Assessing the Impact of Language Disparity on Multilingual Linguistic Ability in Large Language Models

Anonymous ACL submission

Abstract

Claims about the grammatical competence of multilingual language models vary sharply with how competence is measured, yet the interaction between evaluation paradigm, post-training, and language resource availability has not been systematically examined. We evaluate base and post-trained models from six families on MultiBLiMP, a syntactic minimal-pair benchmark covering 101 languages, using four evaluation methods. We report three principal findings. First, post-training degrades grammatical competence, but the magnitude of this effect is determined by post-training recipe rather than model scale, and scaling reduces it unevenly, with low-resource languages bearing the highest cost. Second, post-trained models retain grammatical knowledge they cannot articulate through explicit prompting, yet this is measurable only in high-resource languages, because near-chance baselines in low-resource settings leave little knowledge to hide. Third, native-language prompting recovers otherwise hidden competence on low-resource languages, demonstrating that only high-resource languages can be probed directly from unprompted probabilities. We conclude that multilingual grammatical evaluation must adopt language-informed, multi-paradigm protocols to avoid systematically underestimating low-resource abilities.

1 Introduction

Multilingual Large Language Models (LLMs) have demonstrated remarkable capabilities in processing and generating text across a wide range of languages. Scaling LLMs up multilingually has led to an increase of evaluation benchmarks that evaluate model performance across a broader spectrum of languages, taking fine-grained linguistic and cultural properties into account (Chang et al., 2025; Vayani et al., 2025). These benchmarks have exposed that language disparity persists in NLP, with advanced LLM abilities remaining limited to a small subset of high-resource languages (Blasi et al., 2022; Wendler et al., 2024).

While various works have investigated the impact of language disparity on downstream tasks (Qi et al., 2023; Xuan et al., 2025), a detailed investigation of its impact on *linguistic ability* remains lacking. For this, we leverage MultiBLiMP (Jumelet et al., 2025), a comprehensive multilingual benchmark consisting of linguistic minimal pairs across 101 languages. However, the exact way of measuring linguistic ability in LLMs remains a topic of discussion (Hu et al., 2026), and we therefore measure model performance across a range of metrics.

We conduct detailed regression analyses to understand performance differences at the language level, focusing on two previously underexplored sources of disparity. First, we investigate the *post-training* phase, in which pre-trained LLMs are fine-tuned for instruction following and human preference alignment. Due to the scarcity of language-specific post-training data, this stage predominantly uses high-resource languages (Longpre et al., 2024), potentially degrading low-resource language abilities (Kew et al., 2024). We quantify this degradation through a **linguistic alignment tax**, measuring the performance gap between base and post-trained models. Our analysis reveals that this tax is determined by post-training recipe rather than model scale, and that scaling reduces it unevenly: low-resource languages pay substantially higher costs than high-resource ones.

Second, building on Song et al. (2025)'s finding that LLMs fail to "introspect" about their grammatical knowledge in English, we extend this analysis across all 101 languages in MultiBLiMP. We introduce the **inarticulate gap**, measuring the difference between direct grammaticality judgments and meta-linguistic evaluation. Our results show that post-trained models retain grammatical knowledge they cannot articulate through explicit prompt-

ing, but critically, this gap is only measurable in high-resource languages, where performance exceeds chance levels. In low-resource settings, near-chance baselines obscure any hidden competence. Statistical analysis confirms that the alignment tax and inarticulate gap are moderately negatively correlated.

Finally, we demonstrate that native-language prompting, in which we provide translated instructions in the target language itself, yields disproportionate gains for low-resource languages, revealing grammatical knowledge that remains weakly expressed under direct probability assessment. These findings stress the need for language-informed, multi-paradigm evaluation protocols that avoid systematically underestimating low-resource language capabilities.

2 Related Work

2.1 Evaluating Linguistic Ability

The grammatical competence of language models has commonly been assessed through the minimal pair paradigm, where a model is considered correct if it assigns higher log-probability to a grammatical sentence over a minimally perturbed ungrammatical counterpart (Linzen et al., 2016; Marvin and Linzen, 2018; Warstadt et al., 2020). A parallel line of work evaluates grammaticality through metalinguistic prompting, asking models directly whether a sentence is grammatical. Dentella et al. (2023) found near-chance accuracy and a systematic yes-response bias under this paradigm, but Hu and Levy (2023) argue that such failures reflect auxiliary task demands rather than grammatical incompetence, since answering metalinguistic questions requires instruction-following ability independent of grammatical knowledge. Hu et al. (2024) provide supporting evidence, showing that log-probability comparisons reveal substantially stronger alignment between models and human judgments.

Song et al. (2025) directly compare both paradigms on the same minimal pairs, finding that post-trained models retain grammatical knowledge they fail to articulate through the chat interface—the *inarticulate gap*. However, this line of work is conducted exclusively in English, leaving open whether the same dissociation holds cross-lingually and how it interacts with language resource availability and post-training.

2.2 Multilingual Evaluation

Minimal pair benchmarks have since been extended to languages beyond English, including Chinese (Liu et al., 2025), Russian (Taktasheva et al., 2024), Dutch (Suijkerbuijk et al., 2025), Japanese (Someya and Oseki, 2023), and Icelandic (Ármannsson et al., 2025). Cross-lingual evaluation has been attempted at small scale by Gulordava et al. (2018) and Mueller et al. (2020), and is scaled to 101 languages by Jumelet et al. (2025) with MultiBLiMP. However, all of these efforts evaluate exclusively with log-probability, leaving the effect of evaluation method unexamined across languages.

The choice of prompt language adds a further complication. Etxaniz et al. (2024) show that multilingual models perform better when prompted in English than in the target language, and Poelman and de Lhoneux (2025) argue that using English prompts to evaluate non-English grammatical knowledge conflates task performance with language understanding. No existing work simultaneously applies multiple evaluation methods across a typologically diverse language set. We address this gap by comparing four evaluation methods across 101 languages spanning four resource tiers.

3 Methods

3.1 Evaluation Methods

We utilize the Language Model Evaluation Harness (`lm-eval`) framework to standardize our assessment. For minimal pair evaluation, which consists of a grammatical sentence (S_{cor}) and an ungrammatical sentence (S_{inc}), we evaluate the model's ability to assign a higher probability to the correct sentence. To investigate the impact of post-training and the discrepancy between intrinsic knowledge and instruction following and model's preference on English, we implemented four distinct evaluation strategies, categorized into Discriminative and Meta paradigms.

Discriminative Evaluation (Likelihood-based) In these methods, the model acts as a probability estimator for the linguistic sequences themselves. We compute the log-likelihood of both the correct (S_{cor}) and incorrect (S_{inc}) sentences conditioned on a specific prefix context C. The model is considered correct if $P(S_{cor}|C) > P(S_{inc}|C)$. We define three variations of the prefix to test context sensitivity:

- **Direct Probability (Zero-Shot):** This method assesses the model's intrinsic knowledge without any prompt engineering. The model scores the raw sequence probability of the sentences based purely on its pre-training distribution, serving as a baseline for linguistic competence.

- **English-Prompted Probability:** To evaluate cross-lingual alignment through English-centric priming, we prepend a standardized English context string C_{eng} (e.g., *"The following is a sentence in [Language Name]. "*) before the target sentences. This tests whether English instructions can successfully trigger the retrieval of low-resource language knowledge.

- **Native-Prompted Probability:** Similar to the English prompt, but the context string C_{native} is provided in the target language itself. This evaluates the model's ability to process instructions and linguistic cues within the same low-resource language space, minimizing cross-lingual interference. For modern languages, C_{native} is a direct translation of the English template (e.g., *"Dies ist ein Satz auf Deutsch."* for German), generated by Gemini 3.0 Pro and verified by Claude 4.6 Opus. For ancient and historical languages where meta-linguistic prompts cannot be naturally translated (e.g., Gothic, Ancient Greek, Church Slavonic), we substitute a representative sentence in the target language to serve as a language cue; a full list of substitute sentences is provided in Appendix E.

Meta Evaluation This method shifts the task from sequence modeling to explicit instruction following. We restructure the minimal pair task into a Question-Answering (QA) format. The model is presented with a chat-style prompt: *"Which of the following sentences is grammatically correct in [Language]? Only respond with 1 or 2."*. Unlike the discriminative methods that score the full sentence, here the model must explicitly generate a classification token. This method specifically tests the model's ability to adhere to formatting constraints and explicitly articulate its linguistic knowledge, and has been referred to as '*meta-linguistic* evaluation' in prior work (Song et al., 2025).

Metric For all methods, we report Accuracy as the primary metric, measuring the percentage of instances where the model correctly assigns higher probability to the grammatical sentence (in discriminative evaluation) or generates the correct option label (in meta evaluation). Accuracy is averaged equally across languages, so that each language contributes the same weight regardless of the number of test items it contains.

3.2 Data

Evaluation Benchmark We evaluate our models on **MultiBLiMP** (Jumelet et al., 2025). This dataset facilitates targeted syntactic evaluation across 101 languages covering a diverse set of language families including Indo-European, Uralic, Turkic, and Afro-Asiatic.

MultiBLiMP consists of over 128,000 minimal pairs. Each pair comprises a grammatical sentence (S_{cor}) and a minimally perturbed ungrammatical counterpart (S_{inc}), created automatically by leveraging annotations from Universal Dependencies and UniMorph. The dataset focuses specifically on morpho-syntactic agreement phenomena, isolating the model's ability to verify grammatical rules without confounding semantic factors. Specifically, it covers:

- **Subject-Verb Agreement**: Tests agreement in number, person, and gender between the subject and the finite verb.

- **Subject-Participle Agreement**: Tests agreement features on participles, which is prominent in Romance and Slavic languages.

Resource Stratification: Common Crawl To investigate the correlation between model alignment and pre-training data volume, we utilize language frequency statistics derived from the common crawl corpus, provided by Kargaran et al. (2024). These language statistics serve as a general proxy to language distributions in web-scale corpora: training data has not been released for any of the models we tested. Based on the natural power-law distribution of web data, we categorize the languages into four distinct resource tiers:

Tier	CC Tokens	#Lang	Examples
Super High	$> 10^{11}$	7	eng, rus, spa
High	10^{9}–10^{11}	38	por, nld, tur
Mid	10^{7}–10^{9}	20	guj, cym, bre
Low/Zero	$\leq 10^{7}$	36	abk, got, yrl

These boundaries follow the natural power-law distribution of web data, where order-of-magnitude

intervals correspond to qualitative shifts in resource availability. The resulting tier sizes reflect the skewed real-world distribution of language resources, with a small number of dominant languages and a long tail of low-resource ones.

3.3 Models

We evaluate models from six open-weight LLM families, each providing paired *base* and post-trained (*PT*) checkpoints for multiple model sizes. For models with native reasoning capabilities, the thinking mode is disabled during evaluation to ensure comparability across all model families. Inputs to PT models are constructed using `apply_chat_template` for all evaluation methods.

Gemma 3 Gemma 3 (Team et al., 2025) is a dense decoder-only transformer trained on a mixture of over 140 languages. We evaluate four parameter scales (1B, 4B, 12B, 27B), each available in both a base and PT checkpoint.

Llama 3.1/3.2 Llama 3.1/3.2 (Grattafiori et al., 2024) is a family of dense decoder-only transformers. We evaluate three parameter scales (1B, 3B, 8B), where the 1B and 3B models are drawn from the Llama 3.2 release and the 8B model from Llama 3.1. Both releases officially support 8 languages; however, Meta notes that the models have been trained on a broader multilingual corpus than these 8 languages, making evaluation on the full MultiBLiMP benchmark feasible.

Qwen3 Qwen3 (Yang et al., 2025) is a dense decoder-only transformer pre-trained on approximately 36 trillion tokens spanning 119 languages and dialects. We evaluate four parameter scales (1.7B, 4B, 8B, 14B).

Qwen3.5 Qwen3.5 (Qwen Team, 2026) is a multimodal model family supporting 201 languages and dialects. The smaller variants (0.8B, 2B, 4B, 9B) use a dense transformer architecture; the 35B variant uses a sparse mixture-of-experts (MoE) architecture with 3B active parameters per forward pass. We evaluate all five sizes.

Gemma 4 Gemma 4 (Noyan et al., 2026) is the newest Gemma family, including both dense (E2B, E4B, 31B) and mixture-of-experts (26B-A4B) architectures, with effective parameter counts of 2.3B, 4.5B, 30.7B, and 3.8B active (25.2B total), respectively. We evaluate all four sizes across both base and post-trained checkpoints.

		Evaluation Method			
		direct	**meta**	**native prompt**	**English prompt**
Gemma 3 (27B)	base	84.17	73.17	86.21	85.77
	PT	79.46	74.02	81.63	79.69
Gemma 4 (31B)	base	85.56	84.38	88.36	88.13
	PT	76.09	82.33	78.06	79.77
Llama 3 (8B)	base	80.88	54.51	83.24	82.74
	PT	79.19	57.31	82.07	81.39
Qwen 3 (30B)	base	80.37	71.40	82.56	81.73
	PT	78.14	68.67	80.36	79.80
Qwen 3.5 (35B)	base	83.45	81.91	85.47	85.56
	PT	82.44	80.48	84.89	84.59
Tiny-Aya (3.35B)	base	76.68	55.01	78.06	77.51
	PT	74.61	54.77	76.65	75.52
Average	base	81.85	70.06	83.98	83.57
	PT	78.32	69.60	80.61	80.13

Table 1: Average model accuracy by evaluation method and model variant. Results are reported for the largest model in each model family, a detailed breakdown of all model scores is provided in Appendix B.

TinyAya TinyAya (Salamanca et al., 2026) is a 3.35B-parameter dense decoder-only transformer developed by Cohere Labs, pre-trained on data spanning over 70 languages with a deliberate emphasis on low-resource language representation. Unlike the other model families evaluated here, TinyAya is specifically designed for multilingual coverage under constrained compute, making it a distinctive point of comparison for low-resource languages on MultiBLiMP. We only evaluate the base and the post-trained TinyAya-Global variant.

4 Experiments Results

We present our results in five parts. We first report general performance (§4.1) and direct evaluation accuracy (§4.2), followed by an analysis of differences between evaluation metrics (§4.3–§4.5).

4.1 General Model Performance

We report the general model performance of the largest model in each model family in Table 1, across the four evaluation methods and split out by base/PT version. These results demonstrate that the assessment of linguistic ability is strongly driven by evaluation method and model variant. The base Gemma 4 model obtains the highest score overall, with the *native prompt* method achieving 88.36% accuracy. The combination of native prompt method with base model obtains the highest accuracy on average across models as well, demonstrating that conditioning the model on an instruction in the language we evaluate on improves its

		Resource Tier			
		Low Zero	**Mid**	**High**	**Super High**
Gemma 3 (27B)	base	66.72	86.60	96.77	98.60
	PT	61.70	81.20	92.49	95.15
Gemma 4 (31B)	base	68.61	88.22	97.77	98.86
	PT	61.08	76.96	86.95	91.79
Llama 3 (8B)	base	64.96	82.89	91.91	97.11
	PT	63.42	81.64	89.88	95.31
Qwen 3 (30B)	base	63.26	81.06	93.15	97.03
	PT	61.58	78.01	90.85	94.65
Qwen 3.5 (35B)	base	66.25	86.36	95.56	97.91
	PT	65.62	84.98	94.36	96.96
Tiny-Aya (3.35B)	base	62.07	74.14	88.01	97.60
	PT	61.81	71.19	85.11	93.19
Average	base	65.31	83.21	93.86	97.85
	PT	62.53	79.00	89.94	94.51

Table 2: Average model accuracy by resource tier and model variant, for the *direct* evaluation method.

grammaticality judgment.

Averaged across models, there is substantial divergence between the three discriminative methods and the meta evaluation method, with the discriminative methods ranging from 78.32% (*direct*, PT) to 83.98% (*native prompt*, base), and the meta method at 69.6% (PT) and 70.06% (base). However, the variance across models is substantially higher for the meta method, with Gemma 4 obtaining 84.38% and Llama 3 only 54.51%. In the next sections we will explore in more detail which model and data properties drive these differences.

4.2 Direct Evaluation

Table 2 reports mean Direct accuracy for the largest model in each family, averaged equally across languages within each resource tier. Across all model families, accuracy increases monotonically with resource availability. Base models consistently outperform their PT counterparts across all tiers, with the alignment tax more pronounced in lower-resource tiers. Super High Resource languages yield substantially higher accuracy than Low/Zero Resource languages, with gaps of approximately 30–35 percentage points, reflecting the strong influence of pretraining data availability on grammatical competence. All results for individual model sizes and languages can be found in Appendix D.

Among the six families, Gemma3-27B and Gemma4-31B achieve the highest accuracy across resource tiers, with base models reaching 98.60% and 98.86% respectively on Super High languages. TinyAya shows competitive performance relative to its size of 3.35B parameters, yet despite its claims about more equal language support, it remains heavily skewed towards high resource languages. A more detailed comparison of TinyAya against similarly sized models is provided in Appendix C. For Low/Zero Resource languages, all models cluster in the range of 60–69%, approaching but not reaching the 50% chance baseline.

4.3 The Linguistic Alignment Tax

Figure 1 plots base accuracy against post-trained accuracy for each ⟨language, model size⟩ combination in the Gemma 3 family. We focus on Gemma 3 for scaling analyses throughout §4.3–4.4, as it offers the broadest parameter ladder (1B–27B) among the evaluated families; cross-family comparisons are provided at the end of each subsection. Two patterns emerge: base and PT accuracy are strongly correlated, but the majority of points fall below the diagonal, indicating that post-training consistently *degrades* grammatical judgment. We term this degradation the **linguistic alignment tax**: $\Delta = \mathrm{Acc}_{\mathrm{base}} - \mathrm{Acc}_{\mathrm{PT}}$.

However, raw tax alone can be misleading. English and Hebrew both lose approximately 10 percentage points after post-training, yet the proportional damage differs substantially: English (base: 99.1%) retains 80.9% of its above-chance knowledge, while Hebrew (base: 87.6%) retains only 73.7%. We therefore define the *normalized linguistic alignment tax*:

$$\text{Normalized Tax} = \frac{\mathrm{Acc}_{\mathrm{base}} - \mathrm{Acc}_{\mathrm{PT}}}{\mathrm{Acc}_{\mathrm{base}} - 0.5} \quad (1)$$

Languages where $\mathrm{Acc}_{\mathrm{base}} \leq 0.5$ are excluded from normalization. In the remainder of this section, "tax" refers to the normalized tax.

Tax by Resource Tier Figure 2 reports the normalized tax for Gemma 3 across model scales and resource tiers. Super High and High Resource languages respond well to scaling: their taxes drop from 23.4% and 33.6% at 1B to 7.1% and 9.3% at 27B, respectively. Mid-Resource languages follow a similar but slower trajectory (36.9% → 16.1%), with diminishing returns beyond 12B. In the Low/Zero tier, tax drops from 35.9% at 1B to 12.0% at 12B, before rising to 20.6% at 27B. While scaling reduces the alignment tax, the benefit is uneven: low-resource languages suffer the highest tax and remain substantially above zero even for the largest model size.

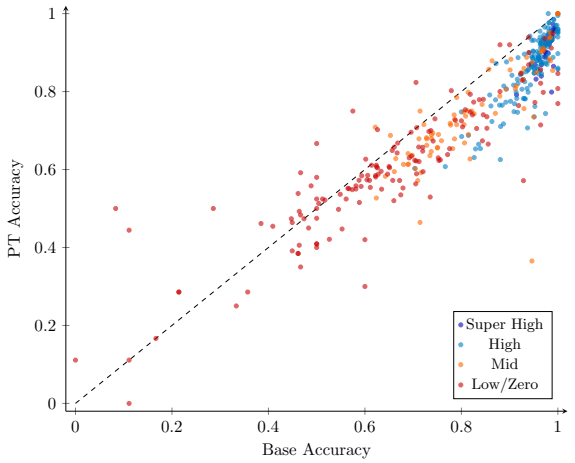

Figure 1: Base vs. PT accuracy for Gemma 3 across 101 languages. Each point is one (language, model size) combination; color indicates resource tier. The dashed diagonal marks parity; points below the diagonal indicate linguistic alignment tax.

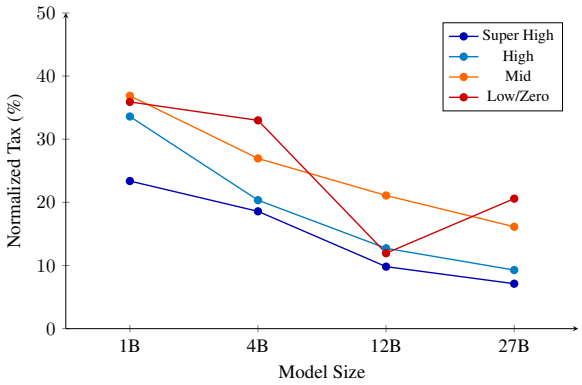

Figure 2: Normalized linguistic alignment tax for Gemma 3 by resource tier and model size. 12 languages where $\mathrm{Acc}_{\mathrm{base}} \leq 0.5$ have been excluded from all sizes.

Cross-Family Comparison Figure 3 reports the normalized tax for the largest model in each family. The ranking reveals that alignment tax is not driven by model size: TinyAya (3.35B) exhibits the lowest tax (1.4%), while Gemma 4-31B shows the highest (30.9%). Since base accuracies are comparable across families (80.4%–85.6%), the divergence arises primarily from post-training recipes rather than base competence, suggesting that alignment tax is a model-family-specific property rather than a universal cost of post-training.

4.4 The Inarticulate Gap

We define the **inarticulate gap** as the difference between a post-trained model's 'implicit' and 'explicit' grammatical knowledge:

$$\mathrm{Gap} = \mathrm{Acc}_{\mathrm{Direct}}(\mathrm{PT}) - \mathrm{Acc}_{\mathrm{Meta}}(\mathrm{PT}) \quad (2)$$

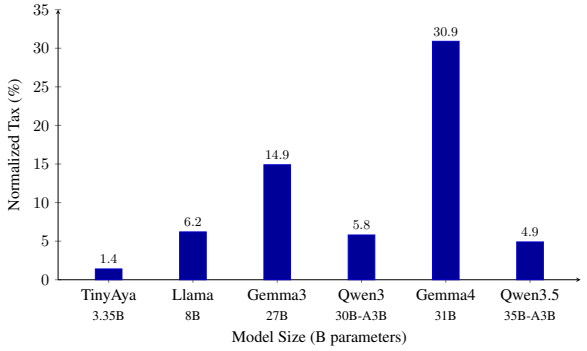

Figure 3: Normalized linguistic alignment tax for the largest model in each family, measured using the Direct method, and averaged equally across languages with above-chance base accuracy (91–93 of 101, varying by model)

A positive gap indicates that the model retains grammatical knowledge it cannot articulate through the chat interface. As in §4.3, we use Gemma 3 for within-family scaling analyses; cross-family comparisons follow at the end of this subsection. This Direct-vs-Meta discrepancy parallels the paradigm used by Song et al. (2025) to study introspection, though our focus is on quantifying the knowledge gap rather than testing privileged self-access.

Base Models as a Format Mismatch Baseline Base models also exhibit a positive gap across all resource tiers, indicating that the gap is not solely an artifact of post-training. For the Gemma 3 family, the gap in Super High languages narrows from 0.473 at 1B to 0.070 at 27B, suggesting that larger pre-trained models are less affected by the format mismatch between Direct probability measurements and Meta prompting. We treat the base gap as a structural baseline against which PT convergence is measured. (see Table 9)

PT Models: Hidden Competence Persists Figure 4 plots Direct vs. Meta accuracy for all ⟨language, model size⟩ pairs under the PT setting, and Figure 5 reports the mean gap by tier and scale. The majority of points lie above the diagonal—89.3% of Super High, 86.8% of High, 92.5% of Mid, and 66.7% of Low/Zero Resource pairs exhibit Direct accuracy exceeding Meta accuracy—confirming the *hidden competence* hypothesis (Song et al., 2025). Scaling reduces the gap, but unevenly: Super High and High Resource languages improve substantially (Super High: 0.228 → 0.038; High: 0.198 → 0.044), while Low/Zero Resource languages remain flat

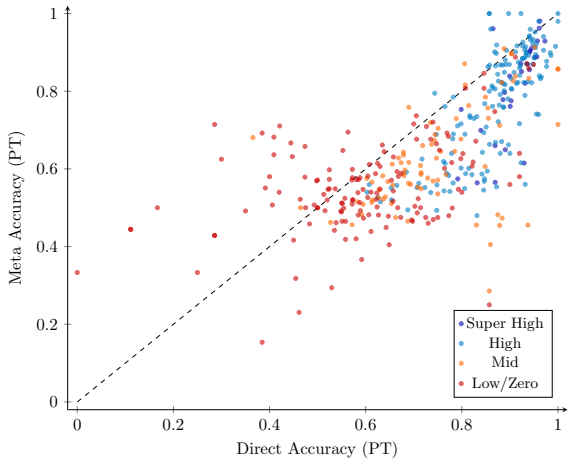

Figure 4: Direct vs. Meta accuracy for Gemma 3 under the PT setting. Each point is one ⟨language, model size⟩ combination; color indicates resource tier. The dashed diagonal marks parity; points above the diagonal indicate a positive inarticulate gap.

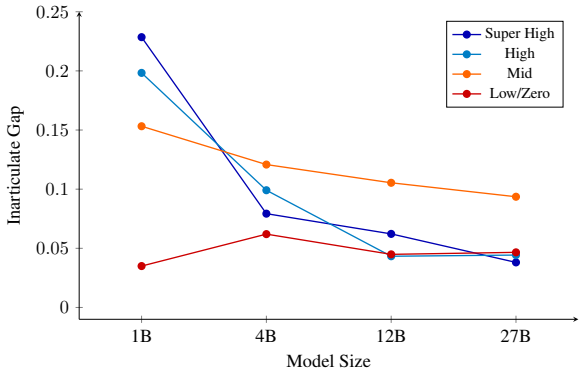

Figure 5: Mean inarticulate gap for Gemma 3 by resource tier and model size. Gap is computed as $\mathrm{Acc}_{\mathrm{Direct}}(\mathrm{PT}) - \mathrm{Acc}_{\mathrm{Meta}}(\mathrm{PT})$, averaged equally across languages within each tier.

(0.035 → 0.047), not because scaling fails but because Direct accuracy in this tier is already close to chance, leaving little room for a gap to form. The inarticulate gap is therefore primarily a high-resource phenomenon.

Cross-Family Comparison Figure 6 reports the mean gap for the largest model in each family. Llama (0.219) and TinyAya (0.198) exhibit the largest gaps, while Gemma 4 shows a *negative* gap (−0.062), indicating that its Meta accuracy exceeds its Direct accuracy. Since base accuracies are comparable across families (80.4%–85.6%), this divergence arises from post-training recipes rather than base competence. The mean pairwise Spearman correlation of language-level gaps across all six families ($\rho = 0.412$) is substantially higher than

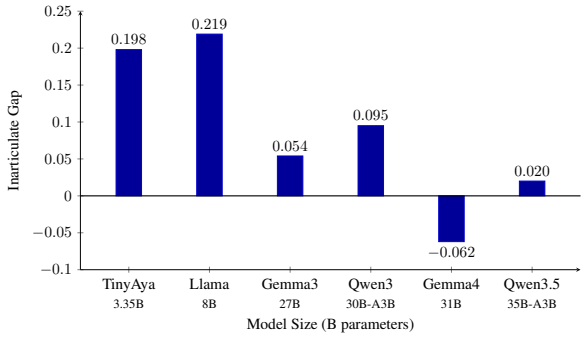

Figure 6: Mean inarticulate gap for the largest model in each family, computed as $\mathrm{Acc}_{\mathrm{Direct}}(\mathrm{PT}) - \mathrm{Acc}_{\mathrm{Meta}}(\mathrm{PT})$, averaged equally across 101 languages. Models are ordered by total parameter count.

the same analysis for alignment tax ($\rho = 0.206$), indicating that the gap is more strongly driven by language-level properties, while alignment tax is more sensitive to model-specific training recipes.

Gap, Alignment Tax, and Pre-training The gap and alignment tax are moderately negatively correlated: their language-level Spearman correlation is $\rho = -0.396$ ($p < 0.001$), both averaged over the same subset of languages with $\mathrm{Acc}_{\mathrm{base}} > 0.5$ at all model sizes), with four of six families (Gemma 3, Gemma4, Llama3, TinyAya) showing moderate negative correlations (ρ ranging from -0.26 to -0.58). This shows that both the gap and alignment tax are impact similarly by language resources. Furthermore, the base gap is a reliable predictor of the PT gap ($\rho = 0.603$, $p < 0.001$, consistent across all families), suggesting that 'inarticulation' arises during pre-training: languages where the base model struggles to express grammatical knowledge in chat format also exhibit larger gaps after post-training.

4.5 Language-specific Prompting

To investigate the role of language-specific prompting in eliciting grammatical knowledge, we introduce a complementary metric: the **Prompt Gain**, defined as the difference between accuracy when using native prompting versus the direct evaluation method:

$$\text{Prompt Gain} = \mathrm{Acc}_{\mathrm{Native}}(\mathrm{PT}) - \mathrm{Acc}_{\mathrm{Direct}}(\mathrm{PT}) \quad (3)$$

While the inarticulate gap captures the failure to articulate latent knowledge under English instructions, the Prompt Gain quantifies how much performance can be improved by providing instructions in the target language itself. A model may

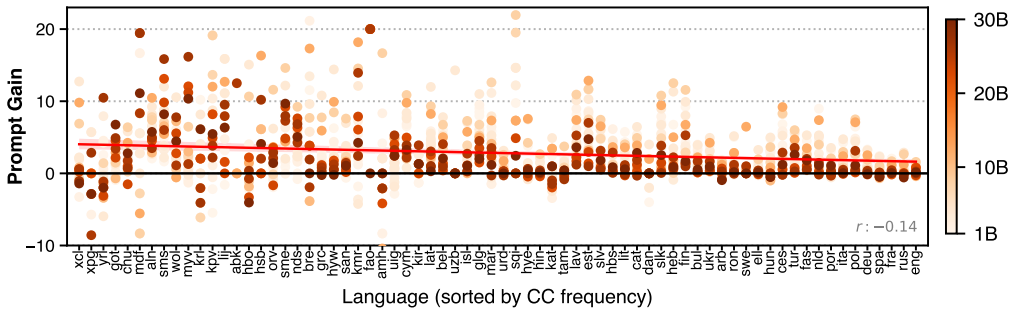

Figure 7: Prompt Gain set out against language frequency across all evaluated models, which are colored by size in billions of parameters.

require a language-specific cue to adapt its probability distribution to the linguistic structure specific to that language. A positive value then indicates that native-language prompting successfully activates linguistic knowledge that remains dormant under direct evaluation, effectively bridging the articulation gap through language-specific contextualization.

Figure 7 reveals a striking cross-lingual gradient in the Prompt Gain. For high-resource languages, the gain approaches zero, indicating that explicit prompting provides minimal benefit over the model's raw probabilistic judgments: the model's base predictions are sufficiently robust and tailored towards these high-resource languages that grammatical ability manifests accurately without language-specific prompting. For low-resource languages, the advantage is substantially more positive, demonstrating that explicit, language-specific instructions help the model apply grammatical knowledge that is present but weakly expressed in raw probabilities. This pattern is stronger in smaller models for mid- to high-resource languages: smaller models thus need explicit prompting to reliably access their grammatical knowledge, while larger models can apply it more directly. This pattern supports Poelman and de Lhoneux (2025)'s argument against using English as an 'evaluation interface': native prompting substantially improves low-resource language performance, indicating that direct probability and English prompting underestimate grammatical competence where language-specific contextualization is absent.

5 Discussion

Our findings reveal two core properties of multilingual grammatical knowledge in LLMs. First, post-training consistently degrades grammatical competence across languages—the alignment tax—with the degradation amplified in low-resource languages. Second, post-trained models retain grammatical knowledge they cannot express through the chat interface—the inarticulate gap—which is driven by language-level properties rather than model-specific training recipes. Together, these patterns show that linguistic knowledge acquired during pre-training is systematically filtered through post-training, leading to losses that depend on both resource availability and evaluation methods.

A methodological implication of our work is that conclusions about a model's grammatical competence are highly evaluation-dependent. Metalinguistic prompting yields accuracy ranging from near-chance (Llama-3 at 55%) to high (Gemma-4 at 84%) across families with comparable base competence, revealing that it conflates grammatical knowledge with instruction-following ability in a model-family-specific way. Future multilingual evaluation should therefore report multiple paradigms rather than treating any single method as definitive.

On the practical side, our native prompt results offer a mitigation: providing evaluation instructions in the target language consistently improves accuracy for low-resource languages, suggesting that direct evaluation may actually be *underestimating* low-resource performance. Understanding the mechanisms that trigger linguistic abilities for such languages is an exciting direction for future work, adapting findings from multilingual interpretability in an actionable way to reduce the detrimental effects of language disparity (Brinkmann et al., 2025; Kryvosheieva et al., 2026).

Limitations

Given all the experiments are conducted on the dataset of MultiBLiMP, which isolates a single class of subject-verb and subject-participle agreement in number, person and gender, the paper's conclusion about grammatical competence therefore generalize from agreement to grammar more broadly. The patterns we report should be read primarily as properties of agreement processing, and may be prone to limitations propagating from MultiBLiMP itself. Low-resource languages have a smaller set of minimal pairs in MultiBLiMP, and therefore result in a more uncertain estimate of model performance.

Another limitation is about the resource tiers that are from Common Crawl corpus counts, since training data has not been released for any of the evaluated models. Actual per-language exposure during pre-training may differ substantially across model families, so a family's behavior within a given tier may reflect its own training mixture rather than the tier itself, and cross-family tier-level comparisons should be interpreted with caution. A comprehensive study training models from scratch on different training data mixtures would provide a valuable insight into the degree to which linguistic abilities are driven by typological features.

References

Bjarki Ármannsson, Finnur Ágúst Ingimundarson, and Einar Freyr Sigurðsson. 2025. An Icelandic linguistic benchmark for large language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 37–47, Tallinn, Estonia. University of Tartu Library.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. 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 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Abdelrahman Eldesokey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, Aditi Gupta, Afitab Iyigun, Afonso Simplício, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akshay Ramesh, Aleksei Dorkin, and 319 others. 2025. Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures. *Preprint*, arXiv:2510.24081.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Jennifer Hu, Ethan Gotlieb Wilcox, Siyuan Song, Kyle Mahowald, and Roger P. Levy. 2026. What can string probability tell us about grammaticality? *Transactions of the Association for Computational Linguistics*, 14:124–146.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. Glotcc: An open broad-coverage commoncrawl corpus and pipeline for minority languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 16983–17005. Curran Associates, Inc.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2024. Turning English-centric LLMs into polyglots: How much multilinguality is needed? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13097–13124, Miami, Florida, USA. Association for Computational Linguistics.

Daria Kryvosheieva, Andrea de Varda, Evelina Fedorenko, and Greta Tuckute. 2026. Different types of syntactic agreement recruit the same units within large language models. *Preprint*, arXiv:2512.03676.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2025. A systematic assessment of language models with linguistic minimal pairs in chinese. *Preprint*, arXiv:2411.06096.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi (Alexis) Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2024. A large-scale audit of dataset licensing and attribution in ai. *Nature Machine Intelligence*, 6(8):975–987.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Merve Noyan, Pedro Cuenca, Sergio Paniego, Ben Burtenshaw, Steven Zheng, Alvaro Bartolome, and Nathan Habib. 2026. Welcome gemma 4: Frontier multimodal intelligence on device.

Wessel Poelman and Miryam de Lhoneux. 2025. The roles of English in evaluating multilingual language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 492–498, Tallinn, Estonia. University of Tartu Library.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Alejandro R. Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, and 7 others. 2026. Tiny aya: Bridging scale and multilingual depth. *Preprint*, arXiv:2603.11510.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. 2025. Language models fail to introspect about their knowledge of language. *Preprint*, arXiv:2503.07513.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, 51(4):1267–1301.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian benchmark of linguistic minimal pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Minkov Mihaylov, Chao

Qin, Abdelrahman M. Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Gian Esplana, Monil Gokani, and 50 others. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19565–19575.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Language List

Super High			High			Mid			Low/Zero		
ISO	Lang	N	ISO	Lang	N	ISO	Lang	N	ISO	Lang	N
deu	German	27576	arb	Arabic	14580	amh	Amharic	1344	abk	Abkhazian	480
eng	English	9240	bel	Belarusian	30840	bre	Breton	3120	aln	Gheg Alb.	8124
fra	French	30576	ben	Bengali	252	bua	Buriat	1236	apu	Apurinã	336
ita	Italian	35988	bul	Bulgarian	29496	cym	Welsh	13440	aqz	Akuntsu	168
pol	Polish	39264	cat	Catalan	27408	fao	Faroese	2784	azz	H-P Nahuatl	2484
rus	Russian	45984	ces	Czech	51072	fro	Old French	23712	bho	Bhojpuri	408
spa	Spanish	30492	dan	Danish	600	gla	Scot. Gaelic	792	bor	Borôro	2892
			ell	Mod. Greek	13152	gle	Irish	336	chu	Church Slav.	49992
			est	Estonian	30900	grc	Anc. Greek	44340	egy	Egyptian	264
			eus	Basque	3276	guj	Gujarati	84	frm	Mid. French	3528
			fas	Persian	30636	hsb	Up. Sorbian	2232	got	Gothic	18948
			fin	Finnish	30840	hyw	W. Armenian	13836	hbo	Anc. Hebrew	11796
			glg	Galician	9036	kir	Kyrgyz	2220	hit	Hittite	600
			hbs	Serbo-Croat.	39432	kmr	N. Kurdish	6528	koi	Komi-Perm.	516
			heb	Hebrew	27960	nds	Low German	21288	kpv	Komi-Zyr.	3840
			hin	Hindi	17364	orv	Old Russian	55380	krl	Karelian	3120
			hun	Hungarian	10140	sah	Yakut	1728	kxh	Karo	1440
			hye	Armenian	16980	san	Sanskrit	53304	lij	Ligurian	3048
			isl	Icelandic	33612	sme	N. Sami	30432	mdf	Moksha	984
			kat	Georgian	2448	uig	Uighur	9096	myv	Erzya	5568
			kaz	Kazakh	2076				nhi	T. Nahuatl	456
			lat	Latin	37788				olo	Livvi	2280
			lav	Latvian	36384				ota	Ott. Turkish	1188
			lit	Lithuanian	14160				pcm	Nig. Pidgin	312
			mar	Marathi	5520				quc	K'iche'	1572
			mkd	Macedonian	468				sms	Skolt Sami	3156
			nld	Dutch	27972				tpn	Tupinambá	108
			por	Portuguese	36576				ttc	Tektiteko	828
			ron	Romanian	24672				urb	Urubú-Kaap.	156
			slk	Slovak	49740				vep	Veps	2244
			slv	Slovenian	53796				wbp	Warlpiri	144
			sqi	Albanian	2916				wol	Wolof	8460
			swe	Swedish	2412				xcl	Class. Armenian	19476
			tam	Tamil	4584				xnr	Kangri	1032
			tur	Turkish	20904				xpg	Phrygian	600
			ukr	Ukrainian	32928				yrl	Nhengatu	8640
			urd	Urdu	6600						
			uzb	Uzbek	600						

Table 3: All 101 languages in the MultiBLiMP benchmark, organized by resource tier. N = number of minimal pairs. SH = Super High ($>10^{11}$ CC tokens), H = High (10^9–10^{11}), M = Mid (10^7–10^9), LZ = Low/Zero ($<10^7$).

B Average Model Accuracy

Table 4 reports the average model accuracy for each model that we evaluated, for the four different evaluation methods.

			Evaluation Method			
	Size	Variant	**direct**	**meta**	**native prompt**	**English prompt**
Gemma 3	1.0	PT	67.98	54.66	70.52	69.75
		base	77.34	49.58	79.52	79.05
	4.0	PT	74.43	65.55	75.26	73.97
		base	81.38	60.03	82.84	82.49
	12.0	PT	77.91	72.16	79.87	77.53
		base	83.54	65.89	84.89	84.65
	27.0	PT	79.46	74.02	81.63	79.69
		base	84.17	73.17	86.21	85.77
Gemma 4	2.0	PT	72.28	63.06	74.05	73.76
		base	80.86	50.15	82.93	82.03
	4.0	PT	74.61	63.03	77.42	76.89
		base	82.63	64.78	85.17	83.90
	26.0	PT	74.56	77.75	76.39	76.59
		base	83.96	76.06	86.86	86.54
	31.0	PT	76.09	82.33	78.06	79.77
		base	85.56	84.38	88.36	88.13
Llama 3	1.0	PT	70.96	49.62	73.32	73.19
		base	75.68	49.46	77.98	76.53
	3.0	PT	74.34	53.76	77.23	76.12
		base	78.59	50.89	81.21	80.26
	8.0	PT	79.19	57.31	82.07	81.39
		base	80.88	54.51	83.24	82.74
Qwen 3	0.6	PT	70.39	50.34	72.32	71.81
		base	73.16	50.90	75.14	74.19
	1.7	PT	72.91	56.47	74.76	73.84
		base	75.50	56.52	77.78	77.35
	4.0	PT	75.34	54.08	77.07	77.34
		base	77.27	64.43	80.35	79.44
	8.0	PT	78.18	50.65	79.79	79.50
		base	72.22	70.17	81.87	81.80
	14.0	PT	78.54	72.09	80.31	80.05
		base	78.25	72.13	82.86	82.15
	30.0	PT	78.14	68.67	80.36	79.80
		base	80.37	71.40	82.56	81.73
Qwen 3.5	0.8	PT	74.31	54.35	76.10	75.88
		base	74.84	56.64	77.93	76.91
	2.0	PT	77.57	61.41	79.23	78.96
		base	78.89	62.56	80.91	80.14
	4.0	PT	78.45	67.72	81.34	80.72
		base	77.82	71.33	82.84	81.61
	9.0	PT	81.52	76.03	83.55	83.49
		base	79.44	78.24	84.87	84.50
	35.0	PT	82.44	80.48	84.89	84.59
		base	83.45	81.91	85.47	85.56
Tiny-Aya	3.35	PT	74.61	54.77	76.65	75.52
		base	76.68	55.01	78.06	77.51

Table 4: Model accuracy by method, size, and base/PT variant.

C Tiny Aya Performance

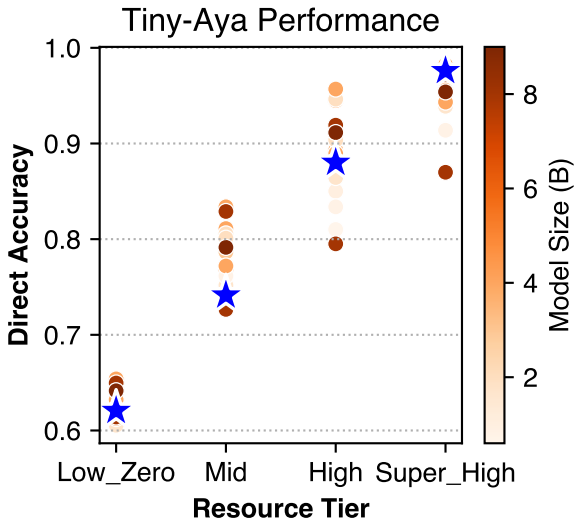

Figure 8: Performance of Tiny-Aya (blue star) against all other models up to 9B parameters. Tiny-Aya demonstrates a strong skew towards the Super High resource languages, and underperforms on low- and mid-resource languages.

D Cross-Family Results

Tables 5 and 6 report per-language Mean Direct accuracy for all model variants.

Table 5: Mean Direct accuracy (%) — Base models.

		Gemma3			Gemma4				Llama			Qwen3							Qwen3.5			TinyAya		
ISO	N	1B	4B	12B	27B	E2B	E4B	26B-A4B	31B	1B	3B	8B	0.6B	1.7B	4B	8B	14B	30B-A3B	0.8B	2B	4B	9B	35B-A3B	3.35B
abk	40	57.5	60.0	50.0	80.0	75.0	80.0	57.5	50.0	75.0	70.0	57.5	52.5	62.5	47.5	67.5	55.0	80.0	75.0	70.0	57.5	80.0	42.5	72.5
aln	677	75.3	75.6	80.8	81.8	76.5	79.9	84.5	86.6	70.9	69.9	74.6	70.3	71.3	74.5	69.4	70.0	73.7	68.4	72.1	71.6	75.2	75.6	74.2
amh	112	92.9	94.6	98.2	98.2	96.4	92.9	99.1	92.9	94.6	94.6	99.1	90.2	90.2	88.4	96.4	97.3	97.3	86.6	95.5	92.9	86.6	96.4	92.0
apu	28	92.9	96.4	96.4	96.4	96.4	92.9	92.9	96.4	96.4	92.9	96.4	96.4	92.9	96.4	96.4	92.9	96.4	92.9	96.4	100.0	96.4	96.4	92.9
aqz	14	35.7	28.6	21.4	21.4	35.7	42.9	14.3	28.6	21.4	42.9	35.7	42.9	21.4	21.4	57.1	28.6	42.9	42.9	42.9	42.9	28.6	28.6	35.7
arb	1215	93.1	93.2	95.1	94.9	92.8	94.6	95.0	95.6	84.0	88.5	90.6	85.3	91.0	90.9	82.6	92.4	92.8	85.4	89.7	89.8	91.5	94.1	94.9
azz	207	72.0	72.0	69.6	66.7	68.6	71.0	71.5	71.0	65.7	67.2	69.1	69.6	67.6	69.6	73.4	69.1	75.4	74.4	67.2	70.5	68.6	70.5	71.0
bel	2570	88.1	94.3	96.0	97.3	93.0	95.1	96.7	97.6	78.8	85.2	88.5	68.2	76.8	81.1	65.6	86.8	87.6	78.9	87.5	86.4	89.1	94.7	73.2
ben	21	95.2	100.0	95.2	100.0	100.0	90.5	90.5	100.0	95.2	90.5	90.5	90.5	95.2	95.2	85.7	95.2	95.2	85.7	95.2	95.2	95.2	100.0	61.9
bho	34	73.5	70.6	85.3	85.3	73.5	76.5	85.3	94.1	50.0	76.5	79.4	55.9	55.9	61.8	52.9	55.9	55.9	50.0	67.7	64.7	70.6	76.5	52.9
bor	241	60.2	60.6	62.2	66.8	61.8	62.7	62.2	66.0	61.0	66.8	67.2	64.7	66.8	62.2	59.8	61.8	62.7	63.5	65.1	69.7	63.9	66.8	61.4
bre	260	94.6	95.0	96.9	96.5	93.1	92.7	93.8	97.3	92.3	90.0	94.6	53.5	61.2	76.1	51.1	68.1	90.8	85.4	92.7	82.7	91.5	96.5	74.6
bua	103	62.1	73.8	73.8	77.7	73.8	75.7	76.7	82.5	68.9	66.0	67.0	67.0	71.8	72.8	68.9	65.0	69.9	70.9	68.9	64.1	72.8	75.7	69.9
bul	2458	93.6	96.9	98.2	99.2	96.5	97.2	98.9	99.2	86.9	90.6	92.8	83.2	87.5	92.8	79.8	94.4	95.4	85.3	91.2	90.4	93.6	97.2	97.2
cat	2284	92.6	95.8	97.0	96.6	95.2	96.8	97.5	98.4	88.3	92.0	94.7	85.2	89.5	92.9	82.8	94.2	95.3	86.5	90.7	91.1	93.5	95.2	96.5
ces	4256	86.3	93.1	96.0	96.9	94.1	96.1	97.2	98.1	84.5	88.3	91.8	77.2	82.8	86.9	72.2	86.1	91.2	79.4	86.7	85.5	87.5	94.1	95.7
chu	4166	62.4	62.7	66.1	69.2	63.7	65.0	68.9	73.7	62.0	63.5	63.4	61.8	62.2	63.2	60.4	62.8	63.6	61.1	62.6	62.3	63.7	66.3	60.3
cym	1120	78.2	84.3	90.1	90.3	88.2	90.8	92.4	94.5	79.0	84.4	91.2	70.0	75.2	79.5	69.5	80.5	86.5	73.8	80.6	79.5	83.3	92.9	85.1
dan	50	98.0	100.0	100.0	98.0	100.0	98.0	100.0	100.0	98.0	100.0	100.0	94.0	96.0	96.0	84.0	98.0	100.0	96.0	98.0	96.0	100.0	100.0	100.0
deu	2298	97.5	98.0	98.8	99.1	98.7	98.7	99.3	99.2	96.4	97.0	98.1	93.8	94.1	96.5	86.3	95.7	98.4	92.9	96.3	95.3	97.1	98.7	98.5
egy	22	40.9	50.0	50.0	50.0	50.0	50.0	54.5	50.0	40.9	50.0	50.0	45.5	45.5	45.5	45.5	54.5	40.9	50.0	45.5	45.5	40.9	40.9	45.5
ell	1096	97.5	99.0	99.6	99.8	99.4	99.5	98.8	99.2	93.7	96.8	97.5	91.1	96.3	97.3	87.6	98.7	98.5	93.1	96.7	94.2	96.2	98.6	99.5
eng	770	98.7	99.1	99.4	99.2	98.8	98.3	97.8	98.4	98.8	98.8	98.3	98.3	98.3	98.6	94.3	98.0	97.9	96.4	98.4	97.5	97.1	98.2	98.7
est	2575	76.7	88.5	93.8	94.6	90.4	93.8	96.3	97.0	67.1	73.0	79.7	60.8	68.3	73.3	67.3	77.0	83.4	69.1	77.8	79.6	81.5	90.0	91.5
eus	273	91.6	94.9	97.1	96.0	95.6	98.9	96.3	98.9	92.3	96.3	95.6	88.3	93.4	92.7	92.7	96.3	91.6	88.6	94.9	96.7	91.9	98.2	93.0
fao	232	75.0	81.5	89.2	93.5	84.5	84.9	89.7	93.5	75.9	78.9	79.3	61.2	66.0	67.2	59.5	71.5	75.0	75.0	86.6	84.0	78.5	90.1	72.8
fas	2553	90.5	94.9	96.3	96.3	95.3	96.1	97.1	97.1	90.4	93.2	95.0	81.1	86.1	89.7	78.1	91.6	93.2	85.7	90.2	88.2	92.8	95.0	89.7
fin	2570	89.2	94.5	96.4	96.6	93.1	95.4	96.6	97.4	79.2	84.5	91.2	69.5	78.2	81.3	72.2	87.1	89.2	73.0	83.2	84.3	84.3	94.1	94.6
fra	2548	98.5	98.7	98.7	99.0	98.9	99.2	99.4	99.5	97.0	97.8	98.9	96.5	97.2	98.4	92.9	98.0	98.7	95.5	98.3	97.9	97.8	99.2	98.7
frm	294	84.4	95.2	98.6	97.3	91.2	96.3	98.6	96.9	92.5	96.6	97.6	81.3	90.8	96.3	88.8	95.9	97.3	90.8	95.2	97.3	96.6	98.3	92.5
fro	1976	68.9	73.5	75.2	77.8	70.8	73.5	77.2	80.0	71.5	75.0	78.7	67.9	70.6	71.5	69.2	76.3	77.4	67.6	71.8	71.6	74.4	81.5	69.1
gla	66	86.4	92.4	98.5	98.5	92.4	89.4	100.0	98.5	95.5	89.4	93.9	97.0	90.9	95.5	83.3	93.9	95.5	93.9	93.9	83.3	86.4	95.5	92.4
gle	28	71.4	71.4	82.1	78.6	64.3	75.0	67.9	71.4	82.1	71.4	82.1	85.7	85.7	71.4	82.1	78.6	75.0	85.7	78.6	71.4	60.7	60.7	67.9
glg	753	88.7	92.8	95.1	94.6	91.1	92.2	95.5	95.3	85.8	87.6	88.4	82.6	85.7	83.9	81.4	90.2	92.2	82.5	90.6	89.9	91.5	93.8	91.5
got	1579	57.4	59.2	62.6	59.3	57.1	61.2	60.8	66.2	60.0	60.9	65.7	57.6	56.5	58.3	56.7	60.0	57.8	57.8	58.3	58.2	60.1	64.1	59.8
grc	3695	71.6	80.7	87.0	88.0	79.9	86.9	90.3	91.7	76.4	80.3	86.6	68.4	75.0	78.0	67.8	81.0	85.6	69.4	78.0	74.1	79.6	91.2	82.1
guj	7	85.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	85.7
hbo	983	70.7	85.5	91.9	92.6	77.3	86.9	93.8	93.5	72.6	79.5	86.4	69.6	69.9	72.6	66.7	84.4	84.6	66.0	69.2	65.3	73.6	91.1	76.2
hbs	3286	90.9	96.5	98.3	98.2	97.1	98.1	99.0	98.9	84.5	89.8	94.0	76.8	83.4	88.7	75.0	90.9	93.0	84.6	92.4	90.8	92.9	96.5	96.6
heb	2330	81.5	87.9	88.8	92.2	86.9	88.3	89.9	91.8	77.3	84.9	88.1	73.0	75.8	78.6	64.8	79.7	81.1	72.4	79.4	73.7	76.5	88.3	88.8
hin	1447	96.4	98.6	99.2	99.1	98.6	98.5	99.1	99.0	97.6	98.2	99.3	93.8	96.8	97.4	91.0	98.8	98.5	90.8	95.8	94.2	94.3	98.8	67.0
hit	50	50.0	50.0	60.0	66.0	56.0	58.0	66.0	74.0	52.0	66.0	64.0	66.0	56.0	52.0	62.0	44.0	54.0	50.0	48.0	54.0	64.0	56.0	74.0
hsb	186	62.4	70.4	79.0	79.0	66.1	72.6	77.4	78.0	66.1	64.5	70.4	61.8	62.4	64.0	55.4	60.8	69.9	64.0	70.4	66.1	75.8	85.5	66.1
hun	845	90.2	97.8	98.5	98.8	97.5	98.6	99.3	99.2	89.3	94.2	97.0	83.4	89.6	93.8	81.8	95.7	96.3	86.3	94.6	91.7	94.9	97.9	98.2
hye	1415	93.7	96.6	97.7	98.4	97.0	98.2	98.7	99.4	87.8	94.9	96.1	83.5	92.4	94.8	84.5	89.9	97.0	88.1	94.0	89.5	93.6	97.2	66.1
hyw	1153	81.8	93.0	95.6	97.1	92.5	95.0	96.6	97.8	80.2	88.6	93.8	69.1	81.5	87.3	78.2	85.8	94.1	83.3	89.8	83.5	89.3	97.8	59.8
isl	2801	82.7	91.6	95.8	95.7	92.8	94.5	96.6	97.5	69.1	77.1	84.4	61.8	67.4	76.9	66.3	83.0	84.4	71.1	82.5	83.9	85.5	93.4	69.5
ita	2999	95.5	96.6	97.4	97.5	96.6	96.9	97.6	97.7	92.2	94.1	96.0	88.0	92.1	94.6	86.1	94.6	96.2	87.0	92.3	91.1	93.6	96.6	96.9
kat	204	95.1	98.0	94.6	94.6	96.6	95.6	96.1	96.6	92.7	92.2	98.0	89.2	92.7	95.6	80.4	89.2	95.6	88.2	93.1	89.7	92.2	95.1	92.7
kaz	173	86.1	93.1	94.2	96.0	94.8	94.8	97.7	97.7	77.5	80.3	85.5	76.9	86.1	89.6	76.9	92.5	94.2	80.9	93.6	84.4	87.3	94.8	72.8
kir	185	93.0	98.9	96.8	98.4	97.8	97.3	98.4	97.8	87.6	91.3	94.6	95.7	93.0	94.6	92.4	95.1	93.5	93.0	94.6	92.4	95.7	96.2	94.0
kmr	544	71.0	80.9	91.4	91.2	82.0	89.7	92.3	95.6	65.8	75.7	83.8	59.7	61.2	64.9	60.3	68.9	75.9	67.5	76.8	78.7	78.9	89.9	59.4
koi	43	67.4	48.8	58.1	65.1	60.5	58.1	58.1	67.4	53.5	51.2	58.1	53.5	48.8	51.2	62.8	60.5	62.8	48.8	53.5	51.2	48.8	72.1	58.1
kpv	320	58.1	62.2	67.2	70.6	62.5	67.8	77.5	78.8	59.7	62.8	65.3	52.8	54.7	53.4	59.7	57.2	53.8	58.4	58.4	58.8	58.4	82.2	53.1
krl	260	58.9	65.0	69.6	71.2	59.6	64.6	77.7	76.1	55.8	59.6	65.0	60.0	58.9	63.5	56.5	61.9	61.5	59.2	62.3	67.3	67.3	66.1	62.7
kxh	120	46.7	46.7	48.3	46.7	46.7	35.8	45.0	44.2	45.8	45.8	45.8	45.8	45.0	48.3	47.5	53.3	44.2	44.2	48.3	38.3	45.0	44.2	45.8
lat	3149	70.3	80.1	87.1	89.4	80.6	85.3	89.2	92.6	75.1	80.5	87.2	63.5	71.8	77.4	66.5	79.1	84.2	69.4	79.5	78.1	83.0	92.0	77.0
lav	3032	83.9	90.8	94.4	95.2	92.2	94.5	96.7	97.2	67.5	75.9	81.7	68.7	77.9	83.5	73.0	87.8	90.5	73.9	82.4	82.9	85.7	92.0	93.7
lij	254	72.0	75.2	77.6	82.3	75.6	74.8	80.3	77.2	74.4	72.0	74.4	66.9	68.5	66.9	61.0	73.2	74.8	75.6	72.0	69.3	75.2	76.0	58.7
lit	1180	91.0	97.5	98.7	98.4	97.3	98.3	98.6	98.9	82.3	88.7	92.4	83.8	90.1	93.1	81.7	93.7	95.9	89.4	93.6	93.0	95.2	98.1	97.5
mar	460	80.0	85.9	89.3	91.7	79.8	87.6	90.2	93.0	70.4	75.9	82.6	75.2	82.2	81.1	79.1	87.2	90.9	70.4	84.6	82.0	87.0	92.8	74.1
mdf	82	50.0	56.1	51.2	54.9	51.2	46.3	45.1	53.7	46.3	52.4	57.3	52.4	45.1	50.0	39.0	54.9	50.0	56.1	59.8	53.7	48.8	52.4	42.7
mkd	39	97.4	87.2	100.0	97.4	97.4	100.0	100.0	100.0	89.7	89.7	92.3	76.9	74.4	92.3	82.0	89.7	94.9	76.9	94.9	87.2	94.9	97.4	87.2
myv	464	56.5	54.5	50.4	51.7	56.0	55.8	61.6	61.6	55.6	52.8	51.5	56.0	55.8	53.0	51.9	53.7	53.9	55.0	57.3	54.7	51.7	56.9	54.5
nds	1774	66.4	68.9	71.2	72.8	68.6	70.3	73.8	74.8	70.1	68.4	70.4	66.7	67.6	67.8	65.9	67.2	70.2	67.6	70.2	68.5	68.2	73.3	67.9
nhi	38	55.3	47.4	52.6	44.7	39.5	57.9	52.6	57.9	47.4	50.0	50.0	50.0	47.4	44.7	55.3	50.0	44.7	50.0	52.6	50.0	52.6	50.0	47.4
nld	2331	95.1	95.5	97.5	98.2	97.0	97.0	98.3	98.5	90.4	94.3	96.2	82.8	90.8	91.9	76.5	89.6	95.6	87.7	91.7	91.0	93.5	96.4	95.8
olo	190	67.9	70.5	73.7	83.2	66.3	72.6	80.0	83.2	64.7	73.2	79.0	60.5	64.2	66.3	63.7	65.3	63.7	63.7	68.4	66.8	70.0	82.6	70.5
orv	4615	69.8	75.1	78.6	80.4	74.2	79.0	81.9	84.3	70.6	71.4	76.5	67.4	70.1	71.9	67.9	74.0	74.2	68.0	70.7	71.9	71.7	81.0	71.0
ota	99	97.0	93.9	99.0	93.9	96.0	96.0	99.0	97.0	89.9	96.0	98.0	88.9	91.9	89.9	83.8	92.9	92.9	89.9	90.9	89.9	93.9	97.0	95.0
pcm	26	92.3	100.0	100.0	100.0	96.2	92.3	96.2	96.2	96.2	92.3	84.6	96.2	100.0	100.0	96.2	92.3	100.0	92.3	96.2	96.2	96.2	96.2	100.0
pol	3272	91.4	96.1	98.2	98.6	97.2	97.8	99.0	99.2	87.6	90.7	94.0	81.4	87.1	89.3	74.7	89.2	93.7	84.5	90.8	89.2	90.3	96.6	95.0
por	3048	96.4	96.7	97.2	97.7	97.3	97.3	97.3	97.6	94.4	95.6	96.9	93.2	94.5	95.8	87.3	97.2	97.5	90.7	95.0	93.1	94.9	96.1	97.7
quc	131	58.8	63.4	62.6	65.6	68.7	68.7	65.6	72.5	70.2	71.8	69.5	71.0	67.2	76.3	63.4	68.7	67.9	74.8	76.3	70.2	62.6	67.2	73.3
ron	2056	91.7	96.3	97.9	98.0	96.5	96.9	97.8	98.7	90.0	92.2	95.5	84.6	90.2	92.9	84.1	93.5	95.6	85.6	92.8	90.1	93.5	96.8	94.3
rus	3832	95.3	97.3	98.3	98.8	97.8	98.4	99.1	99.6	91.8	94.8	97.0	89.3	93.5	95.5	84.8	96.2	97.3	91.6	95.3	94.4	95.4	98.3	97.3
sah	144	65.3	72.2	80.6	82.6	75.0	76.4	87.5	86.8	63.2	71.5	79.2	75.7	77.1	75.0	79.2	75.0	73.6	70.8	73.6	70.8	75.7	86.8	78.5
san	4442	64.2	68.0	73.8	75.9	66.9	72.6	77.1	78.6	68.3	71.9	75.4	65.3	67.0	70.1	67.2	77.0	75.8	63.5	66.5	67.5	68.5	76.2	56.8
slk	4145	83.6	92.9	95.7	96.0	94.0	96.1	97.3	98.8	72.4	80.2	85.8	69.9	77.9	82.9	65.4	84.6	90.8	74.5	82.4	82.6	83.7	92.8	94.3
slv	4483	82.9	91.5	93.9	94.7	91.3	93.0	95.8	96.5	77.0	80.9	86.1	72.3	78.7	82.9	68.8	85.7	88.9	79.9	87.1	85.7	86.5	92.3	93.1
sme	2536	66.1	69.0	72.8	74.1	67.8	72.3	76.3	80.0	65.0	66.6	66.6	65.6	66.3	64.3	63.0	65.7	65.5	65.5	68.3	66.0	69.0	77.8	66.0
sms	263	74.5	74.1	70.7	70.3	78.0	68.8	67.3	70.0	78.7	79.1	76.8	78.0	76.0	79.5	78.7	77.2	76.4	78.3	78.7	73.8	78.7	69.6	79.8
spa	2541	96.6	97.2	98.2	98.0	97.7	98.1	98.4	98.5	95.7	96.4	97.5	92.7	95.0	96.3	89.8	96.5	97.0	91.9	95.3	94.9	96.4	97.7	98.0
sqi	243	88.1	93.8	94.7	96.3	91.0	94.2	92.6	97.1	81.9	79.8	85.2	79.4	86.8	86.0	74.9	88.5	92.6	79.0	90.5	86.8	89.3	93.4	64.6
swe	201	100.0	100.0	99.5	100.0	100.0	100.0	100.0	100.0	99.5	99.5	100.0	99.0	99.0	99.5	94.5	93.5	99.0	99.5	99.5	99.5	100.0	100.0	100.0
tam	382	95.0	97.9	98.2	98.2	96.6	97.9	97.6	99.0	92.2	95.5	97.9	89.8	95.5	97.4	90.8	97.9	97.1	94.0	96.1	94.5	93.7	97.4	72.5
tpn	9	0.0	11.1	11.1	11.1	0.0	0.0	0.0	0.0	0.0	11.1	0.0	0.0	11.1	0.0	0.0	0.0	11.1	0.0	0.0	22.2	22.2	11.1	11.1
ttc	69	44.9	46.4	44.9	46.4	43.5	50.7	44.9	43.5	40.6	40.6	39.1	46.4	42.0	42.0	50.7	42.0	40.6	43.5	40.6	47.8	47.8	44.9	44.9
tur	1742	91.3	93.4	96.7	96.9	94.7	95.7	98.3	97.8	83.0	87.7	89.0	79.5	85.9	87.0	78.5	88.6	89.5	80.4	86.6	85.0	86.4	93.2	92.7
uig	758	72.8	78.5	81.3	81.4	76.5	80.7	83.0	88.5	73.2	73.0	74.4	72.6	71.0	74.9	75.7	74.9	75.5	71.9	74.3	74.9	76.2	82.1	71.6
ukr	2744	92.6	96.3	98.1	98.1	96.9	97.3	98.4	98.2	88.7	92.6	94.8	84.5	89.6	92.2	80.5	93.3	95.3	86.9	91.5	91.8	95.4	96.9	96.3
urb	13	46.2	46.2	46.2	38.5	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2
urd	550	96.0	96.4	97.1	97.3	96.0	96.5	96.7	97.5	92.0	95.8	96.2	90.0	94.5	95.3	88.7	95.6	97.6	88.5	93.6	93.3	94.7	96.7	96.5
uzb	50	94.0	100.0	100.0	100.0	100.0	98.0	100.0	100.0	84.0	90.0	94.0	90.0	94.0	94.0	96.0	98.0	98.0	90.0	94.0	98.0	100.0	100.0	86.0
vep	187	56.7	61.0	64.7	65.8	68.5	63.1	69.5	69.0	58.8	62.0	64.7	58.3	56.1	52.9	54.5	47.6	60.4	56.1	58.3	54.0	60.4	70.0	48.7
wbp	12	8.3	50.0	33.3	16.7	8.3	16.7	8.3	25.0	41.7	25.0	25.0	25.0	25.0	25.0	25.0	16.7	16.7	16.7	16.7	25.0	25.0	25.0	16.7
wol	705	81.8	82.5	84.5	84.5	84.1	84.1	85.1	89.5	83.4	83.0	82.1	83.8	83.5	82.5	78.3	85.0	84.0	88.1	86.0	86.8	82.3	89.8	82.3
xcl	1623	63.5	71.2	77.1	78.7	70.4	76.2	81.3	85.1	62.7	65.1	69.0	64.2	64.2	67.5	65.7	62.9	73.1	63.6	67.7	64.8	69.0	81.7	59.3
xnr	86	73.3	80.2	83.7	77.9	72.1	80.2	77.9	75.6	74.4	74.4	79.1	69.8	73.3	72.1	75.6	77.9	70.9	72.1	67.4	73.3	73.3	82.6	66.3
xpg	50	88.0	90.0	94.0	88.0	82.0	92.0	86.0	90.0	76.0	90.0	90.0	86.0	74.0	90.0	82.0	90.0	84.0	88.0	90.0	92.0	90.0	86.0	86.0
yrl	720	60.4	61.1	62.4	61.4	59.3	62.2	64.7	63.5	64.6	63.1	61.1	64.0	61.4	62.4	56.5	62.5	58.6	64.6	57.9	64.0	61.9	61.4	62.8

Table 6: Mean Direct accuracy (%) — PT models.

		Gemma3				Gemma4			Llama					Qwen3					Qwen3.5				TinyAya	
ISO	N	1B-it	4B-it	12B-it	27B-it	E2B-it	E4B-it	26B-A4B-it	31B-it	1B-it	3B-it	8B-it	0.6B-it	1.7B-it	4B-it	8B-it	14B-it	30B-A3B-it	0.8B-it	2B-it	4B-it	9B-it	35B-A3B-it	3.35B-it
abk	40	75.0	30.0	47.5	72.5	62.5	65.0	72.5	60.0	72.5	75.0	72.5	42.5	45.0	50.0	72.5	50.0	70.0	75.0	62.5	62.5	67.5	52.5	60.0
aln	677	62.3	62.5	75.3	74.7	65.0	68.0	65.6	67.4	68.1	66.3	66.9	68.0	68.5	69.4	70.3	66.9	68.2	69.3	73.0	73.1	74.2	77.2	71.3
amh	112	83.0	87.5	88.4	93.8	93.8	95.5	92.9	91.1	94.6	91.1	98.2	94.6	92.9	95.5	97.3	93.8	97.3	82.1	92.0	93.8	97.3	94.6	89.3
apu	28	57.1	85.7	82.1	71.4	89.3	75.0	71.4	64.3	96.4	85.7	96.4	96.4	92.9	100.0	96.4	96.4	96.4	96.4	96.4	96.4	96.4	96.4	85.7
aqz	14	28.6	50.0	28.6	28.6	21.4	42.9	50.0	28.6	28.6	21.4	50.0	28.6	35.7	35.7	28.6	50.0	21.4	28.6	50.0	21.4	57.1	57.1	50.0
arb	1215	78.2	78.1	84.8	90.3	79.9	84.4	81.7	88.4	82.1	84.9	88.9	81.9	86.1	87.7	90.5	89.8	90.2	85.3	87.8	90.1	92.5	93.2	90.0
azz	207	64.7	65.7	69.6	64.2	66.7	70.5	73.4	61.8	62.3	67.6	70.0	68.1	63.8	69.1	70.0	67.6	69.6	66.7	68.1	66.2	65.2	73.4	73.9
bel	2570	71.0	81.7	88.5	91.8	73.9	79.8	79.8	83.8	63.5	73.4	86.0	63.8	73.0	77.2	81.7	84.5	85.1	78.6	84.7	89.1	91.8	92.8	80.2
ben	21	76.2	95.2	85.7	85.7	85.7	81.0	90.5	90.5	90.5	100.0	100.0	76.2	95.2	95.2	100.0	100.0	100.0	76.2	95.2	95.2	100.0	95.2	47.6
bho	34	52.9	82.3	73.5	70.6	79.4	79.4	85.3	85.3	58.8	52.9	67.7	58.8	64.7	64.7	73.5	79.4	64.7	55.9	67.7	73.5	76.5	79.4	70.6
bor	241	62.7	58.5	58.1	66.8	63.9	59.8	58.5	62.2	58.9	64.3	59.8	66.4	64.7	56.9	62.7	61.0	56.9	64.3	64.3	62.2	62.7	64.7	65.6
bre	260	36.5	73.5	91.1	90.4	55.0	65.0	72.3	67.7	85.8	86.9	94.6	55.0	60.4	66.9	74.2	85.0	81.5	87.7	93.5	93.5	94.2	92.7	55.0
bua	103	70.9	68.0	68.9	71.8	67.0	72.8	71.8	71.8	69.9	66.0	67.0	67.0	67.0	72.8	67.0	65.0	71.8	68.0	66.0	68.9	72.8	76.7	68.9
bul	2458	80.8	87.5	92.7	94.1	82.0	82.9	85.6	86.1	76.3	83.9	89.7	79.1	82.6	89.4	92.2	93.1	93.1	83.6	88.0	91.0	92.6	95.4	93.4
cat	2284	80.6	85.8	91.4	93.0	80.5	85.7	82.5	86.8	79.8	86.1	92.2	80.8	87.6	90.9	91.6	92.7	92.2	86.9	90.1	93.2	95.3	94.8	92.2
ces	4256	72.3	82.5	88.9	90.9	75.7	82.0	79.7	83.8	70.8	80.2	89.0	72.3	78.8	82.9	88.8	90.5	88.9	77.2	83.1	86.8	90.4	93.1	89.5
chu	4166	60.7	58.8	65.8	63.0	60.9	61.6	63.4	63.0	59.6	60.8	60.8	61.0	60.9	59.6	61.9	61.5	61.6	59.3	61.6	61.7	63.9	67.1	59.7
cym	1120	65.4	76.7	85.5	88.3	74.1	79.8	82.9	85.1	75.5	77.4	86.5	75.1	77.2	80.6	86.2	86.5	88.8	76.2	78.8	84.7	87.5	91.6	90.5
dan	50	88.0	86.0	94.0	100.0	92.0	94.0	84.0	94.0	94.0	94.0	98.0	92.0	94.0	92.0	96.0	96.0	100.0	94.0	98.0	96.0	98.0	98.0	92.0
deu	2298	89.0	90.0	94.1	95.7	90.2	91.3	90.1	92.6	91.6	94.1	96.4	91.4	93.0	94.4	96.8	95.8	96.3	92.7	95.5	96.7	98.3	98.0	94.9
egy	22	45.5	50.0	40.9	40.9	45.5	50.0	45.5	45.5	40.9	45.5	45.5	40.9	45.5	50.0	54.5	50.0	50.0	40.9	45.5	40.9	50.0	45.5	40.9
ell	1096	87.2	88.8	95.3	97.0	89.0	89.8	90.1	90.2	86.6	90.8	95.8	85.6	91.8	94.2	95.3	97.1	95.1	93.1	96.2	96.8	98.3	98.1	96.4
eng	770	79.9	86.5	96.1	96.2	92.5	96.0	93.5	93.5	94.0	96.2	98.3	96.8	90.8	94.0	96.1	94.5	95.1	97.3	97.4	98.4	98.4	98.0	96.4
est	2575	60.7	76.5	85.4	89.0	63.9	68.3	73.0	76.3	59.5	66.9	76.3	58.7	64.7	70.2	74.6	77.4	80.6	66.0	71.6	79.6	85.0	86.4	84.2
eus	273	83.2	88.3	93.8	89.0	86.8	88.6	85.4	81.0	87.2	93.4	95.6	87.2	83.5	86.1	91.2	91.6	92.7	94.5	92.7	93.4	95.6	97.8	94.5
fao	232	65.5	69.0	80.6	88.8	69.8	75.0	69.0	75.4	66.8	74.6	80.2	59.1	61.6	75.0	75.4	74.1	74.1	73.7	80.2	87.5	87.5	89.7	69.4
fas	2553	84.2	90.4	93.9	95.5	90.1	92.6	92.2	92.9	82.1	87.2	93.4	73.1	80.0	87.4	91.3	93.3	92.7	87.4	91.5	92.7	95.1	96.0	87.7
fin	2570	77.3	86.5	91.2	92.6	78.0	84.5	84.4	87.0	73.7	81.0	88.4	66.6	76.6	80.9	86.6	86.8	87.5	72.6	81.7	87.7	91.2	92.1	89.7
fra	2548	93.0	93.2	95.2	97.3	94.2	95.9	95.3	95.0	94.8	96.4	97.1	94.1	95.0	96.3	97.8	97.0	97.5	94.4	96.5	97.8	98.7	98.5	95.0
frm	294	67.3	77.2	91.2	93.5	76.2	81.6	87.4	87.4	82.3	91.2	90.8	78.9	84.7	89.8	93.9	93.9	91.8	85.0	92.2	94.2	95.9	98.6	85.7
fro	1976	61.3	63.6	69.0	71.4	64.1	66.0	63.6	68.2	65.0	68.7	75.7	67.7	64.5	64.3	70.5	71.0	70.8	66.8	68.7	71.0	72.9	76.1	60.3
gla	66	87.9	84.9	87.9	93.9	87.9	90.9	92.4	84.9	90.9	93.9	92.4	95.5	92.4	92.4	97.0	95.5	92.4	90.9	95.5	97.0	90.9	95.5	95.5
gle	28	46.4	75.0	75.0	67.9	71.4	46.4	60.7	75.0	75.0	75.0	89.3	78.6	75.0	67.9	71.4	71.4	67.9	78.6	78.6	50.0	71.4	64.3	53.6
glg	753	71.7	78.2	85.8	87.2	70.7	74.6	76.1	78.9	78.8	86.5	87.2	82.2	76.9	79.0	85.1	86.5	87.4	82.7	89.4	89.9	90.8	91.5	91.4
got	1579	54.8	55.7	60.4	58.4	57.4	57.8	56.8	58.8	56.9	56.4	61.5	56.2	56.2	58.0	57.8	56.7	55.9	58.0	58.3	59.4	59.9	62.3	56.9
grc	3695	59.2	66.8	74.4	78.8	71.6	72.0	72.3	75.9	67.8	73.9	84.8	63.7	71.3	74.8	79.7	83.1	80.1	70.5	76.2	82.2	85.3	88.1	75.6
guj	7	85.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	85.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	85.7
hbo	983	58.9	66.5	76.4	82.9	66.2	70.1	66.1	76.6	68.8	73.6	82.8	64.2	68.0	72.1	76.6	81.2	81.2	67.8	72.2	78.8	83.3	92.0	69.2
hbs	3286	74.8	87.6	92.1	94.8	80.4	85.5	86.5	89.1	75.5	83.4	91.1	74.2	80.4	85.7	89.0	91.0	90.5	83.7	89.1	92.8	94.2	95.8	92.2
heb	2330	68.8	76.5	81.4	84.3	71.2	74.0	72.4	74.1	63.3	70.7	85.9	70.3	73.0	75.3	78.9	82.4	77.7	75.1	77.1	81.4	84.4	85.5	83.5
hin	1447	92.3	96.1	98.4	98.6	97.4	98.0	96.9	98.1	95.0	97.2	98.5	87.7	94.0	95.7	96.9	98.1	98.1	90.3	94.6	97.2	97.9	98.3	66.7
hit	50	40.0	58.0	42.0	52.0	48.0	54.0	62.0	54.0	50.0	62.0	56.0	56.0	58.0	54.0	60.0	64.0	58.0	48.0	56.0	40.0	58.0	52.0	72.0
hsb	186	52.7	60.2	71.5	78.5	57.0	58.6	49.5	57.5	60.8	64.5	64.5	58.6	58.1	60.8	55.9	61.3	66.1	63.4	73.7	71.5	78.0	82.3	67.2
hun	845	77.5	87.0	92.9	93.4	84.0	87.6	85.9	88.8	84.0	90.1	94.9	78.8	85.8	91.7	92.8	94.7	94.0	83.4	89.5	94.0	96.2	96.7	93.5
hye	1415	70.7	84.6	92.0	93.5	87.8	90.5	86.8	89.8	70.1	83.6	94.2	73.4	86.6	92.3	93.3	94.3	94.4	87.1	91.8	95.3	96.7	96.5	59.1
hyw	1153	62.7	75.6	84.9	90.7	78.8	82.8	79.8	83.1	62.5	72.6	88.3	66.7	78.1	83.2	86.5	88.9	87.9	82.5	87.2	93.1	95.1	96.6	58.8
isl	2801	63.1	74.7	86.4	90.2	73.4	77.6	79.5	82.2	63.2	71.5	83.1	63.7	65.0	72.6	80.1	81.0	80.9	69.3	80.5	86.5	91.4	92.2	71.1
ita	2999	83.2	84.0	90.3	92.4	84.4	88.3	85.6	86.9	87.0	91.1	93.5	83.5	86.7	90.3	92.6	92.8	92.6	86.1	89.9	92.3	95.0	94.8	89.3
kat	204	73.5	89.7	86.3	93.1	95.1	90.2	85.3	85.8	90.2	89.7	91.2	85.8	90.2	90.2	96.1	98.0	95.1	87.8	88.2	91.7	91.2	94.6	89.2
kaz	173	76.3	89.6	94.2	93.1	89.0	90.2	90.8	94.8	72.2	76.9	89.0	73.4	83.8	89.0	93.1	90.8	93.1	82.7	86.7	90.2	92.5	94.2	84.4
kir	185	86.0	85.4	90.3	95.1	89.2	83.8	92.4	95.1	89.2	89.2	95.1	94.0	90.3	88.1	85.4	89.7	90.8	84.9	91.9	96.2	96.8	93.5	90.3
kmr	544	57.2	73.7	81.4	80.7	67.5	74.3	79.0	80.0	60.7	69.8	81.1	62.5	63.6	67.7	70.0	77.9	76.5	65.6	73.5	78.5	84.9	88.6	58.3
koi	43	53.5	55.8	55.8	60.5	62.8	60.5	62.8	65.1	58.1	41.9	51.2	48.8	51.2	51.2	58.1	51.2	55.8	51.2	55.8	46.5	55.8	62.8	62.8
kpv	320	55.0	58.4	57.2	62.2	61.6	50.9	55.3	65.6	58.8	60.6	61.3	52.2	54.4	54.1	55.6	56.9	55.0	55.3	60.3	64.7	71.2	77.8	56.2
krl	260	51.5	60.8	63.5	65.8	55.4	61.5	48.9	57.7	56.1	61.9	63.8	58.1	55.8	58.9	60.8	55.4	61.2	60.4	52.7	58.9	59.6	65.4	63.8
kxh	120	35.0	59.2	45.0	48.3	40.8	44.2	44.2	45.8	40.0	47.5	45.8	39.2	41.7	47.5	49.2	48.3	51.7	48.3	43.3	45.0	46.7	44.2	38.3
lat	3149	60.3	68.5	77.9	80.7	64.6	71.6	72.7	76.3	69.3	75.5	82.9	60.9	65.1	74.3	78.8	80.8	80.6	67.3	76.4	81.1	87.1	89.1	66.6
lav	3032	65.4	81.7	87.0	89.1	69.6	72.2	75.4	79.0	64.3	67.6	77.8	65.2	76.4	80.4	85.2	88.0	85.8	72.2	79.2	85.0	88.4	91.8	86.9
lij	254	59.5	63.0	65.3	74.0	62.2	64.6	62.6	63.0	67.7	71.7	74.0	65.0	70.1	66.9	67.7	74.8	69.3	73.2	75.2	71.7	73.6	77.6	51.6
lit	1180	76.9	88.6	93.3	94.8	80.9	85.8	83.3	87.5	74.5	82.3	89.9	80.0	86.4	91.5	92.5	95.2	94.2	87.8	90.1	93.6	96.7	96.6	94.6
mar	460	74.8	79.1	87.6	88.7	79.1	80.4	83.0	82.0	64.6	72.4	83.5	76.7	79.3	80.0	86.7	86.5	90.0	75.0	82.8	82.6	87.2	88.0	77.0
mdf	82	52.4	52.4	52.4	53.7	40.2	52.4	52.4	45.1	50.0	54.9	50.0	50.0	47.6	57.3	52.4	45.1	48.8	45.1	56.1	58.5	56.1	52.4	56.1
mkd	39	74.4	87.2	89.7	92.3	79.5	87.2	74.4	89.7	76.9	74.4	87.2	71.8	66.7	87.2	79.5	92.3	79.5	89.7	87.2	87.2	100.0	100.0	87.2
myv	464	55.2	49.8	51.3	52.2	53.7	58.4	53.2	50.4	55.8	55.0	51.9	53.9	56.2	56.2	54.3	50.9	56.0	54.1	55.6	48.9	55.0	53.0	55.8
nds	1774	62.8	64.1	67.9	68.2	64.4	66.3	64.1	68.5	65.9	65.3	70.0	64.9	65.0	64.2	66.6	67.6	66.1	67.9	68.5	69.2	69.8	73.2	67.9
nhi	38	44.7	47.4	42.1	47.4	42.1	42.1	52.6	47.4	44.7	50.0	47.4	39.5	47.4	42.1	57.9	47.4	44.7	47.4	52.6	44.7	42.1	42.1	39.5
nld	2331	84.7	88.0	93.5	94.8	84.6	88.4	86.4	89.3	83.9	88.3	93.4	78.6	86.5	89.5	92.6	93.8	92.5	83.4	89.1	93.1	94.8	95.8	89.6
olo	190	55.3	56.3	70.0	70.5	52.6	57.9	55.3	58.4	60.0	70.0	76.8	62.1	60.5	63.2	61.1	61.6	57.9	61.6	65.8	66.8	72.6	76.3	64.2
orv	4615	63.6	65.7	72.9	74.7	65.7	66.4	67.3	70.3	65.0	68.0	74.0	64.8	67.7	69.4	71.1	71.4	70.9	68.5	70.5	73.5	76.9	79.7	67.6
ota	99	79.8	86.9	95.0	93.9	90.9	89.9	92.9	91.9	90.9	83.8	93.9	88.9	87.9	87.9	93.9	96.0	95.0	90.9	85.9	89.9	97.0	95.0	92.9
pcm	26	84.6	80.8	76.9	84.6	88.5	88.5	92.3	84.6	96.2	92.3	92.3	100.0	96.2	92.3	88.5	96.2	96.2	100.0	100.0	88.5	96.2	88.5	100.0
pol	3272	79.2	88.6	92.3	94.2	85.5	88.9	88.6	90.6	79.9	85.4	90.8	76.7	84.2	88.1	90.7	92.0	91.5	84.2	88.8	92.2	94.6	94.9	90.0
por	3048	85.3	86.9	91.7	92.6	87.5	91.2	90.2	89.7	89.5	93.6	94.9	89.6	90.2	91.6	94.9	93.7	93.8	89.8	92.1	94.4	95.7	95.2	92.5
quc	131	57.2	59.5	70.2	64.9	68.7	58.8	66.4	70.2	65.6	68.7	69.5	72.5	60.3	58.8	68.7	67.2	66.4	64.9	73.3	67.2	67.9	69.5	67.2
ron	2056	82.1	88.3	91.5	94.3	80.2	84.8	85.4	86.7	83.4	88.3	90.9	81.6	87.4	89.6	93.5	93.0	92.1	83.5	89.1	92.5	94.6	95.6	91.6
rus	3832	86.9	90.1	94.1	95.6	88.9	91.9	92.5	92.8	82.7	89.7	94.9	86.8	89.6	92.3	94.5	94.8	94.5	89.8	93.1	95.0	97.2	97.4	92.9
sah	144	68.8	71.5	71.5	74.3	80.6	66.0	76.4	75.7	77.8	74.3	75.7	76.4	75.7	75.0	77.8	68.1	67.4	70.1	75.0	70.8	86.8	81.9	79.9
san	4442	58.3	61.3	64.6	67.4	63.7	68.6	67.8	67.7	63.1	66.2	71.4	61.1	64.4	70.6	70.2	71.8	71.1	62.4	65.7	68.9	72.0	74.8	58.3
slk	4145	66.5	83.1	88.0	89.9	72.5	78.1	78.1	83.2	62.8	71.4	81.8	67.0	74.6	81.0	85.1	88.8	87.9	72.9	79.5	85.2	90.2	92.2	86.3
slv	4483	65.3	78.2	85.7	88.7	71.0	74.2	73.5	77.6	68.8	75.6	82.2	70.2	76.1	80.1	82.4	86.1	85.8	78.0	82.5	85.6	88.8	91.5	87.9
sme	2536	59.5	61.4	67.9	69.4	61.8	63.4	62.3	65.2	62.1	66.8	68.1	66.4	63.4	64.5	63.5	66.8	64.2	67.8	68.0	69.2	73.3	77.6	62.4
sms	263	73.0	67.3	70.7	62.4	73.4	69.6	65.8	68.8	76.4	78.0	78.0	75.7	78.0	74.5	79.5	78.0	74.5	78.7	76.4	77.6	74.9	71.1	77.2
spa	2541	86.9	88.9	93.7	94.6	88.5	91.5	91.6	91.2	92.1	93.9	96.1	91.1	92.2	93.3	95.2	94.9	95.1	91.4	94.4	95.7	96.5	97.1	93.9
sqi	243	62.5	79.8	86.8	89.3	74.5	82.3	70.8	81.9	69.1	70.8	80.7	73.7	78.6	83.1	89.3	85.6	91.4	81.9	84.0	89.3	92.6	89.7	81.1
swe	201	95.5	95.0	97.0	99.5	95.5	97.5	96.0	91.5	96.5	99.0	99.5	96.5	96.0	95.0	99.0	97.5	98.0	99.0	99.0	98.5	99.0	99.0	100.0
tam	382	93.2	96.3	97.6	97.6	92.7	95.8	93.2	95.0	96.1	95.3	98.2	84.5	94.8	94.8	96.6	96.9	97.1	95.0	96.1	96.3	96.9	96.9	74.4
tpn	9	11.1	11.1	44.4	0.0	0.0	0.0	0.0	11.1	0.0	11.1	11.1	0.0	0.0	11.1	22.2	0.0	11.1	11.1	11.1	0.0	0.0	0.0	11.1
ttc	69	39.1	40.6	46.4	49.3	43.5	53.6	44.9	42.0	43.5	46.4	46.4	47.8	42.0	42.0	46.4	42.0	42.0	44.9	43.5	37.7	46.4	47.8	46.4
tur	1742	79.8	85.1	92.2	93.3	84.6	90.3	88.7	92.1	77.0	81.7	85.8	76.9	83.1	87.4	88.6	89.5	88.5	80.0	86.3	88.8	92.4	93.9	89.1
uig	758	66.2	74.0	75.1	79.8	67.7	71.8	71.1	81.0	73.5	71.5	75.9	72.0	68.9	73.5	74.4	78.5	74.5	69.4	73.4	75.5	77.2	82.1	69.5
ukr	2744	84.7	87.8	92.6	95.4	86.7	90.5	87.5	88.4	73.1	83.4	92.5	79.6	86.2	89.6	91.4	93.2	91.8	86.4	90.7	93.2	95.7	96.1	93.6
urb	13	38.5	53.8	38.5	46.2	38.5	38.5	46.2	46.2	38.5	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	46.2	53.8
urd	550	84.5	92.0	94.5	96.5	92.4	96.0	92.5	94.0	89.5	93.3	95.8	88.5	91.6	93.3	96.7	96.5	96.2	87.3	93.6	94.5	96.0	96.0	95.5
uzb	50	90.0	96.0	96.0	100.0	88.0	96.0	94.0	98.0	82.0	86.0	90.0	86.0	88.0	90.0	96.0	94.0	98.0	88.0	96.0	94.0	98.0	100.0	72.0
vep	187	55.1	57.2	58.8	57.2	54.5	60.4	52.9	57.8	54.0	55.6	62.0	50.8	46.0	58.3	54.0	53.5	59.9	56.7	57.8	61.5	62.6	63.6	55.6
wbp	12	50.0	66.7	25.0	16.7	16.7	33.3	33.3	25.0	16.7	8.3	8.3	25.0	41.7	16.7	25.0	25.0	25.0	33.3	8.3	33.3	16.7	25.0	16.7
wol	705	79.3	77.3	74.0	78.0	76.6	77.3	80.0	81.3	81.1	81.4	81.4	84.7	80.1	77.2	79.3	78.6	78.4	88.4	84.1	82.8	85.2	88.1	87.2
xcl	1623	55.0	62.1	63.8	72.6	63.0	65.5	64.9	69.5	57.4	58.9	66.6	61.9	63.3	66.1	67.1	69.4	67.0	63.5	65.9	69.8	73.1	76.3	57.6
xnr	86	69.8	70.9	77.9	80.2	76.7	68.6	80.2	80.2	66.3	66.3	75.6	69.8	74.4	69.8	72.1	74.4	69.8	76.7	72.1	70.9	73.3	82.6	66.3
xpg	50	92.0	92.0	90.0	80.0	92.0	90.0	82.0	92.0	72.0	82.0	90.0	88.0	76.0	84.0	80.0	86.0	84.0	88.0	88.0	90.0	86.0	82.0	88.0
yrl	720	57.4	55.4	61.1	61.1	58.2	60.1	64.7	64.6	59.2	60.1	58.9	59.3	53.3	58.2	59.6	60.1	55.4	62.6	63.3	63.1	60.3	56.7	55.3

D.1 CC-100 Frequency and Base Accuracy

Table 7 reports within-family OLS regression results predicting base accuracy from log-transformed CC-100 corpus size and model scale. The cross-family mixed-effects model (with model as random

intercept) is reported in §4.2.

Table 7: Within-family OLS: $\text{Accuracy} \sim \log_{10}(\text{CC count}) + \text{Size\_Order}$. All models are PT, Direct method.

Family	N	β_{CC}	p	R^2
Gemma 3	356	0.0610	< 0.001	0.653
Gemma 4	356	0.0613	< 0.001	0.628
Llama 3.1	89	0.0578	< 0.001	0.629
Llama 3.2	178	0.0535	< 0.001	0.569
Qwen 3	534	0.0597	< 0.001	0.571
Qwen 3.5	445	0.0555	< 0.001	0.591
TinyAya	89	0.0619	< 0.001	0.538
Cross-family MixedLM (model as random intercept): $\beta_{\text{CC}} = 0.059$, $p < 0.001$; ICC $= 0.078$				

D.2 Gap by Resource Tier and Family

Table 8 reports the mean inarticulate gap for the largest model in each family, broken down by resource tier.

Table 8: Mean inarticulate gap by resource tier and family (largest model only).

Family	Super High	High	Mid	Low/Zero
Gemma 3 (27B)	0.038	0.044	0.094	0.047
Gemma 4 (31B)	-0.066	-0.091	-0.044	-0.042
Llama (8B)	0.233	0.272	0.292	0.120
Qwen 3 (30B-A3B)	0.074	0.088	0.133	0.085
Qwen 3.5 (35B-A3B)	0.003	0.004	0.020	0.039
TinyAya (3.35B)	0.277	0.244	0.209	0.129

	Size	Variant	*Resource Tier*			
			Low Zero	**Mid**	**High**	**Super High**
Gemma 3	1.0	PT	3.50	15.32	19.83	22.85
		base	12.13	23.98	40.97	47.25
	4.0	PT	6.20	12.08	9.91	7.93
		base	11.52	28.72	25.67	27.41
	12.0	PT	4.49	10.54	4.32	6.22
		base	13.70	24.00	19.13	11.79
	27.0	PT	4.66	9.36	4.43	3.81
		base	11.86	16.22	8.18	6.97
Gemma 4	2.0	PT	6.78	15.91	8.28	7.76
		base	13.66	30.11	44.67	44.34
	4.0	PT	7.68	16.57	13.10	9.23
		base	13.43	18.66	21.99	15.80
	26.0	PT	1.81	-1.41	-8.38	-5.79
		base	8.38	10.84	6.51	4.65
	31.0	PT	**-4.20**	**-4.38**	**-9.10**	**-6.61**
		base	-0.59	3.69	1.70	0.33
Llama 3	1.0	PT	9.38	23.35	28.63	37.50
		base	11.88	27.01	35.89	45.22
	3.0	PT	8.47	24.40	28.28	30.07
		base	14.73	28.03	37.30	41.42
	8.0	PT	12.00	29.17	27.15	23.28
		base	14.00	31.09	33.98	35.24
Qwen 3	0.6	PT	10.75	21.53	25.82	32.29
		base	11.84	21.92	29.38	38.08
	1.7	PT	7.88	19.41	22.69	18.11
		base	10.59	23.44	23.98	22.24
	4.0	PT	9.28	24.77	28.64	32.81
		base	4.63	19.29	17.42	11.77
	8.0	PT	12.72	27.32	38.78	43.29
		base	-0.93	2.60	4.52	4.97
	14.0	PT	3.97	11.57	6.40	4.88
		base	5.32	10.32	5.31	2.66
	30.0	PT	8.47	13.25	8.82	7.40
		base	7.44	12.89	8.53	7.98
Qwen 3.5	0.8	PT	12.17	22.13	24.36	30.00
		base	11.87	21.33	21.63	23.27
	2.0	PT	11.48	22.79	17.37	14.72
		base	10.67	21.21	18.88	17.69
	4.0	PT	8.68	13.69	11.09	10.91
		base	6.48	8.24	5.88	4.88
	9.0	PT	5.75	5.49	5.37	4.85
		base	4.88	-1.61	-0.63	0.19
	35.0	PT	3.90	2.02	0.39	0.30
		base	1.97	2.83	0.56	1.10
Tiny-Aya	3.35	PT	12.91	20.94	24.37	27.70
		base	9.97	23.68	29.59	33.08

Table 9: Mean inarticulate gap ($\mathrm{Acc}_{\mathrm{Direct}} - \mathrm{Acc}_{\mathrm{Meta}}$, %) by resource tier for all model variants. Bold=lowest gap per column.

Native Prompts Used in Native-Prompted Evaluation

Table: Native-language prompts used as language cues in the native-prompted evaluation method, for all 101 languages in MultiBLiMP. Rows highlighted in yellow indicate languages where a direct translation of the English template ("This is a sentence in [Language].") was not available; a representative sentence in the target language was used as a language cue instead.

Code	Language	Native Prompt	Note
abk	Abkhaz	Ари аҳәоу абаза бызшәала иҟоуп.	
aln	Gheg Albanian	Kjo asht nji fjali në gegnisht.	
amh	Amharic	ይህ ዓረፍተ ነገር በአማርኛ ነው።	
apu	Apurinã	Iiye iatiki Apurinã sã.	
aqz	Akuntsu	I'ãẽũ õ ita akuntsú.	Language cue (no direct translation available)
arb	Arabic	هذه جملة باللغة العربية.	
azz	Highland Puebla Nahuatl	Inin tlahtolli tlen nahuatlahtolli.	
bel	Belarusian	Гэта сказ на беларускай мове.	
ben	Bengali	এটি বাংলা ভাষার একটি বাক্য।	
bho	Bhojpuri	ई भोजपुरी भाषा के वाक्य ह।	
bor	Borôro	Awu Boe barae.	Language cue (no direct translation available)
bre	Breton	Ur frazenn e brezhoneg eo homañ.	
bua	Buryat	Энэ буряад хэлэн дээрэ бэшэгдэһэн мэдүүлэл.	
bul	Bulgarian	Това е изречение на български език.	
cat	Catalan	Aquesta és una frase en català.	
ces	Czech	Toto je věta v češtině.	
chu	Church Slavonic	Сє ѥстъ словѣньскъ ѩꙁыкъ.	Language cue (ancient language; no direct translation available)
cym	Welsh	Dyma frawddeg yn y Gymraeg.	
dan	Danish	Dette er en sætning på dansk.	
deu	German	Dies ist ein Satz auf Deutsch.	
egy	Ancient Egyptian	𓇋𓊪𓈖 𓅓 𓂋𓏤𓈖 𓆎𓅓𓏏𓊖	Language cue (ancient language; no direct translation available)
ell	Modern Greek	Αυτή είναι μια πρόταση στα ελληνικά.	
eng	English	The following is a sentence in English.	
est	Estonian	See on lause eesti keeles.	
eus	Basque	Hau euskarazko esaldi bat da.	
fao	Faroese	Hetta er ein setningur á føroyskum.	
fas	Persian	این جمله‌ای به زبان فارسی است.	
fin	Finnish	Tämä on lause suomeksi.	
fra	French	Ceci est une phrase en français.	
frm	Middle French	Cestui est un devis en moyen françois.	

Code	Language	Native Prompt	Note
fro	Old French	Cest est une phrase en ancien français.	
gla	Scottish Gaelic	Is e seo seantans ann an Gàidhlig.	
gle	Irish	Is abairt í seo i nGaeilge.	
glg	Galician	Esta é unha frase en galego.	
got	Gothic	𐌷𐌰𐌸𐌰 𐌹𐌳𐌲 𐌹𐌽𐌳𐌴𐌹𐌽𐌰𐌼𐌰𐌽𐌽𐌰𐌽.	Language cue (ancient language; no direct translation available)
grc	Ancient Greek	Τοῦτό ἐστι ῥῆμα Ἑλληνιστί.	
guj	Gujarati	આ ગુજરાતી ભાષાનું વાક્ય છે.	
hbo	Ancient Hebrew	זֶה מִשְׁפָּט בִּלְשׁוֹן הַקֹּדֶשׁ.	
hbs	Serbo-Croatian	Ovo je rečenica na srpskohrvatskom jeziku.	
heb	Hebrew	זהו משפט בעברית.	
hin	Hindi	यह हिंदी भाषा का एक वाक्य है।	
hit	Hittite	Kī uttar nešumnili.	Language cue (ancient language; no direct translation available)
hsb	Upper Sorbian	To je sadźba w hornjoserbšćinje.	
hun	Hungarian	Ez egy mondat magyarul.	
hye	Eastern Armenian	Սա նախադասություն է հայերենով:	
hyw	Western Armenian	Ասիկա արեւմտահայերէն նախադասութիւն մըն է:	
isl	Icelandic	Þetta er setning á íslensku.	
ita	Italian	Questa è una frase in italiano.	
kat	Georgian	ეს არის წინადადება ქართულ ენაზე.	
kaz	Kazakh	Бұл қазақ тіліндегі сөйлем.	
kir	Kyrgyz	Бул кыргыз тилиндеги сүйлөм.	
kmr	Kurmanji	Ev hevokek bi Kurmancî ye.	
koi	Komi-Permyak	Тайö сёрникузяыс перем коми кыв вылын.	
kpv	Komi-Zyrian	Тайö сёрникузяыс коми кыв вылын.	
krl	Karelian	Tämä on virke karjalaksi.	
kxh	Karo	Una ałqsianek Kusquqviim.	Language cue (no direct translation available)
lat	Latin	Haec sententia Latine scripta est.	
lav	Latvian	Šis ir teikums latviešu valodā.	
lij	Ligurian	Sta chi a l'é 'na frâze in lìgure.	
lit	Lithuanian	Tai yra sakinys lietuvių kalba.	
mar	Marathi	हे मराठी भाषेतील एक वाक्य आहे.	
mdf	Moksha	Тя валрисьмось мокшень кяльса.	
mkd	Macedonian	Ова е реченица на македонски јазик.	
myv	Erzya	Те валрисьмесь эрзянь кельсэ.	
nds	Low German	Dat is en Satz op Plattdüütsch.	
nhi	Zacatlán-Ahuacatlán-Tepetzintla Nahuatl	Ni se tlahtoli ika nawatlahtoli.	
nld	Dutch	Dit is een zin in het Nederlands.	

Code	Language	Native Prompt	Note
olo	Livvi-Karelian	Tämä on virke livvinkarjalaksi.	
orv	Old Russian	Се ѥстъ слово словѣньско.	Language cue (ancient language; no direct translation available)
ota	Ottoman Turkish	بو جمله لسان عثمانیدەدر	
pcm	Nigerian Pidgin	Dis na sentence for Pidgin.	
pol	Polish	To jest zdanie w języku polskim.	
por	Portuguese	Esta é uma frase em português.	
quc	K'iche'	Are wa' jun tzij pa K'iche'.	
ron	Romanian	Aceasta este o propoziție în română.	
rus	Russian	Это предложение на русском языке.	
sah	Yakut	Бу сахалыы этии.	
san	Sanskrit	इदं संस्कृतवाक्यमस्ति।	
slk	Slovak	Toto je veta v slovenčine.	
slv	Slovenian	To je stavek v slovenščini.	
sme	Northern Sami	Dát lea cealkka sámegillii.	
sms	Skolt Sami	Tät lij sää′mkiõllsaž ciälkk.	
spa	Spanish	Esta es una oración en español.	
sqi	Albanian	Kjo është një fjali në shqip.	
swe	Swedish	Detta är en mening på svenska.	
tam	Tamil	இது தமிழ் வாக்கியம்.	
tpn	Tupinambá	Kó nhe'eng Tupinambá.	Language cue (no direct translation available)
ttc	Tektiteko	Ba'n qyol Tektiteko.	Language cue (no direct translation available)
tur	Turkish	Bu, Türkçe bir cümledir.	
uig	Uyghur	بۇ ئۇيغۇر تىلىدىكى جۈملە.	
ukr	Ukrainian	Це речення українською мовою.	
urb	Ka'apor	Kó ne'e Ka'apor.	Language cue (no direct translation available)
urd	Urdu	یہ اردو زبان کا ایک جملہ ہے۔	
uzb	Uzbek	Bu o'zbek tilidagi jumla.	
vep	Veps	Nece om lause vepsän kelel.	
wbp	Warlpiri	Nyampu-ju Warlpiri.	Language cue (no direct translation available)
wol	Wolof	Lii benn kàddu Wolof la.	
xcl	Classical Armenian	Այս է բան հայերէն:	
xnr	Kangri	इह कांगड़ी च बाक है।	
xpg	Phrygian	ιος νι σεμουν κνουμανεκακον.	Language cue (ancient language; no direct translation available)
yrl	Nheengatu	Kwayé nhe'eng Nheengatu.	Language cue (no direct translation available)

Yellow rows (13 languages): egy, hit, got, chu, orv, xpg, bor, wbp, tpn, urb, yrl, aqz, kxh, ttc — a representative sentence in the target language was used as a language cue in place of a direct translation.