbone such as a support vector machine (Castillo et al., 2013) or an LSTM (Chan et al., 2023). Most recently, FilBench (Miranda et al., 2025) and Batayan (Montalan et al., 2025) support Filipino evaluation on NLP tasks for LLMs.

Meta-learning for Pretraining. Although most work applies meta-learning at fine-tuning time, a growing line of research embeds meta-objectives directly into pretraining. (Raghu et al., 2021) showed that framing parameter-efficient adapter learning as a bilevel problem yields representations that fine-tune more effectively than standard PEFT. (Hou et al., 2022) extend this to full transformers. (Miranda et al., 2023) argue that explicit MAML objectives can outperform fixed pretraining on highly diverse task distributions. (Ke et al., 2021) integrate a MAML-style inner loop into a multi-criteria Chinese Word Segmentation pretraining task.

7 Conclusion

This paper shows that MAML-based meta-pretraining, even when applied to small decoder-only language models, can meaningfully improve zero-shot transfer to low-resource languages, as demonstrated on Tagalog and Cebuano NER. The gains are most pronounced for person entities and head-only finetuning, and scale best with larger model capacities. Our qualitative and word-level analyses reveal that the mechanism of improvement centers on the sharpening of lexical prototypes and better anchoring to surface cues like Tagalog case particles. Hence, we do not expect these improvements to fully generalize to multi-token or highly contextual entity types.

These findings suggest that meta-learning can provide a principled route to more adaptable small models, but also highlight key limitations: the benefits are capacity- and task-dependent, and the current approach struggles with richer entity structures. Future work should explore alternative meta-learning objectives, extend to more diverse tasks

and languages, and investigate the dynamics of prototype formation in even lower-resource settings.

Limitations

The gains are most pronounced for person entities and head-only finetuning, and scale best with larger model capacities. All training runs stop at exactly six thousand outer steps, a horizon that may be too short for the largest model, so the conclusions derived only cover a fraction of the training budget a corporate setup might have. A more diverse and multilingual corpus may alter both quantitative and qualitative conclusions, and varying languages in the meta-task is a natural way to extend this work. Qualitative analysis was conducted on a single configuration and single seed due to cost and GPU constraints. Qualitative analysis was conducted by a native Tagalog speaker with a register typical of Manila, and a wide variety of perspectives would improve the robustness of the analysis. Finally (and most naturally), our focus on only two Austronesian languages controls for certain lexical and syntactic divergences but limits the generality of the typological conclusions; extending to a broader set of Philippine and Malayo-Polynesian languages is a natural next step.

Acknowledgments

This work was supported by a grant from the Accelerate Programme for Scientific Discovery, made possible by a donation from Schmidt Futures. David Demitri Africa is supported by the Cambridge Trust and the Jardine Foundation. Suchir Salhan is supported by Cambridge University Press & Assessment. Richard Diehl Martinez is supported by the Gates Cambridge Trust (grant OPP1144 from the Bill & Melinda Gates Foundation). It was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council.

References

David Demitri Africa, Yuval Weiss, Paula Buttery, and Richard Diehl Martinez. 2025. Learning dynamics of meta-learning in small model pretraining. *Preprint*, arXiv:2508.02189.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901.

Ting Wai Terence Au, Ingemar J Cox, and Vasileios Lampos. 2022. E-ner–an annotated named entity recognition corpus of legal text. *arXiv preprint arXiv:2212.09306*.

Ekaterina Baklanova. 2019. The impact of spanish and english hybrids on contemporary tagalog.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534, Online. Association for Computational Linguistics.

Maria Lourdes S Bautista. 2004. Tagalog-english code switching as a mode of discourse. *Asia Pacific Education Review*, 5(2):226–233.

Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. 2008. Extraction of semantic biomedical relations from text using conditional random fields. *BMC bioinformatics*, 9(1):207.

Rhandley D Cajote, Rowena Cristina L Guevara, Michael Gringo Angelo R Bayona, and Crisron Rudolf G Lucas. 2024. Philippine Languages Database: A Multilingual Speech Corpora for Developing Systems for Philippine Spoken Languages. *LREC-COLING 2024*, page 264.

Jonalyn M Castillo, Marck Augustus L Mateo, Antonio DC Paras, Ria A Sagum, and Vina Danica F Santos. 2013. Named entity recognition using support vector machine for filipino text documents. *International Journal of Future Computer and Communication*, 2(5):530.

Kyle Chan, Kaye Ann De Las Alas, Charles Orcena, Dan John Velasco, Qyle John San Juan, and Charibeth Cheng. 2023. Practical approaches for low-resource named entity recognition of filipino telecom-

munications domain. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 234–242.

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, volume 29, pages 1–21.

Cyprien de Lichy, Hadrien Glaude, and William Campbell. 2021. Meta-learning for few-shot named entity recognition. In *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*, pages 44–58.

Bern Maris Dela Cruz, Cyril Montalla, Allysa Manansala, Ramon Rodriguez, Manolito Octaviano, and Bernie S. Fabito. 2018. Named-entity recognition for disaster related filipino news articles. In *TENCON 2018 - 2018 IEEE Region 10 Conference*, pages 1633–1636.

Richard Diehl Martinez. 2025. Pico: A lightweight framework for studying language model learning dynamics.

Richard Diehl Martinez, Pietro Lesci, and Paula Buttery. 2024. Tending towards stability: Convergence challenges in small language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3275–3286, Miami, Florida, USA. Association for Computational Linguistics.

Shirley N Dita, Rachel Edita O Roxas, and Paul Inventado. 2009. Building online corpora of Philippine languages. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, pages 646–653. Waseda University.

Karen Mae L Eboña, Orlando S Llorca Jr, Genrev P Perez, Jhustine M Roldan, Iluminda Vivien R Domingo, and Ria A Sagum. 2013. Named-entity recognizer (ner) for filipino novel excerpts using maximum entropy approach. *Journal of Industrial and Intelligent Information Vol*, 1(1).

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Zejiang Hou, Julian Salazar, and George Polovets. 2022. Meta-learning the difference: Preparing large language models for efficient adaptation. *Transactions of the Association for Computational Linguistics*, 10:1249–1265.

Zhen Ke, Liang Shi, Songtao Sun, Erli Meng, Bin Wang, and Xipeng Qiu. 2021. Pre-training with meta learning for Chinese word segmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5514–5523, Online. Association for Computational Linguistics.

Srinivasa Rao Kundeti, J Vijayananda, Srikanth Mujjiga, and M Kalyan. 2016. Clinical named entity recognition: Challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1937–1945. IEEE.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *International conference on semantic systems*, pages 272–287. Springer.

Jing Li, Billy Chiu, Shanshan Feng, and Hao Wang. 2020. Few-shot named entity recognition via meta-learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(9):4245–4256.

Tingting Ma, Huiqiang Jiang, Qianhui Wu, Tiejun Zhao, and Chin-Yew Lin. 2022. Decomposed meta-learning for few-shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1584–1596.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Walsh, Yanai Elazar, Kyle Lo, and 1 others. 2024. Paloma: A benchmark for evaluating language model fit. *Advances in Neural Information Processing Systems*, 37:64338–64376.

Richard Diehl Martinez, David Demitri Africa, Yuval Weiss, Suchir Salhan, Ryan Daniels, and Paula Buttery. 2025. Pico: A modular framework for hypothesis-driven small language model research. *arXiv preprint arXiv:2509.16413*.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Brando Miranda, Patrick Yu, Saumya Goyal, Yu-Xiong Wang, and Sanmi Koyejo. 2023. Is pretraining truly better than meta-learning? *Preprint*, arXiv:2306.13841.

Lester James V. Miranda. 2023. Developing a named entity recognition dataset for Tagalog. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 13–20, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Lester James V. Miranda, Elyanah Aco, Conner Manuel, Jan Christian Blaise Cruz, and Joseph Marvin Imperial. 2025. Filbench: Can llms understand and generate filipino? *Preprint*, arXiv:2508.03523.

Jann Railey Montalan, Jimson Paulo Layacan, David Demitri Africa, Richell Isaiah S. Flores, Michael T. Lopez Ii, Theresa Denise Magsajo, Anjanette Cayabyab, and William Chandra Tjhi. 2025. Batayan: A Filipino NLP benchmark for evaluating large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31239–31273, Vienna, Austria. Association for Computational Linguistics.

Varsha Naik, Purvang Patel, and Rajeswari Kannan. 2023. Legal entity extraction: An experimental study of ner approach for legal documents. *International Journal of Advanced Computer Science and Applications*, 14(3).

Ma. Beatrice Emanuela Pilar, Dane Dedoroy, Ellyza Mari Papas, Mary Loise Buenaventura, Myron Darrel Montefalcon, Jay Rhald Padilla, Joseph Marvin Imperial, Mideth Abisado, and Lany Maceda. 2023. CebuaNER: A new baseline Cebuano named entity recognition model. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 792–800, Hong Kong, China. Association for Computational Linguistics.

Marco Polignano, Marco de Gemmis, Giovanni Semeraro, and 1 others. 2021. Comparing transformer-based ner approaches for analysing textual medical diagnoses. In *CLEF (Working Notes)*, pages 818–833.

J Stephen Quakenbush. 2005. Philippine linguistics from an SIL perspective: Trends and prospects. *Current issues in Philippine linguistics and anthropology: Parangal kay Lawrence A. Reid*, pages 3–27.

Aniruddh Raghu, Jonathan Lorraine, Simon Kornblith, Matthew McDermott, and David K Duvenaud. 2021. Meta-learning to improve pre-training. *Advances in Neural Information Processing Systems*, 34:23231–23244.

Robin Kamille Ramos and John Paul Vergara. 2023. Tf-nerd: Tagalog fine-grained named entity recognition dataset. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval*, pages 222–227.

Ria A. Sagum and Janelle Kyra A. Sagum. 2025. Parallel ensemble approach for named entity recognition in filipino text. In *Proceedings of the 2024 7th Artificial Intelligence and Cloud Computing Conference*, AICCC '24, page 409–413, New York, NY, USA. Association for Computing Machinery.

Sarah Shafqat, Hammad Majeed, Qaisar Javaid, and Hafiz Farooq Ahmad. 2022. Standard ner tagging scheme for big data healthcare analytics built on unified medical corpora. *Journal of Artificial Intelligence and Technology*, 2(4):152–157.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, and 1 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Michael Tanangkingsing. 2011. *A Functional Reference Grammar of Cebuano: A Discourse-Based Perspective*. Peter Lang, Berlin.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F Karlsson, Biqing Huang, and Chin-Yew Lin. 2020. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9274–9281.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

A NER-Relevant Typological Features of Cebuano and Tagalog

This extended table highlights how morphosyntactic and discourse-level differences between the two languages interact with the challenges of named entity recognition (NER). We lay out feature-by-feature contrasts to illustrate that even closely related Philippine languages present distinct hurdles for tasks like NER. The table emphasizes that while Tagalog offers overt morphosyntactic cues (e.g., case particles, topic marking), Cebuano relies more heavily on discourse inference, thereby requiring different modeling strategies for effective NER.

Typological Feature	**Tagalog**	**Cebuano**	**Challenge for NER**
Voice system	Four-way actor/non-actor voice paradigm	Reduced two-way system	Tagalog's rich voice alternations encode argument roles morphologically, complicating alignment of entities with semantic roles. Cebuano's reduced system lowers redundancy, making cues for role identification less explicit.
Case marking	Obligatory case particles (si, ni, ang, ng, sa)	Case particles often dropped or fused	Tagalog provides reliable morphosyntactic signals for entity boundaries/roles. Cebuano forces reliance on discourse, requiring coreference and contextual inference.
Lexical borrowing / code-switching	High density of Spanish loans and English code-switching	More conservative Austronesian lexicon	Tagalog NER must cope with OOV issues, language-mixing, and orthographic variation. Cebuano NER must handle morphologically complex Austronesian stems, underrepresented in multilingual embeddings.
Morphological richness	Productive affixation (focus, aspect, causatives)	Similarly rich, but slightly more regular	Surface forms for named entities may be inflected or derivationally complex, increasing sparsity for training data.
Word order flexibility	Relatively free (voice and particles constrain roles)	Even freer, especially without explicit case markers	Named entities may appear in non-canonical positions, reducing the utility of positional cues.
Pronominal systems	Rich system of clitic pronouns that attach to verbs or particles	Similar system but with different distributions	Entities can be referred to obliquely or dropped entirely; clitic attachment blurs tokenization boundaries, confusing NER pipelines.
Reduplication	Common for aspect, plurality, intensification	Widespread and productive	Reduplicated forms of named entities (nicknames, reduplicated roots) may not be recognized as related to the canonical form.
Orthography & variation	Spanish-influenced orthography, multiple spelling conventions	More phonologically consistent, but dialectal spelling variation persists	Orthographic inconsistency makes lexicon-based NER brittle, especially in noisy social media text.
Discourse prominence / topic marking	Ang-marked topic influences salience	Topic is often inferred from discourse, less explicit marking	Tagalog gives overt topic marking, aiding salience detection; Cebuano relies on pragmatics, requiring discourse-level modeling.

Table 3: Detailed typological contrasts between Tagalog and Cebuano and their implications for NER.

Tag.	Pumunta	si	**Maria**	sa	**Cebu**.
Gloss	go.PFV	NOM	Maria	OBL	Cebu
NER	O	O	B-PER	O	B-LOC
Ceb. (with marker)	Miadto	si	**Juan**	sa	**Sugbo**.
Gloss	go.PST	NOM	Juan	OBL	Cebu
NER	O	O	B-PER	O	B-LOC
Ceb. (zero-marked)	Miadto	**Juan**	sa	**Sugbo**.	
Gloss	go.PST	Juan	OBL	Cebu	
NER	O	B-PER	O	B-LOC	

Figure 8: Surface cues for named entities. Tagalog typically provides an overt personal article (*si*/*ni*) before names; Cebuano may show the same article, but zero-marked variants also occur in some registers/contexts, reducing overt anchors.

B Case Studies

To illustrate the mechanisms underlying MAML's improvements, we present two contrasting examples that demonstrate how meta-pretraining affects different types of linguistic patterns in Tagalog NER. We measure Δ log-prob as the change in surprisal ($-!\log p$) for the gold label between the vanilla and MAML model. A negative Δ means the model is more confident after MAML; a positive Δ means less confident.

Case 1: Prototype Amplification. Sentence: "Inahit ni John ang sarili niya." (Gloss: "John shaved himself.")

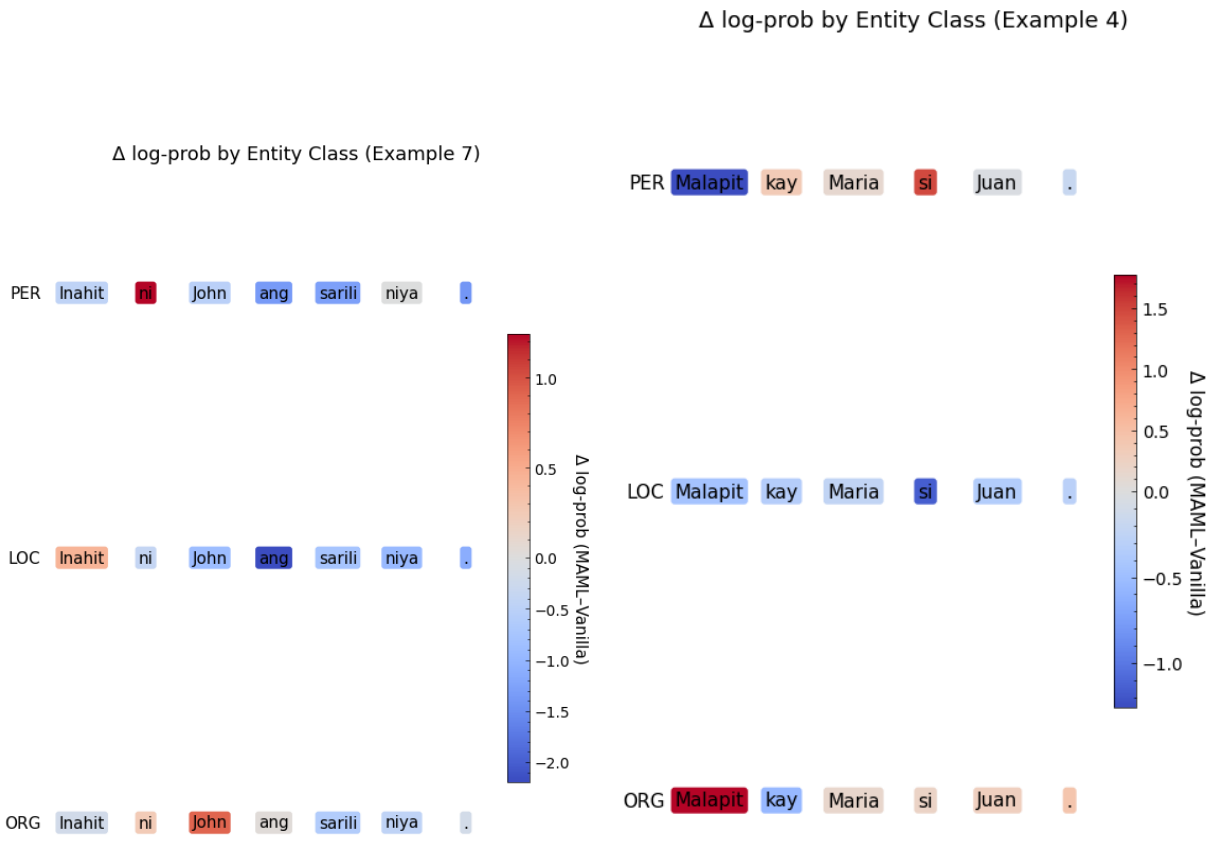
The first case study demonstrates how MAML strengthens recognition of cross-linguistically common proper names. In this example, MAML sharply reduces surprisal on "John," indicating stronger prototype activation.

We suspect improvement operates at two levels: (1) lexical level, in the sense that the token "John" becomes more strongly associated with person entities through meta-learning's emphasis on rapid adaptation to new entities, and (2) contextual level, in the sense that the *ni* + proper-name pattern gets reinforced as a reliable PER indicator during meta-training episodes.

Case 2: Contextual Suppression (Loss). Sentence: "Malapit kay Maria si Juan." (Gloss: "Juan is close to Maria.")

The second case study reveals MAML's limitations with complex multi-token constructions. Here, Δ is positive for key tokens, showing that MAML reduces confidence in the correct label. In "Malapit kay Maria si Juan" (Juan is close to Maria), both the locative adverb "Malapit" (close/near) and the oblique case marker "kay" show substantially decreased confidence for location labeling under MAML (combined decrease of approximately -3.3 log-probability points).

We suspect this occurs due to: (1) capacity constraints, in the sense that the frozen backbone has limited representational capacity, and strengthening PER features may crowd out LOC/ORG representations, and (2) training signal imbalance, in the sense that finetuning contained more person-like entities than complex locative expressions, biasing the learned representations toward single-token person recognition.



(a) Prototype Amplification.

(b) Contextual Suppression.

Figure 9: MAML's impact on (a) single-token prototype confidence and (b) multi-token contextual cue sensitivity.

C Pseudocode

Below is the pseudocode for the MAML and vanilla pretraining setup.

Distributed Subset Masked Language Modeling Tasks (SMLMT) Training

Algorithm 1 Distributed SMLMT Loop

// *Initialization:* same as Alg. 2, plus
initialize inner-optimizer SGD on head h_ϕ
step $\leftarrow 0$
for each sub_batch in dataloader **do**
    // gather across GPUs
    $X \leftarrow$ `fabric.all_gather`(sub_batch["input_ids"])
    // sync random branch decision
    $r \leftarrow \text{Uniform}(0, 1)$; $r \leftarrow$ `fabric.broadcast`(r)
    if $r < \rho$ **then**
        // *Meta-learning episode*
        $(S, Q), \text{labels}_S, \text{labels}_Q \leftarrow \text{mask\_tokens}(X)$
        $\phi_0 \leftarrow \phi$ ▷ snapshot head params
        for $t = 1$ **to** T_{inner} **do**
            $\ell_S \leftarrow \text{CE}(h_{\phi_{t-1}}(f_\theta(S)), \text{labels}_S)$
            $\phi_t \leftarrow \phi_{t-1} - \alpha \bullet \ell_S$ ▷ inner SGD
        end for
        $\ell_Q \leftarrow \text{CE}(h_{\phi_T}(f_\theta(Q)), \text{labels}_Q)$
        $\phi \leftarrow \phi_0$ ▷ restore head
        `fabric.backward`(ℓ_Q/accum_steps)
    else
        // *Standard AR*
        $X_{\text{in}}, Y \leftarrow X[:, :-1], X[:, 1:]$
        $\ell_{\text{AR}} \leftarrow \text{CE}(f_\theta(X_{\text{in}}), Y)$
        `fabric.backward`(ℓ_{AR}/accum_steps)
    end if
    // outer-step and logging
    if (step+1) % accum_steps == 0 **then**
        opt.step(); scheduler.step(); opt.zero_grad()
        // aggregate metrics across GPUs
        log_loss $\leftarrow$ `fabric.all_reduce`(ℓ)
        `fabric.log`(...)
        `fabric.barrier`()
    end if
    step + = 1
end for

Distributed Autoregressive (AR) Training

Algorithm 2 Distributed AR Loop

// Initialization (in `Trainer.__init__`*):*
Load configs; initialize Fabric, tokenizer, model f_θ
$(\text{model}, \text{opt}) \leftarrow$ `fabric.setup`$(f_\theta,\ \text{AdamW})$
dl $\leftarrow$ base dataloader; dl $\leftarrow$ `fabric.setup_dataloaders`(dl)
step $\leftarrow$ 0; zero gradients
for each sub_batch in dl **do**
    // Gather full batch across GPUs if needed:
    $X \leftarrow$ `fabric.all_gather`(sub_batch["input_ids"])
    $X_{\text{in}}, Y \leftarrow X[:, :-1],\ X[:, 1:]$
    // forward + loss
    $\ell \leftarrow \text{CE}\big(f_\theta(X_{\text{in}}),\ Y\big)$
    // backward (handles synchronization)
    `fabric.backward`$\big(\ell$/accum_steps$\big)$
    // outer-step when accumulated
    if (step+1) % accum_steps == 0 **then**
        opt.step(); scheduler.step(); opt.zero_grad()
        // optional barrier
        `fabric.barrier`()
    end if
    step + = 1
end for

C.1 Multi-GPU processing

Pico already uses Lightning-Fabric data parallelism but meta-learning introduces various demands that make multi-GPU processing complicated. A Bernoulli draw is done on one GPU and broadcast so all ranks choose the same objective. Support and query tensors are constructed on rank 0 then scattered, because per-rank random masks would destroy gradient equivalence. Every GPU performs the same ten head updates before any gradient is communicated. A stray early `all_reduce` would mix gradients from different inner steps, so we place an explicit `barrier` between inner and outer phases.

D Universal NER Datasets

To comprehensively evaluate the pretraining method, each permutation of finetuning setup ({head-only, full}, finetuning dataset ({`da_ddt`, ..., `zh_gsdsimp`, `all`}) (where `all` consists of all available training sets), model size ({tiny, small, medium, large}), and pretraining setup ({vanilla, MAML}) is evaluated, for a total of 160 evaluation runs.

- **Publicly Available In-language treebanks** (9 langs): full `train`/`dev`/`test` splits, identical to the official UD partitions.
  - `da_ddt`, `en_ewt`, `hr_set`, `pt_bosque`, `sk_snk`, `sr_set`, `sv_talbanken`, `zh_gsd`, `zh_gsdsimp`
- **Parallel UD (PUD) evaluation** (6 langs): single `test.txt` files, all sentence-aligned across German, English, Portuguese, Russian, Swedish and Chinese.
  - `de_pud`, `en_pud`, `pt_pud`, `ru_pud`, `sv_pud`, `zh_pud`
- **Other eval-only sets** (3 langs): small test splits for low-resource languages.
  - `ceb_gja` (Cebuano), `tl_trg` (Tagalog TRG), `tl_ugnayan` (Tagalog Ugnayan)

D.1 Slovak Fine-Tune Token Statistics

Entity	# spans	% single-token
PER	2 277	87.6 %
LOC	277	75.1 %
ORG	153	56.9 %

Table 4: Span statistics for the Slovak finetune set (`sk_snk` train). The data are strongly person-heavy and person spans are almost always single words, whereas locations and organisations are both rarer and more often multi-token.

D.2 Tagalog and Cebuano Particle and Out-of-Vocabulary Statistics

Language	Particle recall	OOV rate
Tagalog	0.113 ± 0.000	0.523 ± 0.000
Cebuano	0.058 ± 0.000	0.534 ± 0.000

Table 5: Mean (± s.d. across checkpoints) of particle–preceding-span recall and token out-of-vocabulary rate, measured on the zero-shot evaluation sets after Slovak head-only tuning. "Particle recall" is the fraction of gold PER entities whose left context token is a Filipino case particle recognised by the model.

E Pretraining Results

We present the unedited pretraining indicators for each `pico-maml-decoder` model below, as logged on WandB.

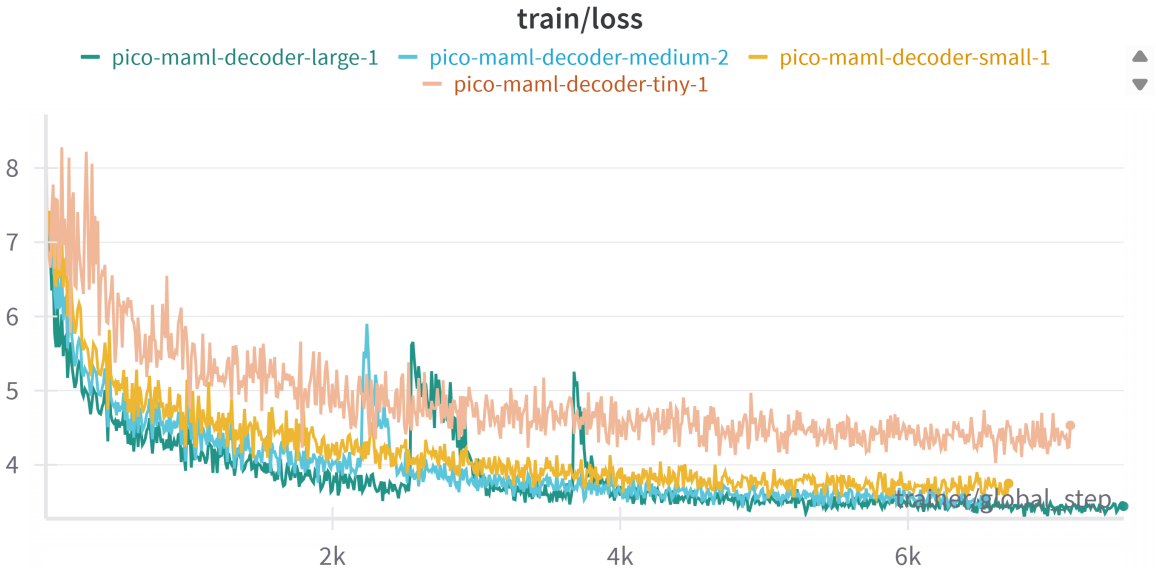


Figure 10: Pretraining training loss curve.

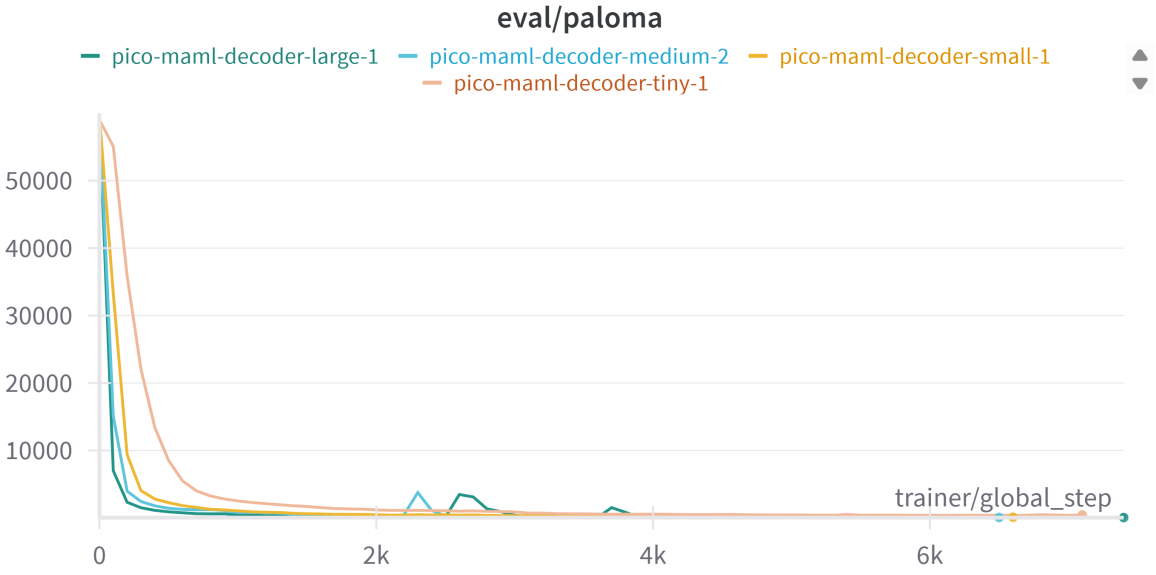


Figure 11: PALOMA score over pretraining steps.

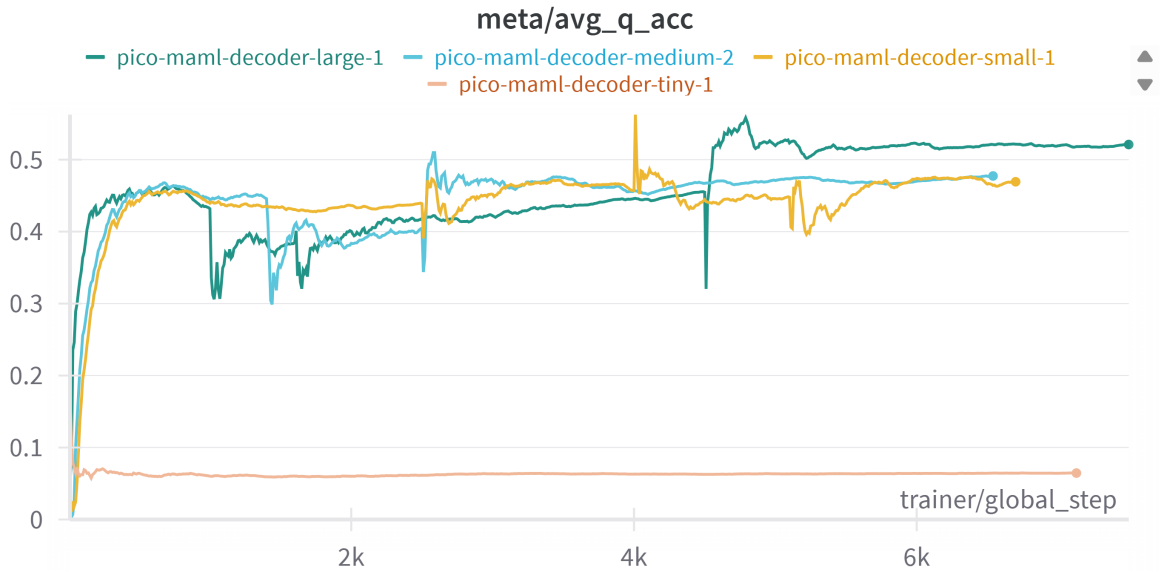


Figure 12: Query accuracy during pretraining.

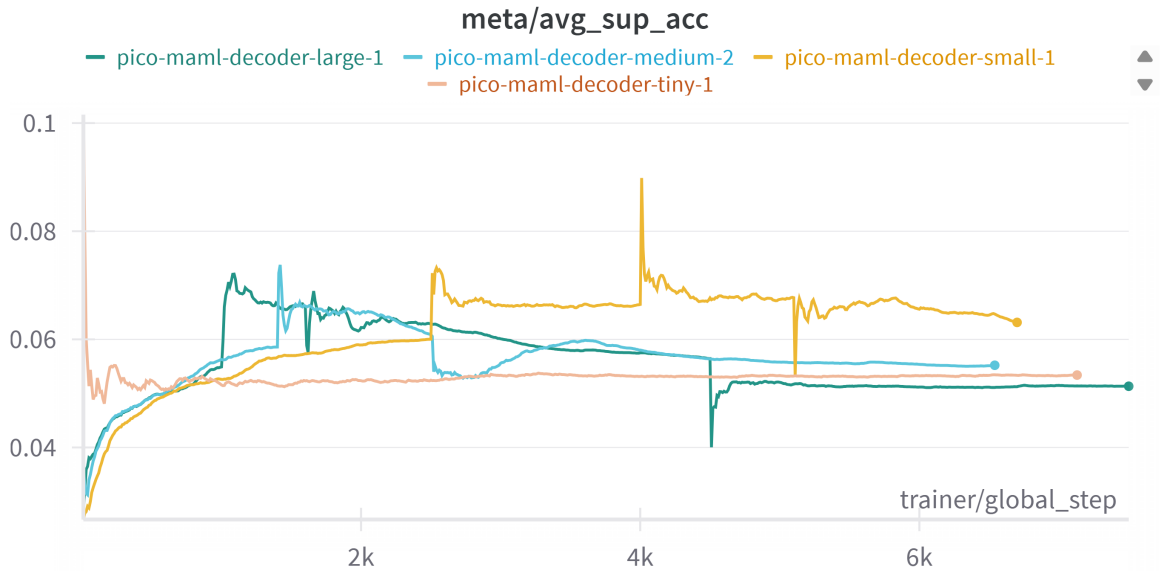


Figure 13: Support accuracy over pretraining.

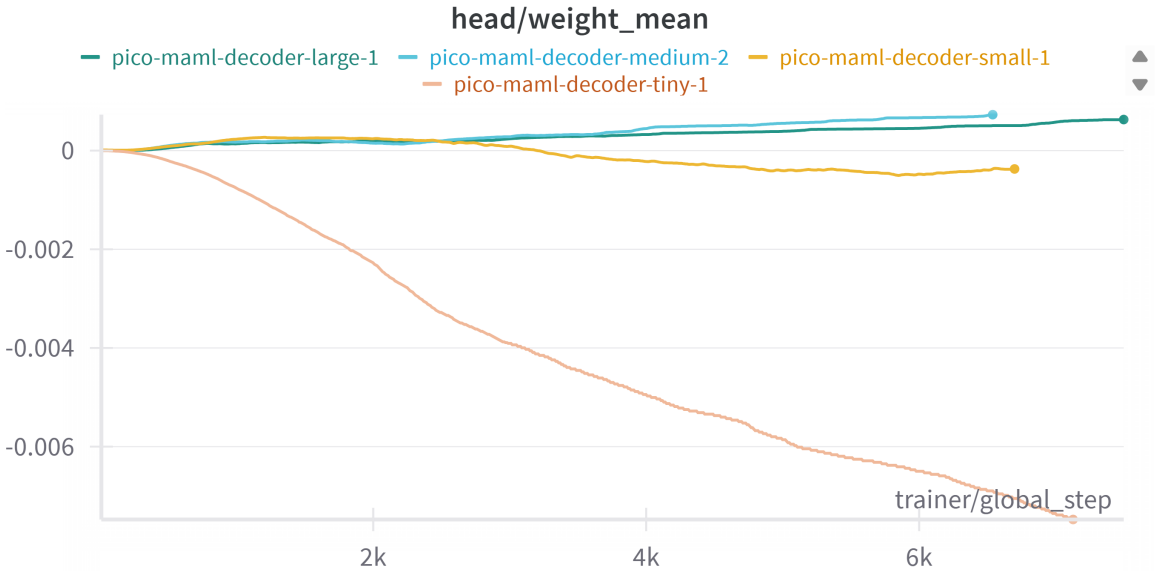


Figure 14: Mean of weights in classifier head over pretraining.

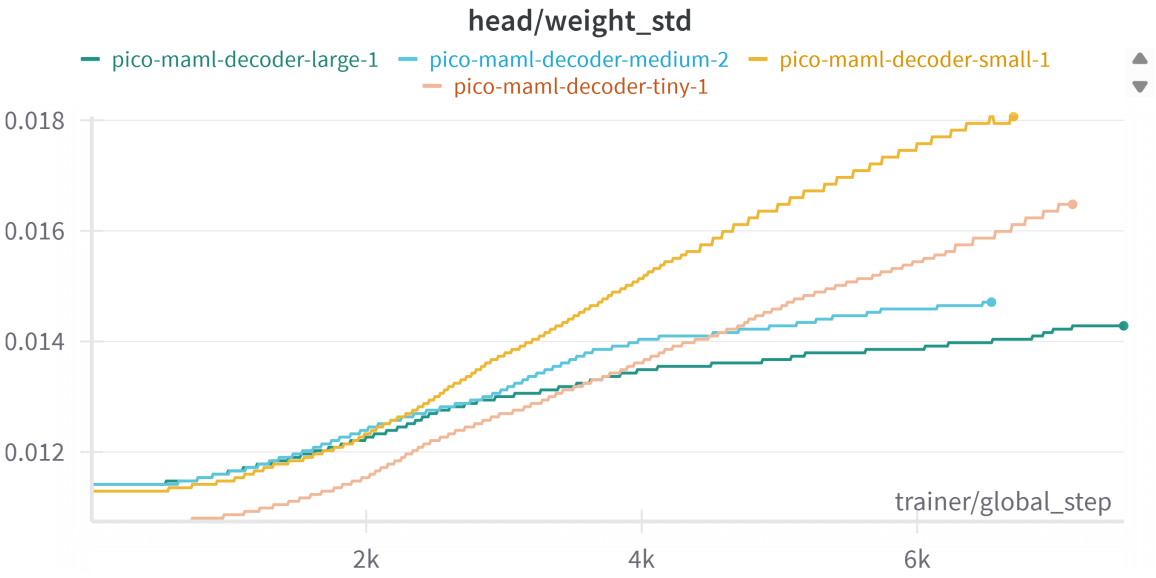


Figure 15: Standard deviation of weights in classifier head over pretraining.

F Default `pico-maml-train` Configurations

Category	Parameter	Default Value
Model	Model Type	`pico_decoder`
	Hidden Dimension (d_{model})	768
	Number of Layers (n_{layers})	12
	Vocabulary Size	50,304
	Sequence Length	2,048
	Attention Heads	12
	Key/Value Heads	4
	Activation Hidden Dim	3,072
	Normalization Epsilon	1×10^{-6}
	Positional Embedding Theta	10,000.0
Training	Optimizer	AdamW
	Learning Rate	3×10^{-4}
	LR Scheduler	Linear w/ Warmup
	Warmup Steps	2,500
	Gradient Accumulation Steps	128
	Max Training Steps	200,000
	Precision	BF16 Mixed
Data	Dataset Name	`pico-lm/pretokenized-dolma`
	Batch Size	1,024
	Tokenizer	`allenai/OLMo-7B-0724-hf`
Checkpointing	Auto Resume	True
	Save Every N Steps	100
	Learning Dynamics Layers	`"attention.v_proj"`, `"attention.o_proj"`, `"swiglu.w_2"`
	Learning Dynamics Eval Data	`pico-lm/pretokenized-paloma-tinsy`
Evaluation	Metrics	`["paloma"]`
	Paloma Dataset Name	`pico-lm/pretokenized-paloma-tinsy`
	Eval Batch Size	16
Monitoring	Logging Level	INFO
	Log Every N Steps	100
Meta-Learning	Enabled	True
	Hybrid Ratio	0.5
	Inner Steps (k)	10
	Inner Learning Rate	0.001
	Support Shots (k)	4
	Query Ways (n)	32
	Classifier Head Layers	4
	Classifier Head Hidden Dim	128
	Classifier Head Dropout	0.1
	Classifier Head Init Method	`xavier`
Monitoring	Logging Level	INFO
	Log Every N Steps	100

Table 6: Default configuration settings used in `pico-maml-train`.

Pico-MAML-Decoder Model Comparison				
Attribute	`tiny`	`small`	`medium`	`large`
Parameter Count	11M	65M	181M	570M
Hidden Dimension (d_{model})	96	384	768	1536
Feed-forward Dim	384	1536	3072	6144
Training Time (6k steps)	10h	15h	16h	25h

Table 7: Comparison of `pico-maml-decoder` model variants trained with default `pico-maml-train` configurations. Except for hidden and feed-forward dimension, all models share the training settings detailed in 6. Models were trained for 6000 training steps on 4 NVIDIA A100-SXM4-80GB GPUs; the listed training times correspond to the initial 6000 steps.