Improving Cross-Lingual Factual Recall via Consistency-Driven Reinforcement Learning

Anonymous EMNLP submission

Abstract

Large language models (LLMs) trained predominantly on English data encode substantial world knowledge, yet often fail to express it reliably in other languages, known as *cross-lingual factual inconsistency*. To study and address this, we introduce POLYFACT, a large-scale parallel multilingual factual QA dataset of 100K Wikidata-grounded facts across 12 typologically diverse languages, and use it to compare light continual pretraining (CPT), supervised fine-tuning (SFT), and reinforcement learning via Group Relative Policy Optimization (GRPO) for improving cross-lingual factual recall in Qwen-2.5-7B and OLMo-2-1124-7B. GRPO consistently outperforms SFT, improving both cross-lingual consistency and generalization to unseen languages, while CPT on parallel data yields limited additional gains. Mechanistic analyses reveal that GRPO reorganizes multilingual routing by reducing language specialization in MLP layers and attention heads and promoting shared cross-lingual representations. We release our code, models, and dataset upon publication.

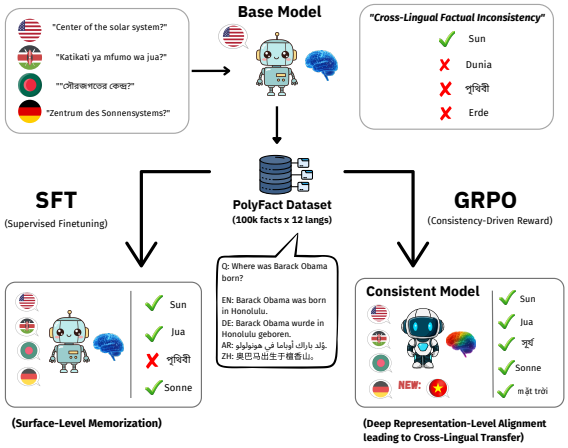


Figure 1: Incentivizing cross-lingual factual consistency through post-training on our multilingual question-answering dataset, POLYFACT. GRPO-based reinforcement learning promotes shared internal representations that yield consistent factual predictions across languages, whereas SFT primarily leads to surface-level memorization.

1 Introduction

Large language models (LLMs) trained predominantly on English data encode vast amounts of world knowledge, yet struggle to reliably access this knowledge in other languages, leading to *cross-lingual factual inconsistency* (Wang et al., 2025a; Schut et al., 2025). This raises a key question: how can models be enabled to access their already present latent knowledge through non-English interfaces without the need for large-scale additional pretraining?

Recent work suggests that multilingual models may rely on shared internal representations (Schut et al., 2025; Wendler et al., 2024), where reasoning is performed in a shared latent space before being translated into the target language. In line with this perspective, prior work has shown that cross-lingual factual inconsistency often does not stem from missing knowledge, but rather emerges during the *language transition phase* (Wang et al., 2025a; Gekhman et al., 2025; Lu et al., 2025; Liu et al., 2025b). Specifically, models may correctly retrieve the answer in intermediate layers, yet fail to map it reliably into the target language in later layers, leading to inconsistent or incorrect outputs across languages.

More recently, parallel data has been identified as a key driver of multilingual capabilities during pretraining (Qorib et al., 2025; Shao et al., 2026; Wang et al., 2025b; Qorib et al., 2025; Fu et al., 2024; Lin et al., 2025; Wu et al., 2024). However, while continual pretraining (CPT) on parallel corpora improves translation fluency, it often fails to substantially improve performance on more demanding tasks such as multilingual factual recall (Shen et al., 2025). This suggests that parallel data

primarily improves the alignment of internal representations, while the model still struggles to reliably access knowledge encoded by the aligned representations through non-English language interfaces, resulting in inconsistent multilingual outputs.

Building on this insight, we hypothesize that multilingual factual recall in English-dominant LLMs can be improved without large-scale retraining, by separating *representation alignment* from *cross-lingual knowledge access*. Concretely, our contributions are as follows:

(i) We show that light CPT on parallel data provides limited gains for cross-lingual factual recall, motivating post-training on multilingual factual QA as a more direct mechanism for improving latent factual knowledge access in English-dominant LLMs.

(ii) We demonstrate that, among factual QA post-training methods, consistency-driven RL via GRPO consistently outperforms SFT, improving cross-lingual factual consistency and generalization to unseen languages while reshaping internal representations and multilingual routing, as visualized in Figure 1.

(iii) We create and open-source POLYFACT, a large fully parallel multilingual factual QA dataset grounded in Wikidata, spanning 100K facts across 12 typologically diverse high- and low-resource languages listed in Figure 2.

2 Related Work

Cross-Lingual Factual Recall. Recent mechanistic studies of Large Language Models (LLMs) have discovered that the primary bottleneck for cross-lingual factual recall tasks is not a knowledge deficit but rather failures during language transition (Wang et al., 2025a). This breakdown can either occur in early layers, where the model fails to map prompts into its shared English-like language agnostic conceptual space, or, more commonly, in the final layers, where latent concepts fail to decode into the correct target-language tokens (Liu et al., 2025b; Wang et al., 2025a). While, query-level interventions such as "subject injection" or English pivoting (Bandarkar et al., 2026; Liu et al., 2025b) can temporarily alleviate these inconsistencies, they function as inference-time patches. Their success suggests that cross-lingual alignment is an important bottleneck for multilingual consistency, motivating more persistent model-level adaptation.

Parallel Data. A common method of extending multilingual capabilities of English-centric language models relies on Continual Pretraining (CPT) (Fujii et al., 2024; Kuulmets et al., 2024; Shao et al., 2026). However, CPT is computationally expensive and often leads to catastrophic forgetting of the original model's English capabilities (Fujii et al., 2024). More recently, parallel data has been identified as the most salient source of mulitlingual capabilities during pretraining (Qorib et al., 2025; Fu et al., 2024; Lin et al., 2025; Wu et al., 2024). However, Shen et al. (2025) notes a significant limitation; while CPT on parallel corpora improves translation ability, it often fails to significantly boost performance in harder tasks, such as cross-lingual factual recall. This suggests that while CPT succeeds in creating a performative interface, giving the illusion of multilingual capability through surface-level fluency, it remains largely disconnected from the model's internal knowledge.

Post-Training via Reinforcement Learning. The emergence of Reinforcement Learning (RL) as a post-training method offers a novel path for adapting models to new domains and improving alignment with task-specific objectives. Mechanistic analysis by Matsutani et al. (2025) suggests that Supervised Fine-Tuning (SFT) and RL play complementary roles during post-training; while SFT expands the model's behavioural search space, RL "squeezes" it, concentrating probability mass on consistent and correct reasoning paths. While current works in monolingual reasoning consistency, such as DeReason (Hu et al., 2026) and CC-Learn (Ye et al., 2025), emphasise that RL is most effective when preceded by an SFT "warm-up" to mitigate its cold-start problem, the effect of RL in cross-lingual consistency remains largely unexplored. GRPO was popularized by DeepSeek-R1 incentivizing reasoning ability (Guo et al., 2025) with a verifiable consistency reward. This method recently was validated in the multilingual domain applied to RAG by Qi et al. (2026), allowing for a fundamental shift in multilingual optimisation.

3 Method

PolyFact Dataset. We construct POLYFACT, a fully parallel multilingual multiple-choice QA dataset for studying cross-lingual factual consistency. Starting from Wikidata truthy triples,[1] we retain 22 factual relations spanning geography, biography, creative works, and organizational or cultural ties, and extract labels in twelve high- and low-resource languages (Figure 2). For each fact, we sample three type- and length-matched distractors from co-occurring objects of the same property, then use round-robin balanced sampling to obtain 100,000 facts. We generate parallel MCQ bundles with Gemma-3-27B-IT (Kamath et al., 2025). Quality is assessed using a web-grounded GPT-4o judge and manual review, yielding 91% LLM-human agreement and a recommended POLYFACT-CLEAN filter for high-ambiguity relations. The final corpus contains 95,000 training, 2,500 validation, and 2,500 test facts, with verification labels and quality tiers included. Further details are provided in Appendix D.

Continual Pretraining. We continually pretrain on TED2025 (Shen et al., 2025), a multi-way parallel corpus covering the 12 languages in Figure 2. We retain talks containing at least two target languages and format each row as a multilingual block where available translations appear once, in randomized order, followed by an end-of-sequence token. Adjacent rows from the same talk are packed into ∼512-token chunks and truncated at 1024 tokens. To improve coverage for Swahili and Bengali, we augment TED2025 with Rogendo English–Swahili and AI4Bharat Samanantar English–Bengali sentence pairs. The final CPT corpus contains 325,134 packed chunks totalling 235.5M tokens (Table 3).

Post-Training via GRPO. We apply a multilingual variant of GRPO (Shao et al., 2024) on the POLYFACT dataset, where each training item is a single factual multiple-choice question available in parallel across all twelve languages. For every fact we sample $G = 8$ grouped rollouts; each rollout consists of twelve *independent* generations, one per language, produced from language-specific prompts that instruct the model to return the answer in the target language. For a rollout

[1] https://dumps.wikimedia.org/wikidatawiki/entities/latest-all.json.bz2

Code	Full Name	Language Family	Native Speakers (M)
en	English	Indo-European	1,500
zh	Chinese	Sino-Tibetan	1,400
es	Spanish	Indo-European	595
ar	Arabic	Afro-Asiatic	400
fr	French	Indo-European	300
bn	Bengali	Indo-European	300
pt	Portuguese	Indo-European	270
ru	Russian	Indo-European	260
id	Indonesian	Austronesian	200
de	German	Indo-European	135
ja	Japanese	Japonic	130
sw	Swahili	Niger-Congo	100

Figure 2: Expanding language coverage from English-only capability to the 12 most widely spoken languages ($18.5\% \rightarrow 70\%$ of the global population).

producing answers $\{\hat{y}_\ell\}_{\ell=1}^{L}$ across $L = 12$ languages, the reward is computed as

$$R = \sum_{\ell=1}^{L} r_\ell \; + \; 1[\forall \ell,\; r_\ell = 1]\,, \tag{1}$$

$$r_\ell = \begin{cases} +1 & \hat{y}_\ell \text{ correct option} \\ -0.5 & \hat{y}_\ell \text{ hallucination} \\ \;\;0 & \hat{y}_\ell \text{ incorrect option} \end{cases} \tag{2}$$

where an answer is considered correct if $\hat{y}_\ell$ matches the gold answer $y^\star_\ell$, and invalid if it does not match any option in $\mathcal{O}_\ell$. The final term adds a bonus of $+1$ if all languages are answered correctly, encouraging cross-lingual consistency.

Post-Training via SFT. As a supervised counterpart to GRPO, we finetune Qwen-2.5-7B and OLMo-2-1124-7B on POLYFACT with a joint classification-plus-consistency objective:

$$\mathcal{L} = -\tfrac{1}{L}\sum_{\ell=1}^{L} \log p_\ell(y^\star_\ell) \,+\, \lambda\cdot\tfrac{1}{L}\sum_{\ell=1}^{L} \mathrm{KL}(p_\ell \,\|\, \mathrm{sg}(\bar{p}))\,, \tag{3}$$

where $p_\ell \in \Delta^4$ is the model's distribution over $\{A, B, C, D\}$ in language ℓ, $\bar{p}$ is the group mean across the L=12 parallel copies of a fact (all processed in the same forward pass), $\mathrm{sg}(\cdot)$ is stop-gradient, and $\lambda = 0.5$. We introduce the consistency term because preliminary GRPO experiments showed that its joint reward reshapes the model's internal representations toward a more language-agnostic space; to isolate whether this effect is driven by the *objective* (rewarding cross-lingual agreement) versus the *algorithm* (on-policy RL), our SFT baseline must share the former. Results for a pure SFT variant with $\lambda = 0$ reported in Appendix E.1.

Mechanistic Interpretability - LAHIS To examine how finetuning affects internal language processing, we apply LAHIS (Liu et al., 2025a) to the base, SFT-, and GRPO-finetuned models. LAHIS estimates the contribution of individual attention heads to language-specific processing using a first-order Taylor approximation with respect to a learned head mask. We run the analysis across all 12 languages using parallel sentences from the TED2025 corpus (Shen et al., 2025). For each language, attention heads are ranked by their resulting importance scores, and the top 2% (20 out of 1,024 in OLMo-2-7B's 32×32 configuration) are defined as language-important. We then track how their distribution across layers and overlap across languages evolve across models.

Mechanistic Interpretability – LAPE To identify language-specific neurons in MLP layers, we use LAPE (Tang et al., 2024). We define a neuron as one intermediate feed-forward dimension, corresponding to a row of the up-projection matrix and the matching column of the down-projection matrix, and treat it as active when its post-activation value is positive.

Using parallel TED2025 text across our twelve target languages (Shen et al., 2025), we estimate each neuron's activation frequency per language, normalize these frequencies across languages, and compute the Shannon entropy:

$$LAPE_{i,j} = -\sum_{k} p'_{i,j,k} \log(p'_{i,j,k}).$$

Low-entropy neurons are more language-specialized. Following Tang et al. (2024), we take the bottom 1% as language-specific, assign each neuron to the language with its highest activation frequency, and analyze their distribution across layers and languages.

4 Experimental Setup

Models. We evaluate our approach on OLMo-2-1124-7B (OLMo et al., 2024) and Qwen-2.5-7B (Qwen et al., 2025), two 7B-scale decoder-only language models with different pretraining distributions and multilingual capabilities. OLMo-2-1124-7B serves as an English-dominant base model (documented in its open-source pretraining corpus) with a large English–non-English performance gap, while Qwen-2.5-7B provides a stronger multilingual comparison point.

Training Pipeline. To disentangle the effects of continual pretraining (CPT) from those of post-training, we study six model variants: (i) the base 7B models, (ii) a CPT model obtained by adapting the base model on 235.5M tokens of balanced parallel data, (iii-iv) SFT and GRPO models trained directly from the base model, and (v-vi) SFT and GRPO models initialized from the CPT checkpoint. This setup allows us to isolate the contribution of light multilingual continual pretraining and consistency-driven post-training, as well as their interaction.

Training Setup and Evaluation. Experiments were conducted on A100 GPUs. SFT and GRPO use LoRA finetuning (Hu et al., 2022) with rank $r = 64$ and $\alpha = 128$. Continual pretraining uses 235.5M tokens from the balanced TED2025 parallel corpus. Detailed CPT training pipeline and hyperparameters are listed in Appendix B.

For post-training, we train both SFT and GRPO on POLYFACT, either directly from the base model or from the CPT checkpoint. GRPO uses grouped rollouts with group size $G = 8$ over $L = 12$ languages and optimizes a reward that favors correct answers, penalizes hallucinated outputs, and adds a bonus when all languages are answered correctly. Full GRPO and SFT hyperparameters are reported in Appendix Tables 4 and 5.

We evaluate multilingual factual recall on POLYFACT, KLAR, and Global-MMLU (Wang et al., 2025a; Singh et al., 2025). We evaluate performance benchmarks using LightEval (Habib et al., 2023) and lm-evaluation-harness (Gao et al., 2024) with vLLM (Kwon et al., 2023) serving as the backend. We further analyze our results using mechanistic interpretability tools LAHIS (Liu et al., 2025a) and LAPE (Tang et al., 2024) as discussed in Section 3 to compare how SFT and GRPO reshape language routing and specialization.

5 Results

We first present the main performance results across datasets, before analysing cross-lingual transfer and mechanistic interpretability results to understand the roots of improved cross-lingual factual consistency.

5.1 Performance

Reinforcement Learning dominates. Table 1 shows a clear difference between supervised fine-

tuning and GRPO-based post-training. Across all three evaluation settings, GRPO delivers the strongest overall gains relative to the base model, while SFT provides weaker and less consistent improvements. On POLYFACT, both GRPO variants outperform the baseline in both high- and low-resource languages, with only GRPO achieving the best overall results. This indicates that GRPO is highly effective at improving performance on the training-style task itself, and that alignment (in our light CPT setup on parallel data) before GRPO does not further improve in-domain factual accuracy.

Transfer to KLAR. Evaluating the SFT and GRPO finetuned models on the KLAR dataset (Wang et al., 2025a) highlights the different capabilities each training method actually develops (see Table 1). Crucially, KLAR requires free-form answer generation, whereas the POLYFACT training data is multiple-choice — so transfer to KLAR is a test of whether the finetuned skills generalise beyond candidate selection. GRPO finetuning improves all languages except Russian, while SFT regresses in every language. This asymmetry suggests that SFT's training-set gains reflect pattern-matching over the answer candidates rather than genuine improvements in cross-lingual retrieval: when the candidates are removed, the learned behaviour offers no benefit and even actively hurts performance. In fact, SFT's main failure mode is to output "1" or "2" rather than make any attempt to answer the prompt. GRPO, by contrast, appears to use the POLYFACT data in a manner that develops real cross-lingual retrieval and target-language generation.

Generalisation to Global-MMLU. On Global-MMLU, the pattern is slightly different but still favours GRPO. The GRPO models achieve the highest scores for both high- and low-resource languages, while the aligned variants regress somewhat relative to the baseline. This suggests that alignment alone can reduce multilingual knowledge performance, and that while GRPO can recover and exceed baseline performance, combining alignment with post-training does not always produce additive gains. It is worth noting that Global-MMLU is substantially more challenging than pure factual recall benchmarks, as it requires multi-step reasoning and domain knowledge rather than direct retrieval. Overall, the results indicate that GRPO is the most robust method across datasets: it improves in-domain performance, transfers more effectively to held-out languages, and generalises better to broader multilingual knowledge evaluation than SFT.

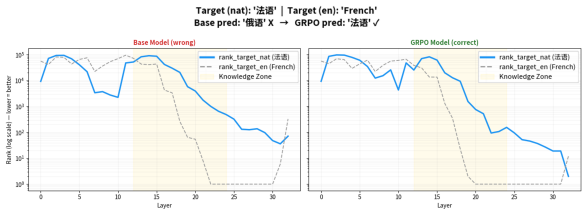
	POLYFACT		KLAR		Global-MMLU	
Method	High	Low	Trained	OOD	High	Low
OLMo-2-1124-7B						
Baseline	57.93	51.80	24.6	13.3	38.72	31.79
CPT	57.89	51.82	17.0	8.3	37.41	29.32
SFT	56.33	50.04	18.1	7.8	35.40	30.32
CPT + SFT	59.20	52.02	20.7	8.6	36.15	30.88
GRPO	**64.21**	**54.48**	29.0	16.7	**39.22**	**32.00**
CPT + GRPO	61.26	54.41	**29.8**	**17.6**	36.34	29.61
Qwen-2.5-7B						
Baseline	66.69	52.26	47.68	35.76	68.15	52.19
CPT	61.29	51.09	38.99	23.75	63.95	49.19
SFT	63.25	48.88	43.23	29.96	64.45	46.75
CPT + SFT	58.10	47.53	45.43	30.21	64.20	51.19
GRPO	**73.15**	**56.93**	**49.69**	**38.13**	**68.35**	**52.31**
CPT + GRPO	64.11	49.33	45.97	30.75	65.20	50.00

Table 1: Knowledge performance across POLYFACT, KLAR and Global-MMLU. CPT denotes light continual pretraining on parallel text before post-training. High and Low refer to high- and low-resource languages. KLAR includes evaluation on trained languages and 11 out-of-distribution (OOD) languages not present in either CPT or post-training. Language breakdown in Appendix E.4.

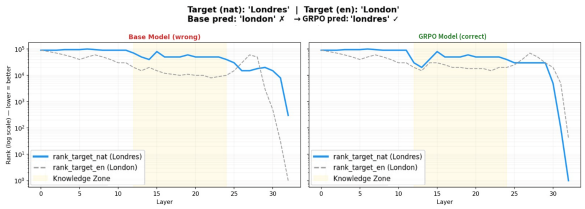
5.2 Language Transition

To understand how GRPO improves cross-lingual behaviour, we analyse the layer-wise ranking of the correct answer in both the target language and English (Figures 3b and 3a). These plots show where the model retrieves the correct fact and how it transitions into the target language.

In the base model, a consistent failure mode emerges: the correct answer is retrieved in English (grey dashed line ranks highly in later layers), while the target-language form (blue line) remains poorly ranked. This indicates that the model accesses the correct knowledge but fails to convert it into the required language. In the Spanish example (Figure 3b), the model retrieves *London* but fails to produce *Londres*; similarly, in the Chinese example (Figure 3a), the concept is accessible but not properly transferred. This reflects a breakdown in the final stage of processing: the transition from a shared or English-centric representation to the target-language output, aligned with

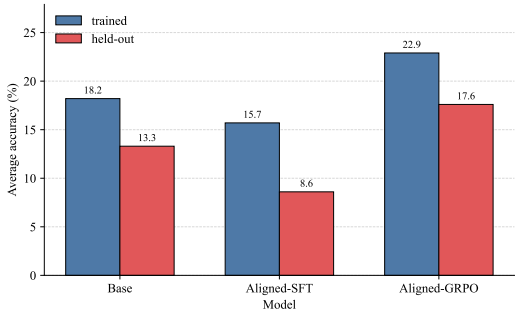


(a) [ZH] Official language: "French Polynesia"

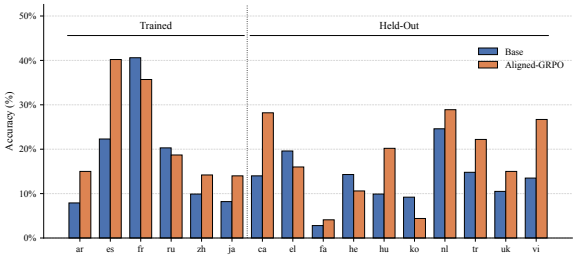


(b) [ES] Place of birth: "Mark Strong"

Figure 3: Layer-rank analysis of Base (left) and GRPO finetuned OLMO-2 (right) models, showing rank of correct English and target language tokens. (a) Example in non-Latin script language. (b) Example in Latin-script language.



(a) Accuracy on trained vs. held-out languages.



(b) Per-language accuracy comparison.

Figure 4: OLMo-2-7B KLAR performance on trained and held-out languages across (a) models and (b) languages.

findings from (Wang et al., 2025a).

Consistency-driven GRPO fine-tuning mitigates this issue. In both examples, the target-language answer rises in early layers and overtakes the English form in later layers. This suggests that GRPO strengthens the pathway between language-agnostic knowledge representations and target-language decoding, enabling the model to directly produce the correct linguistic form rather than relying on an English intermediate. A more detailed analysis of failure modes during factual retrieval can be found in Appendix C.

5.3 Cross-lingual Transfer

We evaluate cross-lingual transfer by measuring performance on 11 other languages held out during post-training such as Greek, Ukrainian and Hebrew. As shown in Figure 4a, GRPO improves accuracy not only on previously seen languages but also on held-out languages, indicating that its gains are not limited to the training distribution. In contrast, SFT shows weaker improvements and limited transfer, suggesting that it primarily captures language-specific memorization rather than generalizable cross-lingual behaviour. The consistent gains of GRPO on held-out languages indicate that it promotes more language-agnostic representations, enabling the model to access and express knowledge across languages beyond those seen during training.

A more fine-grained per-language analysis (Figure 4b) reveals that these gains are not uniformly distributed. Instead, CPT combined with GRPO yields the largest improvements for seen languages with relatively low baseline performance, such as Arabic and Japanese, as well as other unseen languages with scripts similar to the ones seen during training. This is most evident in languages like Catalan and Vietnamese (Latin-script) and Farsi which likely benefits from the alignment of Perso-Arabic subwords during the training on Arabic. Meanwhile, the most pronounced performance drops occur in unseen "script-isolated" languages such as Greek, Korean, and Hebrew, where the absence of similar tokens during training means language transition failures persist.

This pattern suggests that GRPO improves factual access through shared language interfaces, particularly for languages with lexical or orthographic overlap with the training languages, such as Latin-script languages and languages sharing Perso-Arabic subwords. However, these gains remain constrained by script and tokenization boundaries: languages with more isolated scripts, such as Greek, Korean, and Hebrew, benefit less. At the same time, small drops for some high-performing seen languages, such as French, suggest a mild trade-off between peak per-language accuracy and broader cross-lingual regularization.

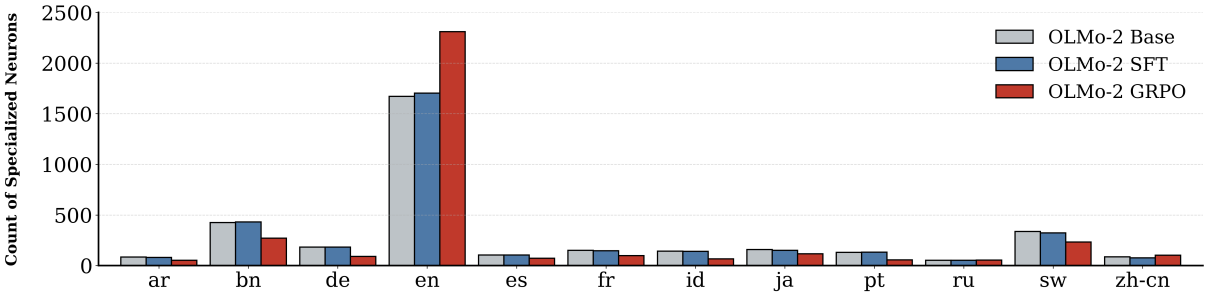


Figure 5: Language-specific neurons across languages; GRPO increases English alignment.

Taken together, these results support the view that GRPO encourages a more balanced allocation of model capacity across languages, reducing performance disparities and improving cross-lingual consistency by prioritizing shared, rather than isolated, retrieval pathways.

5.4 Language-Specific Neurons

Following the LAPE methodology described in Section 3, we identify and evaluate the distribution of language-specific neurons in three configurations of OLMo-2-7B: the base model and POLYFACT fine-tuned model variants using SFT and GRPO.

Total neuron count shifts towards English-specificity. An analysis of the language-specific neuron distribution reveals a striking and somewhat counterintuitive effect. While the SFT model remains nearly identical to the Base model, GRPO exhibits a pronounced shift toward English-specific neurons. As shown in Figure 5, GRPO increases English-specific neurons by 38.2% (1671 to 2310), at the expense of other languages, with large reductions for Bengali (-36.1%, 424 to 271) and Swahili (-31.2%, 337 to 232), while Chinese is a notable exception (+18.6%, 86 to 102).

Notably, this shift cannot be explained by improved English performance. The SFT model achieves stronger English POLYFACT accuracy than GRPO, and GRPO itself performs slightly better than the baseline (Appendix 9). This suggests the effect reflects a structural reorganisation rather than performance gains. One interpretation is that the RL objective "squeezes" the model's behavioural space, encouraging reliance on the most stable representational backbone (here: English), consistent with the work of Matsutani et al. (2025).

GRPO delays linguistic specialisation in non-Latin scripts. To quantify the layer-wise distribution of language-specific neurons, we analyse the empirical cumulative distribution function (ECDF) across layers (Figure 15 in Appendix E.3). We observe a clear "late-discovery" effect unique to GRPO: while the base and SFT models show nearly identical distributions, the GRPO curve is systematically lower, indicating that specialised neurons are recruited later in the network. This is reflected in a negligible Kolmogorov-Smirnov distance for SFT ($D_{KS} = 0.005$) and a substantial shift for GRPO ($D_{KS} = 0.089$).

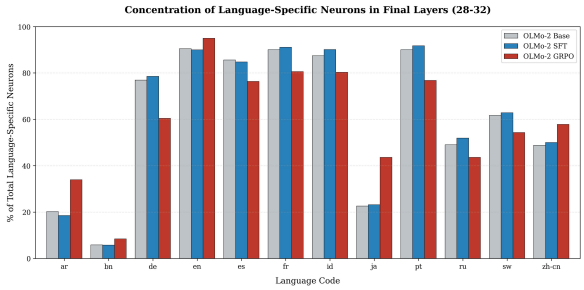
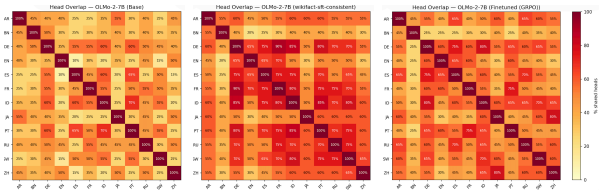


Figure 6: Concentration of specialised neurons in the final layers (28–32). GRPO exhibits a bifurcation: Latin scripts shift toward intermediate layers, while distinct scripts consolidate in the final layers.

This suggests that GRPO preserves a larger language-agnostic space in intermediate layers and defers linguistic specialisation. Notably, Latin-script languages and Russian (which shares a large portion of its alphabet with Latin scripts) shift specialisation earlier, while languages of non-Latin script (e.g. Arabic, Japanese) concentrate it in the final layers, as visualized in Figure 6. A full breakdown can be found in Appendix E.3.

5.5 Mechanistic Analysis via LAHIS

Cross-lingual head sharing. Head overlap between language pairs increases substantially after both finetuning methods (Figure 7). SFT produces the strongest gains in pairwise overlap, particularly among Indo-European languages (e.g. DE–



(a) Base (b) SFT (c) GRPO

Figure 7: Pairwise head overlap across language pairs for Base, SFT-, and GRPO-finetuned models.

FR: 25% → 90%, DE–ID: 35% → 85%). GRPO also increases overlap but more moderately, with notable gains for typologically distant pairs such as JA–ZH (50% → 80%). This suggests that SFT primarily encourages shared representations across languages, while GRPO's key effect lies in redistributing processing deeper into the network rather than maximising head sharing.

Language processing shifts deeper into the network. In the base model, up to 50% of language-important heads for some languages concentrate at layer 0, indicating near-immediate language-routing decisions at the first layer. Both finetuning methods redistribute this: under SFT the peak drops to around 40% with a slight spread across early layers, while GRPO reduces it further to 20% and distributes heads across layers 0–10 (Figure 8). This suggests GRPO drives a more substantial reorganisation of language processing particularly in the early routing layers. More detailed analysis of language-important attention head reshaping can be found in Appendix E.2. Overall, the mechanistic interpretability results are complementary and reveal a qualitatively distinct effect of GRPO compared to SFT: while SFT primarily increases cross-lingual sharing, GRPO fundamentally restructures multilingual computation, shifting it away from early routing and toward a shared, language-agnostic intermediate space with delayed language-specific specialisation.

6 Conclusion

We study cross-lingual factual recall in predominantly English-pretrained LLMs and introduce PolyFact, a fully parallel multilingual factual QA dataset spanning 100K Wikidata-grounded facts across 12 languages. Across Qwen-2.5-7B and OLMo-2-1124-7B, GRPO consistently outperforms supervised finetuning, improving factual recall, transfer to held-out languages, and

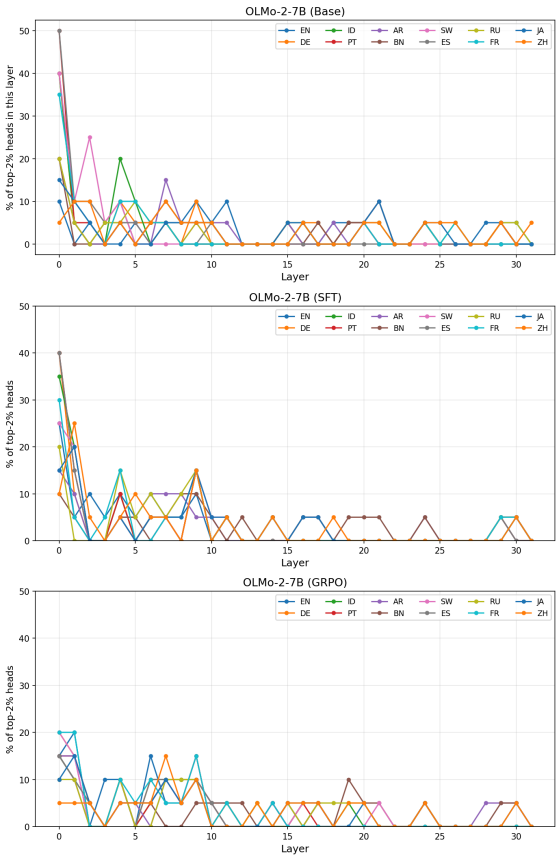


Figure 8: Percentage of language-important attention heads in OLMo-2-1124-7B across Base, SFT and GRPO.

generalisation beyond the multiple-choice setting, while light CPT on parallel data provides limited gains. Mechanistically, our analyses suggest that these improvements are not merely surface-level: GRPO improves the transition from shared factual representations into target-language outputs, restructures language routing, and reduces language-specific specialization in favour of more shared cross-lingual computation. Overall, our results suggest that cross-lingual factual inconsistency is less a problem of missing knowledge than of unreliable access, and that consistency-driven RL provides an effective way to improve cross-lingual factual recall without large-scale retraining.

7 Limitations

Our work has several limitations. First, our experiments are conducted on two 7B model families, OLMo-2-7B and Qwen-2.5-7B. While these models cover different training setups and multilingual capabilities, it is not certain whether our

findings generalise to larger models, smaller models, or substantially different architectures such as mixture-of-experts models.

Second, training on POLYFACT does not fully generalise to more challenging benchmarks such as Global-MMLU, indicating limitations on tasks that require deeper reasoning beyond factual retrieval. Our method is therefore best understood as improving cross-lingual factual access rather than solving multilingual reasoning more broadly.

Third, our evaluation is limited to a relatively small set of benchmarks and languages, which may not fully capture all aspects of real-world multilingual performance. Broader evaluation across more tasks, domains, and languages would strengthen our conclusions.

Finally, our work introduces potential risks. Since POLYFACT is derived from Wikidata, it may inherit coverage biases, factual errors, or uneven representation across languages, regions, and entities. Moreover, improving cross-lingual factual recall may also make incorrect or biased factual associations more consistently expressed across languages if such information is present in the underlying model or dataset. We mitigate these risks through relation filtering, automatic and manual quality checks, and by releasing the dataset and code to support transparency, auditing, and future corrections.

References

Lucas Bandarkar, Alan Ansell, and Trevor Cohn. 2026. Large reasoning models struggle to transfer parametric knowledge across scripts. *arXiv preprint arXiv:2603.17070*.

Tingchen Fu, Lemao Liu, Deng Cai, Guoping Huang, Shuming Shi, and Rui Yan. 2024. The reasonableness behind unreasonable translation capability of large language model. In *The Twelfth International Conference on Learning Representations*.

Kazuki Fujii, Taishi Nakamura, Genta Indra Winata, et al. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. *arXiv preprint arXiv:2404.17790*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Zorik Gekhman, Eyal Ben-David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. Inside-out: Hidden factual knowledge in LLMs. In *Second Conference on Language Modeling*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, and Shirong Ma. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hanxu Hu, Yuxuan Wang, Maggie Huan, Jannis Vamvas, Yinya Huang, Zhijiang Guo, and Rico Sennrich. 2026. DeReason: A difficulty-aware curriculum improves decoupled SFT-then-RL training for general reasoning. *arXiv preprint arXiv:2603.11193*.

Lucie-Aimée Kaffee, Alessandro Piscopo, Pavlos Vougiouklis, Elena Simperl, Leslie Carr, and Lydia Pintscher. 2017. A glimpse into babel: An analysis of multilinguality in Wikidata. In *Proceedings of the 13th International Symposium on Open Collaboration (OpenSym '17)*, pages 14:1–14:5. ACM.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching Llama a new language through cross-lingual knowledge transfer. In *Findings of NAACL 2024*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Haobin Lin, Yan Zhao, Wenhan Han, Ping Guo, BINBINLIU, Yifan Zhang, Bingni Zhang, Taifeng Wang, and Yin Zheng. 2025. From translation to multilinguality: Revisit the role of parallel data in multilingual LLM pretraining.

Xin Liu, Qiyang Song, Qihang Zhou, Haichao Du, Shaowen Xu, Wenbo Jiang, Weijuan Zhang, and Xiaoqi Jia. 2025a. Focusing on language: Revealing and exploiting language attention heads in multilingual large language models.

Yihong Liu, Mingyang Wang, François Yvon, and Hinrich Schütze. 2025b. On the entity-level alignment in crosslingual consistency. *arXiv preprint arXiv:2510.10280*.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. 2025. Paths not taken: Understanding and mending the multilingual factual recall pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15066–15096, Suzhou, China. Association for Computational Linguistics.

Kohsei Matsutani, Shota Takashiro, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2025. RL squeezes, SFT expands: A comparative study of reasoning LLMs. *arXiv preprint arXiv:2509.21128*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 olmo 2 furious.

Rui Qi, Fengran Mo, Yufeng Chen, Xue Zhang, Shuo Wang, Hongliang Li, Jinan Xu, Meng Jiang, Jian-Yun Nie, and Kaiyu Huang. 2026. Language-coupled reinforcement learning for multilingual retrieval-augmented generation. *arXiv preprint arXiv:2601.14896*.

Muhammad Reza Qorib, Junyi Li, and Hwee Tou Ng. 2025. Just go parallel: Improving the multilingual capabilities of large language models. In *Proceedings of ACL 2025*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Jiandong Shao, Raphael Tang, Crystina Zhang, Karin Sevegnani, Pontus Stenetorp, Jianfei Yang, and Yao Lu. 2026. The role of mixed-language documents for multilingual large language model pretraining.

Zhihong Shao, Zhibin Gou, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Yingli Shen, Wen Lai, Shuo Wang, Ge Gao, Kangyang Luo, Alexander Fraser, and Maosong Sun. 2025. From unaligned to aligned: Scaling multilingual llms with multi-way parallel corpora. In *Proceedings of EMNLP 2025*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. 2025a. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of ACL 2025*.

Zhijun Wang, Jiahuan Li, Hao Zhou, Rongxiang Weng, Jingang Wang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025b. Investigating and scaling up code-switching for multilingual language model pre-training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11032–11046, Vienna, Austria. Association for Computational Linguistics.

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.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data. In *Findings of the Association*

for Computational Linguistics: ACL 2024, pages 15092–15108, Bangkok, Thailand. Association for Computational Linguistics.

Xiao Ye, Shaswat Shrivastava, Zhaonan Li, Jacob Dineen, Shijie Lu, Avneet Ahuja, Ming Shen, Zhikun Xu, and Ben Zhou. 2025. CC-LEARN: Cohort-based consistency learning. *arXiv preprint arXiv:2506.15662*.

A Benchmark details

We evaluate across three benchmarks targeting complementary aspects of multilingual factual recall: factual recall in the training-task format (PolyFact), free-form factual recall (KLAR-CLC) and broader multilingual knowledge and reasoning (Global-MMLU). All inference uses bf16 precision.

PolyFact. We evaluate on the 2,500-fact test split of PolyFact using a custom evaluator (`evaluate/evaluate_consistency.py`) that scores each of the four MCQ options by its length-normalised conditional log-likelihood under the model, conditioning on a language-specific prompt that wraps the question and instructs the model to answer in the target language. The option with the highest per-token average log-probability is selected. We report per-language accuracy across all twelve target languages (en, de, es, fr, pt, id, ru, ar, zh, ja, sw, bn). Because facts are fully parallel across languages, per-language differences isolate language-interface effects from underlying knowledge.

KLAR-CLC. For free-form cross-lingual factual recall we evaluate on KLAR-CLC (Wang et al., 2025a) using a custom evaluator (`evaluate/evaluate_klar.py`). The evaluator runs 3-shot prompted greedy generation (max 10 new tokens) and compares the generated answer to the gold using NFC-normalised, case-insensitive, punctuation-stripped string matching, with the punctuation regex covering ASCII, CJK, and typographic quotes/guillemets to handle all 17 KLAR languages. KLAR is structured into 20 factual relations (e.g. *capital*, *place of birth*, *manufacturer*, *occupation*). Unlike PolyFact, KLAR removes the candidate set, testing whether the model can *generate* the correct answer in the target language rather than just discriminate among given options. We evaluate on six seen languages (en, es, fr, ru, zh, ja) and eleven held-out languages (ar, ca, el, fa, he, hu, ko, nl, tr, uk, vi) to measure cross-lingual transfer.

Global-MMLU. For broader multilingual knowledge and reasoning we evaluate on Global-MMLU (Singh et al., 2025) using `lm-evaluation-harness` (Gao et al., 2024) with vLLM (Kwon et al., 2023) as the backend (task identifiers `global_mmlu_full_{lang}`, `batch_size=auto`,

Data	
Max sequence length	1024 tokens
Chunk packing target	512 tokens (talk-internal)
Min languages per row	2
Min languages per talk	2
Validation split	0.5% (language-balanced)
Optimization	
Optimizer	AdamW (fused)
Learning rate	2×10^{-5}
LR schedule	Cosine
Warmup steps	200
Weight decay	0.01
Precision	bf16
Epochs	1
Per-device batch size	6
Gradient accumulation steps	8

Table 2: Continual pretraining hyperparameters.

Language	# lang-lines	Tokens (M)
English (en)	1,778,887	67.2
Spanish (es)	656,921	19.5
Arabic (ar)	615,808	15.3
French (fr)	534,647	16.6
Chinese (zh)	515,553	13.7
Bengali (bn)	500,581	35.3
Russian (ru)	469,035	16.0
Japanese (ja)	356,325	4.9
German (de)	297,306	9.8
Indonesian (id)	219,040	7.3
Portuguese (pt)	210,441	6.7
Swahili (sw)	174,125	4.0
Total	6,328,669	216.4

Table 3: Per-language coverage in the TED 2025 CPT corpus tokenized with the Qwen-2.5 tokenizer. "# lang-lines" counts how often each language appears as a section within a packed chunk; a single TED chunk typically contains 5–12 language sections. Token counts exclude the `{lang}:` format wrapper. The remaining ∼19M tokens in the 235.5M total are end-of-sequence and other framing tokens.

`gpu_memory_utilization=0.85`). Scoring is standard log-likelihood over the four MCQ options. We report per-language accuracy across the twelve target languages. Global-MMLU is substantially harder than direct factual-recall benchmarks because it additionally requires multi-step reasoning and domain knowledge, making it a stricter test of generalisation beyond training-distribution facts.

Inference backends. `lm-evaluation-harness` and LightEval use vLLM for throughput; the POLYFACT and KLAR custom evaluators use HuggingFace Transformers directly because they need access to per-token log-probabilities (POLYFACT) and explicit batched greedy decoding with custom string matching (KLAR), which are awkward to express in the harnesses.

B Hyperparameters

Continual Pretraining. We continually pretrain OLMo-2-1124-7B (OLMo et al., 2024) and Qwen-2.5-7B (Qwen et al., 2025) on the balanced TED2025 dataset using the hyperparameters in Table 2. A detailed per-language of the TED2025 subset used for CPT is listed in Table 3.

GRPO Post-Training. Starting from Olmo-2-1124-7B basemodel and CPT checkpoint, we run multilingual GRPO on POLYFACT with the configuration in Table 4.

Supervised Finetuning. Starting from Olmo-2-1124-7B basemodel and CPT checkpoint, we run multilingual SFT on POLYFACT with the configuration in Table 5.

C Failure Modes during Cross-Lingual Retrieval

Performing a layer-rank analysis of the base and GRPO-finetuned model we can get a better understanding of the specific ways in which we have improved performance. The main failure mode of the base model across all languages is producing incoherent answers including randomly guessed numbers, the word "what" and hybrid combinations of different languages. GRPO is effective in reducing this and contributes most of the gain in performance. Beyond this, the next most common failure modes that we are able to improve upon are largely distinct between languages that do and do not share a Latin script with English. We include Russian in the Latin-script group here: although Cyrillic, it shares enough tokenisation overlap with Latin that it exhibits the same failure mode. In essence, the base model suffers from some languages' representations being too close to that of English, such that their interlingual pathways are not distinct enough; and some being too separated, such that the routes between them and English are too weak. GRPO fine-tuning helps to reduce both contrasting issues.

Those that do share the same script, and likely similar representations within the model, suffer

Parameter	Value
Dataset	POLYFACT
Languages (L)	12
Group size (G)	8
Max train facts	40,000
Epochs	1
Batch size	1 fact
Learning rate	1×10^{-5}
LR schedule	cosine, 3% warmup
Weight decay	0.01
Optimizer	AdamW
Grad clipping	1.0
Max prompt / completion	512 / 48
Temperature / top-p	0.7 / 0.95
Repetition penalty	1.5
No-repeat n-gram	3
Correct reward	$+1$
Hallucination penalty	-0.5
All-correct bonus	$+1$
KL coefficient (β)	0.0
LoRA r / α / dropout	64 / 128 / 0.05
Precision	bf16
Grad checkpointing	enabled
Eval frequency	every 500 steps
Seed	42

Table 4: GRPO post-training hyperparameters.

Parameter	Value
Dataset	POLYFACT
Languages per fact (L)	12
Epochs	1
Facts per device batch	8
Gradient accumulation	8
Effective batch (facts)	64
Learning rate	2×10^{-5}
LR schedule	cosine, 5% warmup
Weight decay	0.01
Optimizer	AdamW (fused)
Consistency weight (λ)	0.5
Consistency loss	KL to stop-grad group mean
Max eval samples/lang	200
LoRA r / α / dropout	64 / 128 / 0.05
Precision	bf16
Grad checkpointing	enabled
Eval / save frequency	every 200 steps

Table 5: SFT post-training hyperparameters.

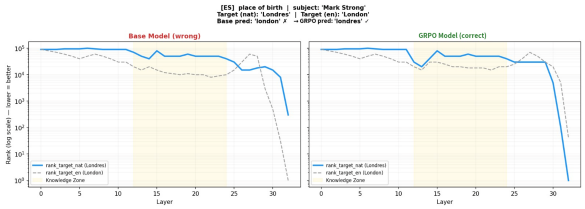


Figure 9: Layer-rank analysis of Base and GRPO fine-tuned OLMO-2 models for a prompt in latin script language.

from responding to prompts with the right answer but in English rather than the target language (Figure 9). Either the base model cannot find the direct translation for the prompt's answer or its prior training has conditioned it to rely too heavily on English-tied pathways within the Latin-script regions of its representations. The former seems unlikely, particularly given the expected prominence of the answer in the specific example shown. Therefore, our GRPO fine-tuning is able to encourage the model to separate pathways for Latin script languages and respond in the correct language. This unfortunately does not apply to all examples and in fact GRPO produces nearly as many regressions of this type as it does improvements. Though this looks discouraging on the surface, a deeper analysis of where these failures occur reveals much about the base model and our fine-tuning process.

The KLAR dataset is structured into *relations* that group questions from the same fields. We break these relations down into two groups, those with and without proper-noun answers (jurisdiction, manufacturer, owned-by etc). Given the global prevalence of English-language proper nouns it is understandable that many of these proper-nouns originate in the English language and may be known as such, even if they have native translations. In some instances this will naturally aid the answering of these questions: target-language answers are often similar or identical to the English forms the model can retrieve more easily. Proper-noun relations account for 61%, 76%, and 73% of base-to-GRPO regressions in Russian, French, and Spanish respectively. While the prevalence of English proper-nouns in the base model's pretraining corpus likely helps it score well on the low-hanging lexical cognates, it also pushes many answers into the wrong language failure state — a tendency our finetuning may have inadvertently reinforced. Wikipedia is well known for exhibiting this bias, where entity labels for people, companies, and many places default to English or Latin-script forms even on non-English Wikipedias, because Wikidata's canonical label is often English and localisers haven't filled in every language (Kaffee et al., 2017). Finetuning on POLYFACT therefore strengthens this bias, worsening the wrong-language failure mode on proper-noun prompts.

This points us to the notion that GRPO fine-tuning is actually stronger than our initial met-

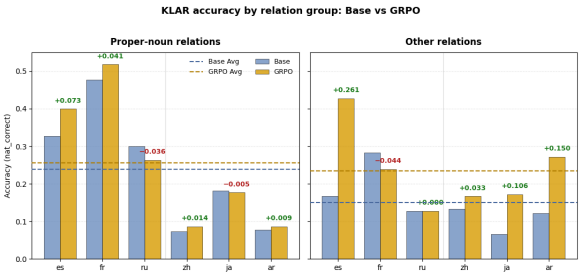


Figure 10: KLAR improvement by relation type. Proper-noun relations account for 55% of the KLAR dataset.

rics suggest. The gains GRPO actually delivers on language abstraction are partly masked in the aggregate by regressions concentrated in proper-noun prompts where dataset bias works against us. Figure 10 makes this visible: improvements on common-noun (non-proper-noun) relations are substantially larger across the board, including in non-Latin-script languages.

We highlight the regression of Russian performance after our fine-tuning, as it is a slight anomaly though we cannot fully explain it. Some gains come from cleaning up incoherent answers, but the overall result is dominated by regressions in the same failure mode as the Latin based languages. This complicates our script-based framing but the aforementioned the English-leak failure mode reaches it too. What we cannot fully explain is its magnitude: Russian is the only language to net-regress on KLAR, and does so more severely than fully Latin-script French, despite sharing the underlying mechanism.

In languages that do not share the same script or have this overlap the failure mode for questions not related to proper-nouns is qualitatively different. These languages perform much worse in general evaluation, in part because of weaker factual retrieval: responses are more often in the correct language but factually wrong (Figure 3a). The model's representations for these languages are clearly distinct from those of English and its language-agnostic regions. This separation means the base model answers in the correct language but the cross-lingual transfer of factual content from the English/agnostic knowledge region into the language-specific output pathway fails. Our GRPO fine-tuning strengthens the pathways between these distinct languages and more commonly allows the model to answer in the correct language with the correct information.

Our GRPO fine-tuning uses a reward function that requires both the correctness of the answer and its production in the target language to assign credit. This is what enables us to correct both failure modes, pushing the model towards answering correctly in the right language from either direction relative to the target. But there is no notion of partial credit and given the distinct failure modes (correct language, wrong information vs wrong language, correct information) it seems that there may be specific benefits tailored to each language type to including this in future iterations. Similarly, an altered training dataset could yield further gains. The reward rewards correctness, and the canonical correct answer is often English, so the reward effectively trains in the English-leak failure on proper-noun prompts The aforementioned biases of the PolyFact database could be corrected to favour more consistently native proper nouns and we would likely see the true gains of this method unhampered by the discussed regressions.

D PolyFact Dataset.

D.1 Detailed Construction.

We describe the full pipeline used to construct PolyFact, a parallel multilingual multiple-choice QA dataset grounded in Wikidata.

Source data and relation selection. We start from the full Wikidata truthy-triples dump and retain only triples whose property belongs to a curated set of 22 factual relations, selected to cover stable, unambiguous facts suitable for MCQ generation. These span geography (*capital*, *country*, *continent*, *official language*, *currency*, *shares border with*), biography (*country of citizenship*, *place of birth*, *place of death*, *educated at*, *employer*), creative works and media (*author*, *director*, *creator*, *developer*, *genre*, *country of origin*, *language of work*, *platform*), and organizational or cultural relations (*manufacturer*, *architect*, *discoverer or inventor*).

Multilingual label extraction. For every subject, property, and object entity, we extract labels in twelve typologically and geographically diverse languages: English, German, Indonesian, Portuguese, Arabic, Bengali, Swahili, Spanish, Russian, French, Japanese, and Chinese. Labels are collected primarily by streaming the compressed Wikidata JSON dump, with missing entries backfilled through the Wikidata `wbgetentities`

API in batches of 50 IDs. We additionally extract each entity's `instance of` (P31) types from the dump for distractor selection.

Triple filtering and distractor construction. We retain only triples whose subject and object have labels in a sufficient subset of the twelve target languages. For each surviving triple, we sample three distractors from objects that appear with the same property elsewhere in the corpus, so distractors are plausible candidates for the relation (e.g., distractors for a *place of birth* question are drawn from other Wikidata-recorded birthplaces). To prevent surface-level shortcuts, distractors must additionally share at least one entity type with the gold object — via Wikidata's `instance of` property (P31), e.g. both being cities — and match its English label length within a small tolerance. We discard any fact whose four options are not all distinct under case- and whitespace-normalised comparison.

Balanced sampling. To prevent high-frequency relations from dominating the corpus, we apply round-robin sampling across properties: on each pass, one fact is drawn from each property's shuffled pool, cycling until 100,000 facts have been selected. This produces a near-uniform distribution over relations regardless of the underlying frequency skew in Wikidata.

Multilingual MCQ generation. For each selected fact, we generate a parallel MCQ bundle covering all twelve languages using gemma-3-27b-it served via vLLM. Entity and property labels missing from Wikidata are prefilled using a dedicated low-temperature translation prompt (temperature 0, max 32 tokens) conditioned on the relation context, with results cached across the batch to amortize cost. Given the localized subject, relation, gold answer, and three distractors, the model generates a question in the target language under explicit instructions to reuse the provided option strings verbatim (temperature 0.1, top-p 0.9, max 192 tokens). The predicted `answer_text` is resolved against the option set via exact match and, on failure, case- and whitespace-normalized string matching against both the options and the known gold label.

Validity and parallelism. Each generated MCQ is validated for well-formedness: exactly four distinct non-empty options, a non-empty question string, and an `answer_text` that resolves to one of the options. A fact is retained only if generation succeeds for *all* twelve languages; partial bundles are discarded entirely. This strict all-or-nothing criterion guarantees that every fact in POLYFACT is fully parallel, so cross-lingual performance differences can be attributed to model behavior rather than to variation in underlying content, question difficulty, or answer set composition.

Splits. The final corpus is partitioned into 95,000 training, 2,500 validation, and 2,500 test facts, with splits applied at the fact level so that all twelve language versions of a given fact remain in the same partition.

D.2 Quality Verification.

To characterise residual ambiguity in the released dataset, we labelled 300 facts per language across all 12 languages (3,600 items total) with an LLM-as-judge (GPT-4o with web-search tool grounding), and an independently human-labelled 100 items per language across six languages (Arabic, English, French, German, Russian, Spanish; 600 items total), using a three-label rubric (correct / ambiguous / incorrect). Using an LLM judge enables per-relation quality estimation at a scale infeasible for purely manual annotation and provides per-item verification flags for a significantly bigger subset. Inter-judge agreement between the LLM judge and the independent human review is 91.0% overall, with per-language values ranging from 86% (Arabic) to 96% (English), as listed in Table 6. The LLM judge identifies an overall ambiguity rate of 13.3% across the 3,600-item sample; rates vary both by Wikidata property and by language. By property, three relations exhibit relatively higher ambiguity: *country of origin* (35%), *place of birth* (25%), and *genre* (25%). The ambiguity in the first two cases is largely driven by person-name queries, where multiple individuals may share the same name, while *genre* is inherently more subjective and underspecified. The remaining 16 relations all have ambiguity rates at or below 20%. By language, Japanese (23%) and Chinese (21%) show roughly twice the ambiguity rate of European languages (9–13%), consistent with the translation-induced ambiguity patterns we observed in qualitative review (e.g., transliterated subjects losing disambiguating information). We release per-item verification labels and a recommended POLYFACT-CLEAN filter (which excludes the three highest-ambiguity relations) alongside the dataset; see Table 7 for the per-relation breakdown.

Language	N	Agreement
English (en)	100	96.0%
French (fr)	100	93.0%
German (de)	100	92.4%
Spanish (es)	100	92.0%
Russian (ru)	100	87%
Chinese (zh)	100	91.9%
Arabic (ar)	100	86.0%
Overall	700	91.0%

Table 6: LLM-human inter-judge agreement on the post-hoc verification sample (100 items per language). N is the number of items with both an LLM-judge label and an independent human label.

E Additional Results

E.1 Pure SFT model

In the following we list our results for the purely SFT trained model ($\lambda = 0$) on POLYFACT (see Table 8. Because this doesn't include any consistency term, this model overindexes on individual per-language performance rather than learning to route the same knowledge consistently through shared pathways.

E.2 LAHIS

Despite similar overlap patterns, the two methods differ in how much they alter the base model's routing. Under SFT, between 45% and 75% of language-important heads change relative to the base, while GRPO replaces between 70% and 85% (Figure 12). Per-language delta maps for both models are provided in Appendix E.2. Pairwise head overlap for the three Qwen-2.5-7B variants is visualized in Figure 13.

Per-language Delta Maps Figures 16 and 17 show per-language delta maps of LAHIS importance scores (finetuned − base) for GRPO and SFT-consistent respectively. Both methods predominantly suppress existing language-specific heads rather than creating new ones, with the strongest suppression at layer 0.

E.3 Per-language Neuron Distributions

Figure 14 provides a detailed visualisation of the language activation probability entropy (LAPE) results, showing the frequency of specialised neurons per layer for all twelve target languages across the Base, SFT, and GRPO configurations. Figure 15 provides the empirical cumulative distribution functions for the base, SFT and GRPO configurations, summed over all languages.

Relation	N	Ambig. %
Excluded from POLYFACT-CLEAN		
country of origin	168	35
genre	240	25
place of birth	216	25
Included in POLYFACT-CLEAN		
director	192	20
continent	216	20
language of work or name	252	18
educated at	240	17
currency	36	14
employer	216	11
architect	180	9
official language	168	8
creator	192	8
place of death	216	8
developer	192	7
country	192	6
country of citizenship	216	5
author	60	3
discoverer or inventor	216	3
manufacturer	192	2

Table 7: Per-relation ambiguity rates from the full 3,600-item LLM-judge verification sample (300 items per language across 12 languages), sorted by the fraction of items labelled ambiguous or incorrect. POLYFACT-CLEAN excludes the three highest-ambiguity relations (top group), removing approximately 17% of items. Three of the 22 curated relations did not appear in the random sample.

	KLAR (Seen)						KLAR (Unseen)										
Model	en	es	fr	ru	zh	ja	ar	ca	el	fa	he	hu	ko	nl	tr	uk	vi
SFT-pure	64.4	24.7	31.0	15.1	9.0	5.2	5.4	16.7	8.4	3.7	3.4	8.8	1.7	19.6	16.1	6.6	12.1

	POLYFACT											
Model	en	de	es	fr	pt	id	ru	zh	ar	ja	sw	bn
SFT-pure	81.37	69.08	66.79	69.96	45.78	41.85	48.16	72.89	54.46	71.78	51.41	48.79

Table 8: SFT-pure performance on KLAR and POLYFACT.

E.4 Per-language Performance

POLYFACT: See Table 9.

KLAR: See Table 10.

Global-MMLU. See Table 11.

Model	en	de	es	fr	pt	id	ru	zh	ar	ja	sw	bn
Baseline	79.94	65.95	64.45	68.61	42.81	41.46	44.31	70.04	54.58	66.19	48.83	47.96
Aligned	79.11	65.68	63.42	68.69	42.96	39.56	47.56	68.21	54.46	68.09	48.59	48.39
GRPO	76.50	70.87	67.97	71.30	49.43	44.83	48.83	71.66	55.81	69.80	51.41	50.18
Aligned + GRPO	79.98	71.03	66.71	70.91	49.23	43.36	51.13	71.74	58.07	71.74	51.09	50.14
SFT	80.90	63.06	65.40	60.36	44.07	40.35	46.29	65.64	49.98	65.60	49.03	45.26
Aligned + SFT	80.62	64.57	66.55	66.75	46.14	43.12	51.61	70.04	54.86	68.65	50.26	47.21

Table 9: POLYFACT accuracy (%) across languages.

	Seen						Unseen										
Model	en	es	fr	ru	zh	ja	ar	ca	el	fa	he	hu	ko	nl	tr	uk	vi
Baseline	62.9	22.3	40.6	20.3	9.9	8.2	7.9	14.0	19.6	2.8	14.3	9.9	9.2	24.6	14.8	10.5	13.5
Aligned	50.2	13.2	21.9	13.7	9.7	3.0	7.4	9.5	10.4	2.4	6.5	8.1	6.2	11.2	14.7	3.7	10.4
GRPO	75.5	37.6	44.3	13.4	11.8	8.2	12.2	26.3	8.6	9.6	2.2	19.5	5.5	35.7	22.0	10.0	27.2
Aligned + GRPO	71.1	40.2	35.7	18.7	14.2	14.0	15.0	28.2	16.0	4.1	10.6	20.2	4.4	28.9	22.2	15.0	26.7
SFT	42.8	16.7	32.4	14.2	8.4	5.4	6.9	8.4	11.8	2.6	3.9	7.1	2.3	11.8	15.0	3.7	11.0
Aligned + SFT	50.2	29.3	35.1	12.1	8.4	2.3	7.3	11.1	7.2	2.3	1.6	8.3	3.7	13.9	20.3	6.3	11.5

Table 10: KLAR accuracy (%) across languages, grouped into seen and held-out (unseen) languages.

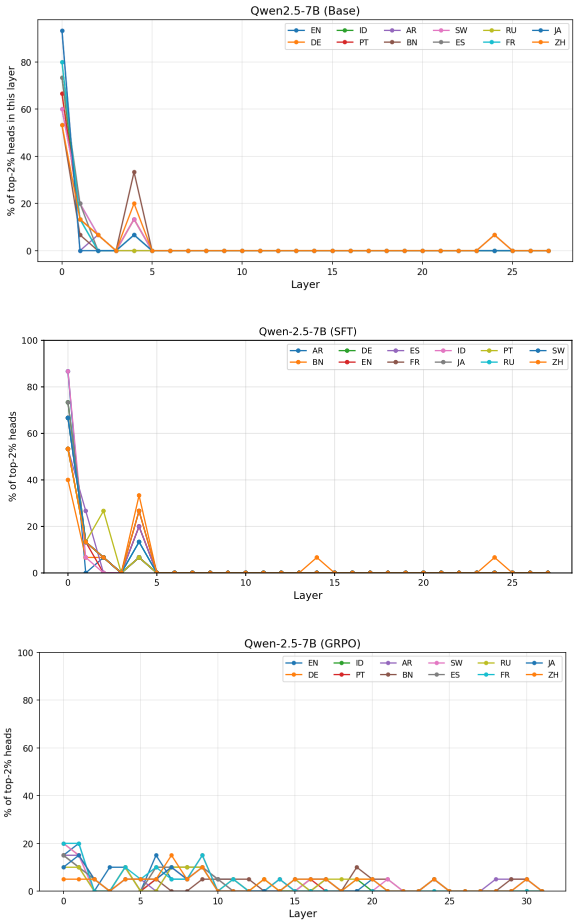
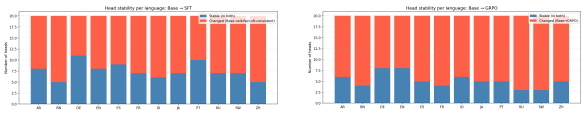
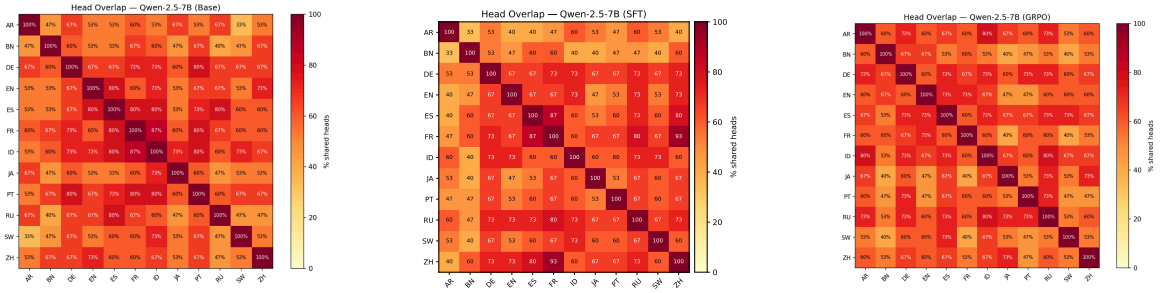


Figure 11: Percentage of language-important attention heads per layer for Qwen-2.5-7B before and after post-training. GRPO reduces the concentration of language-specific heads in the earliest layers and distributes language routing more broadly across the network.



(a) Base → SFT (b) Base → GRPO

Figure 12: Number of stable (blue) and changed (orange) language-important heads per language.



(a) Base (b) SFT (c) GRPO

Figure 13: Pairwise overlap of language-important attention heads across language pairs for Qwen-2.5-7B before and after post-training.

Model	en	de	es	fr	pt	id	ru	zh	ar	ja	sw	bn	High
Baseline	56.49	42.46	40.70	41.75	36.84	37.19	31.58	31.93	33.68	32.98	29.82	25.26	40.25
Aligned	54.04	34.04	32.98	34.74	31.23	29.12	28.42	32.98	26.32	33.68	29.47	25.26	37.41
GRPO	55.79	42.81	38.60	43.86	38.60	38.25	32.98	34.04	29.12	32.98	31.23	28.42	40.95
Aligned + GRPO	54.39	41.75	38.60	41.75	35.09	32.28	33.33	35.09	26.32	32.28	29.82	27.37	39.43
SFT	54.04	40.35	40.35	41.05	37.19	34.39	33.68	32.63	31.58	32.28	29.82	23.51	39.90
Aligned + SFT	54.74	41.75	36.49	41.40	37.89	35.44	32.28	30.88	33.68	31.58	29.82	23.86	39.35

Table 11: Global-MMLU accuracy (%) across languages. High denotes the average over high-resource languages.

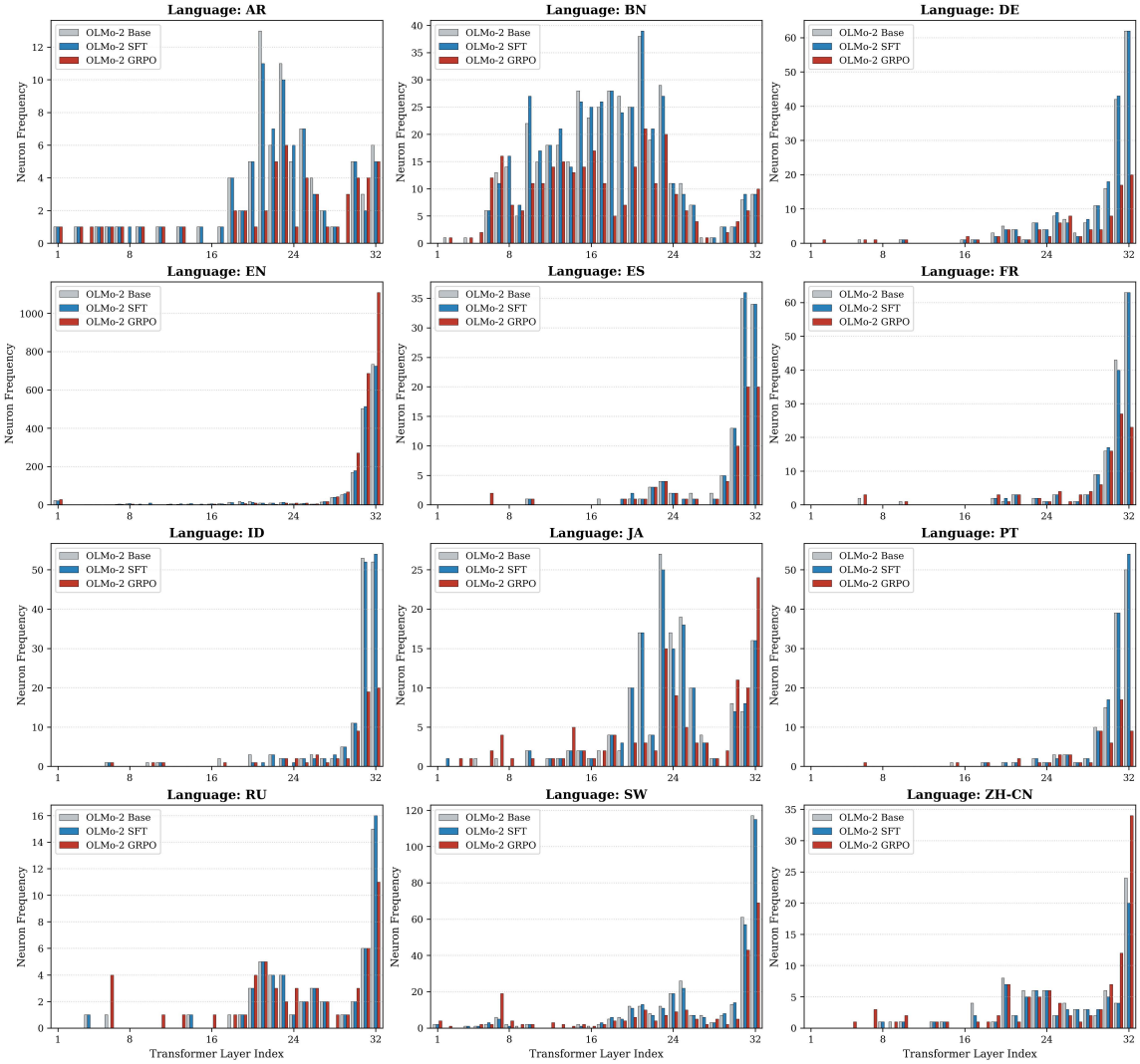


Figure 14: Layer-wise frequency of language-specific neurons for all target languages. Note the distinct English (EN) surge and the redistribution of neurons in non-Latin scripts (AR, JA, ZH-CN) under the GRPO configuration (red).

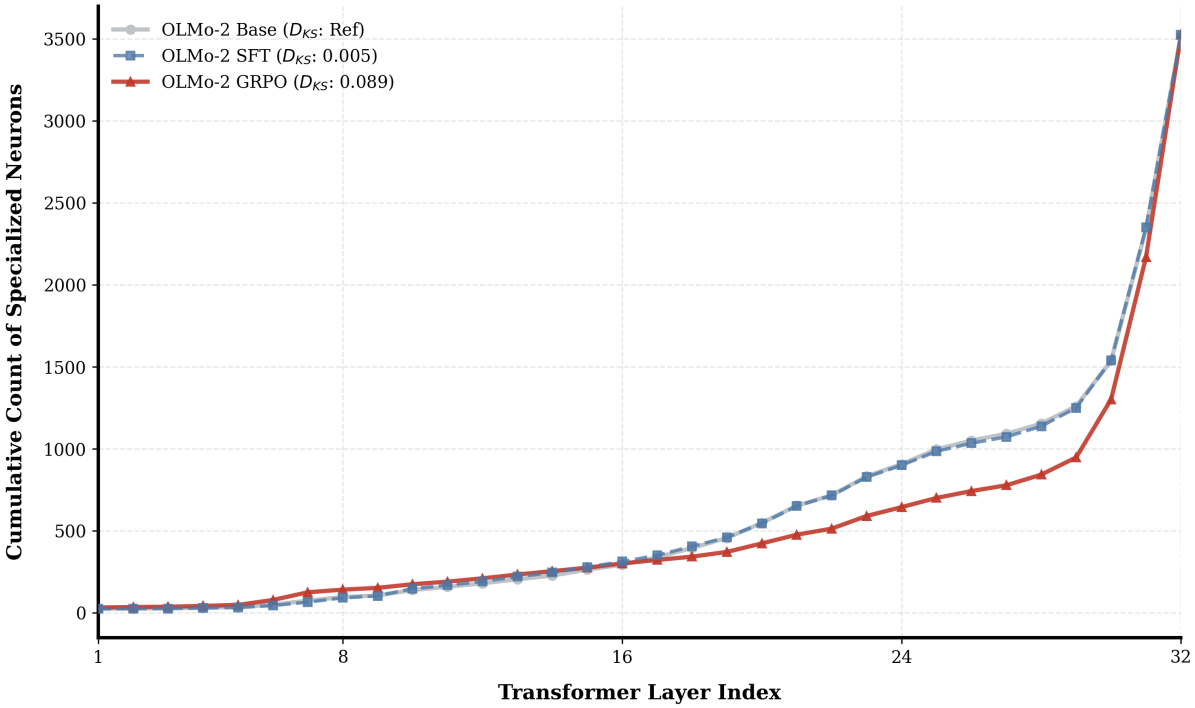


Figure 15: Empirical Cumulative Distribution Function (ECDF) of specialised neuron discovery across layers. The significant D_{KS} shift in GRPO indicates a deferral of linguistic specialisation to later layers.

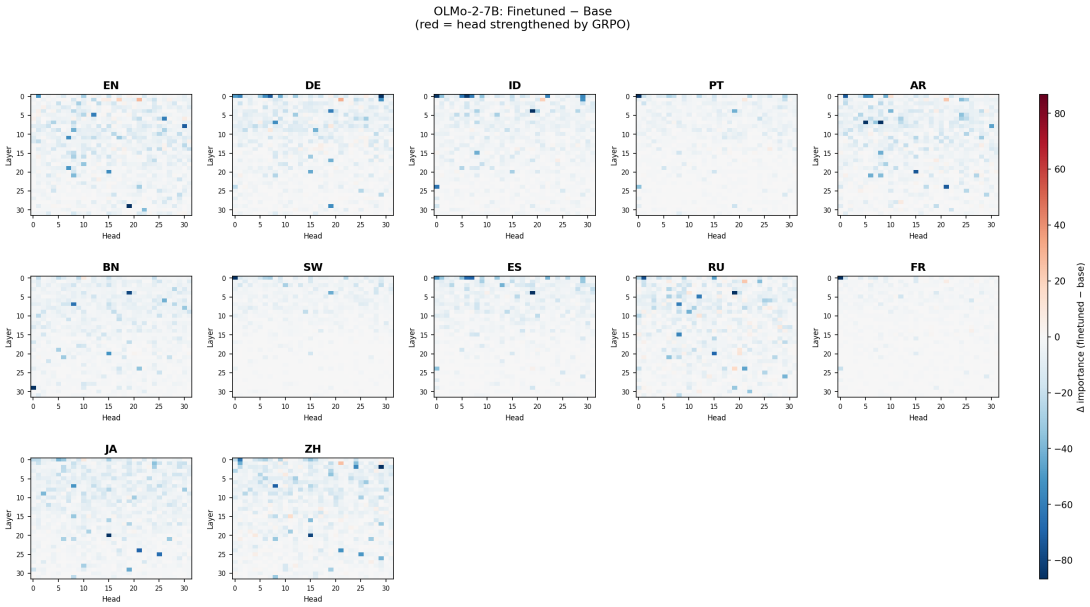


Figure 16: Delta heatmap (GRPO − base). Blue indicates weakened heads, red indicates strengthened heads.

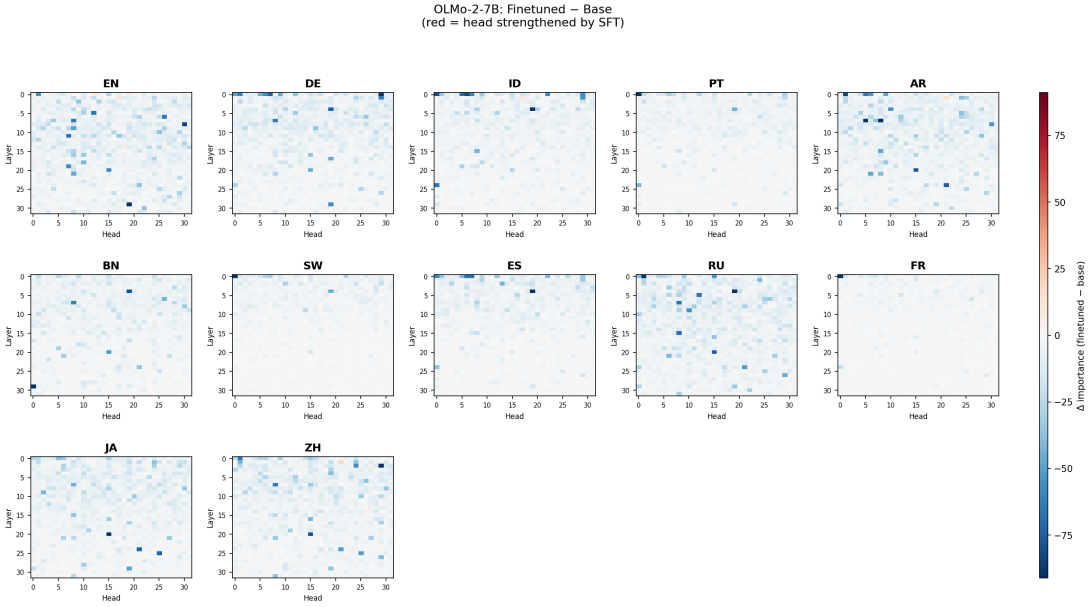


Figure 17: Delta heatmap (SFT-consistent − base). Pattern closely mirrors GRPO, suggesting both methods suppress similar heads.