Aligning LLMs for Multilingual Consistency in Enterprise Applications

Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel,
Tao Sheng, Sujith Ravi, Dan Roth

Oracle AI

Correspondence: amit.h.agarwal@oracle.com

Abstract

Large language models (LLMs) remain unreliable for global enterprise applications due to substantial performance gaps between high-resource and mid/low-resource languages, driven by English-centric pretraining and internal reasoning biases. This inconsistency undermines customer experience and operational reliability in multilingual settings such as customer support, content moderation, and information retrieval. Even with advanced Retrieval-Augmented Generation (RAG) systems, we observe up to an 29% accuracy drop in non-English languages compared to English.

We propose a practical, batch-wise alignment strategy for fine-tuning LLMs, leveraging semantically equivalent multilingual data in each training batch to directly align model outputs across languages. This approach improves non-English accuracy by up to 23.9% without compromising English performance, model reasoning, or retrieval quality. Our method is simple to implement, scalable, and integrates seamlessly with existing LLM training & deployment pipelines, enabling more robust and equitable multilingual AI solutions in industry.

1 Introduction

The demand for multilingual natural language processing systems that perform reliably and consistently across diverse languages has grown rapidly in global industries such as customer support, content delivery networks, and information retrieval (Dua et al., 2025; Agarwal et al., 2024). Despite advances in LLMs, significant performance disparities persist between high-resource languages (e.g., English) and low-resource languages (e.g., French, Arabic), primarily due to the pretraining data, favoring high-resource corpora (Xu et al., 2024). This gap undermines user experience, restricts the accessibility of AI technologies, and limits operational effectiveness for organizations serving multilingual populations worldwide.

In practical deployments, LLMs often exhibit reduced accuracy and reasoning inconsistencies when handling non-English inputs, especially within RAG frameworks (Lewis et al., 2020; Li et al., 2024b). For example, a customer service chatbot may provide precise responses in English but fail to maintain equivalent quality in other languages, thereby eroding user trust and satisfaction. These challenges underscore an urgent need for AI systems that ensure equitable and consistent performance across languages.

Benchmarks like NoMIRACL (Thakur et al., 2023) have advanced RAG across languages; however, these approaches typically optimize task-specific metrics without explicitly aligning the underlying multilingual reasoning and generation processes. Consequently, they fail to guarantee and evaluate consistency across languages. A key factor contributing to this problem is the English-centric internal reasoning of current LLMs (Zhao et al., 2024), where LLMs internally "think" in English, even when processing inputs in other languages; exacerbating cross-lingual inconsistencies, reliability in multilingual applications. Existing approaches typically focus on external retrieval or translation to bridge this gap (Nie et al., 2022).

To overcome these limitations, we propose a novel fine-tuning alignment strategy that explicitly aligns LLMs internal reasoning and generation processes across languages by leveraging semantically equivalent multilingual data within every training batch. Our approach reduces reliance on English-centric reasoning and improves cross-lingual consistency without introducing external translation or retrieval components. Our key contributions are summarized as follows:

- We empirically characterize and quantify the performance gap in LLM reasoning and generation across languages for semantically identical content, highlighting its practical impact.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 117–137
November 4-9, 2025

- We propose a simple yet effective batch-wise alignment strategy that fine-tunes LLMs to align internal multilingual generation processes without relying on external translation or retrieval systems.

- Our method achieves substantial improvements, up to **23.9%**, in non-English task accuracy, narrowing the performance gap with English and enhancing model consistency.

2 Related Work

LLMs have made substantial progress, yet still exhibit significant performance disparities between high-resource and low-resource languages, in part due to English-centric pretraining and internal reasoning biases (Zhao et al., 2024; Xu et al., 2024).

Recent research has further explored this space by penalizing inconsistent response languages (Zhang et al., 2024), applying cross-lingual instruction tuning with translation and mixed-task datasets (Zhu et al., 2023; Lin et al., 2024), and leveraging preference optimization with translation and reward modeling (She et al., 2024; Dang et al., 2024). In contrast, our approach achieves direct alignment of LLM reasoning and generation across languages by leveraging semantically equivalent multilingual data within each training batch, without relying on explicit translation, preference data, or reward models. This enables a simple and scalable alignment strategy suitable for industry (Zeng et al., 2025). See Appendix A.1 for an extended review.

3 Method

We now describe our batch-alignment approach, dataset construction, and the training paradigm.

3.1 Batch Alignment for Cross-Lingual Consistency

The core of our approach is a *batch alignment* strategy that modifies the batching process to explicitly encourage cross-lingual consistency. During training, each batch contains semantically equivalent instances of the same topic presented in different languages (Fig. 1). This setup exposes the model to the same semantic problem across multiple languages simultaneously, promoting aligned internal reasoning & generation, reducing dependence on English-centric intermediate steps.

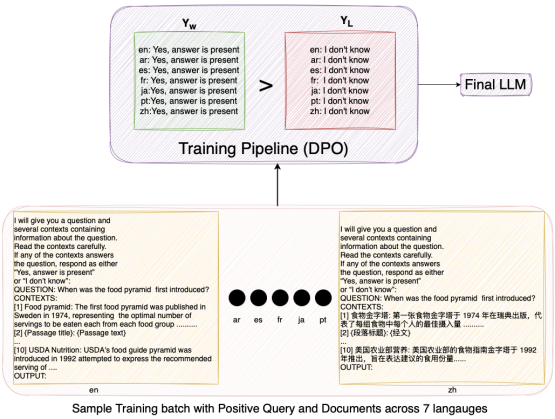

Figure 1: Highlights the training paradigm with the proposed batch-composition technique for a batch with Positive Samples. The batch consists of training samples, with the same Query-Document composition across English and non-English to maintain same semantic consistency during training, where the LLM is expected to respond; "Yes, answer is present" or "I don't know" for the training samples.

Batch Composition Each training batch includes k samples of the same topic, each in a different language. The batch size is determined by the number of languages included and controlled via ablation studies. This composition enforces the model to generate consistent outputs for semantically equivalent inputs across languages within a single update.

3.2 Multilingual Dataset Construction

To enable systematic evaluation of multilingual reasoning and generation, we construct a multilingual question-answering dataset based on the NoMIRACL RAG framework. The dataset consists 500 documents, across diverse business-related topics spanning domains such as science, technology, advertising, and marketing. Each topic is translated by humans into six non-English languages (Arabic, Spanish, French, Japanese, Portuguese, and Chinese) in addition to English, ensuring semantic equivalence across languages.

For each topic, two distinct queries are designed that can be answered using a single relevant document. The dataset is constructed as follows:

- **Positive Samples:** Each sample consists of a query along with one relevant document & nine irrelevant documents selected from the same business domain but unrelated to the query, to simulate challenging retrieval conditions.

- **Negative Samples:** Each sample consists of

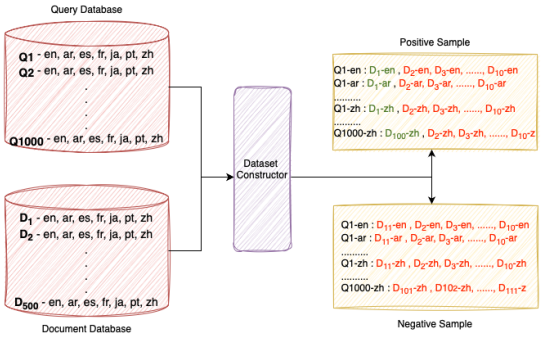

Figure 2: Highlights the high-level data-pipeline used to create the dataset. Each query is mapped by the Dataset Constructor to create a positive & negative sample with corresponding documents of the respective language.

a query along with ten irrelevant documents from the same domain, ensuring that no document directly answers the query.

This design results in 2,000 samples (positive & negative) per language. To ensure a fair cross-lingual evaluation, all contextual information, including queries and documents, is semantically consistent across languages. This controlled data set facilitates a rigorous analysis of multilingual reasoning and generation performance. Figure 2 illustrates our pipeline for constructing datasets, where queries and documents from human-annotated and translated databases are combined to form structured positive and negative samples for batch alignment training in a RAG setup. Additional details of prompt design and design rational are in the Appendix A.3.2 & A.3.3.

3.3 Alignment Finetuning

We adopt recent preference optimization methods: Direct Preference Optimization (DPO) (Rafailov et al., 2024) and Odds-Ratio Preference Optimization (ORPO) (Hong et al., 2024b), to optimize cross-lingual alignment. These approaches use pairwise preference signals derived from multilingual sample rankings within batches to guide the model towards consistent reasoning across languages. Inspired by Shaham et al. (2024), who demonstrated that minimal multilingual instruction tuning yields substantial cross-lingual generalization, we extend this principle to batch-level optimization, leveraging fine-grained multilingual supervision.

Training Procedure Models are fine-tuned under the batch alignment paradigm with DPO and ORPO objectives. Standard finetuned models (uncoupled) use the same optimization methods, but with standard batch composition where samples from different topics and languages are randomly shuffled, providing no explicit multilingual alignment signal. Our method only changes batch composition; there are no extra critics/reward calls, no new parameters, and no inference overhead. Infrastructure details are available at Appendix A.3.

4 Experimental Setup

We describe our controlled RAG experimental setup, evaluation metrics, and the core research questions; see Appendix A.3 for further details.

4.1 Controlled RAG Setup

To isolate the effect of our alignment strategy on multilingual generation, we conduct all evaluations within a RAG framework where the retrieval component is fixed and identical across languages. Specifically, for each query, the same set of documents is used, translated equivalently across languages, ensuring that the quality of the retrieval does not confound the performance of the generation. Furthermore, based on the documents in the context, the LLM has to respond according to the prompted instruction (see Appendix A.3.2).

Instruction Language Control. We fix the instruction/query template to English to control prompt-style variance and isolate cross-lingual conditioning from the content language. Documents are fully localized per language; only the instruction remains constant. This removes prompt-translation confounds in our controlled RAG probe — retrieval is fixed and identical across languages, so differences reflect generation/alignment rather than retriever noise or prompt phrasing. The sampler itself is anchor-agnostic, but including English as a high-resource anchor yields the best non-English gains (see Fig. 4), while English remains stable. Extended discussion and localized-prompt variants appear in Appendix A.3.3.

Evaluation Setup. Our dataset of 2,000 samples per language is split into 70% training & 30% testing, balanced across positive and negative samples. Accuracy is used as the primary evaluation metric due to the binary classification nature of the task: the model must determine if the answer is present or absent based on the query-document pair.

Rationale for Accuracy Metric. Given the controlled binary-response format (A.3.2) (i.e., “Yes,

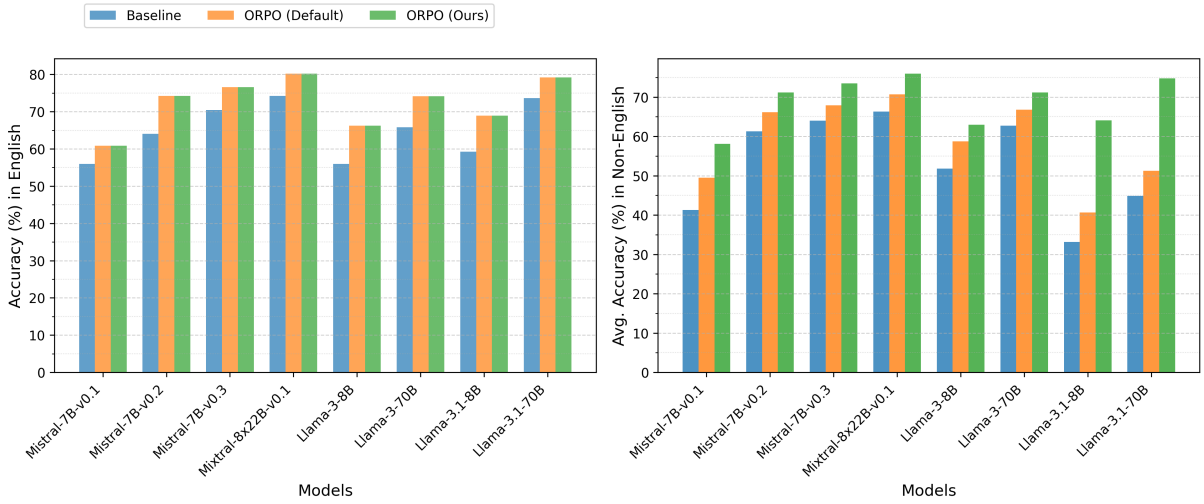

Figure 3: LLM (instruct versions) accuracy in English (left) and averaged over six non-English languages (right) for Baseline, Default ORPO (uncoupled), and ORPO with batch alignment (Ours). Batch alignment substantially improves non-English accuracy and cross-lingual consistency. See Table 8 for full results.

the answer is present" or "I don't know") uniformly applied across languages, exact match accuracy provides a direct, interpretable measure of reasoning and generation correctness. Since the retrieval is fixed, accuracy differences directly reflect the model's multilingual reasoning and generation consistency. Thus, additional multilingual consistency metrics are unnecessary in this controlled setting.

Statistical Reporting. We report point estimates (accuracy/exact match (EM) %, perplexity) and compute 95% bootstrap confidence intervals (B=1000) for all results. Pairwise comparisons on the same test items use McNemar's test for accuracy/EM (significance at $p < 0.01$) and paired bootstrap CIs for perplexity deltas. Claims of improvement are made only when the corresponding test is significant. *Rest (avg.)* is the macro-average over non-English languages. CIs are displayed in Table [X]; due to space, other tables/figures report point estimates, computed with the same procedure.

4.2 Research Questions

To rigorously evaluate the effectiveness and generalizability of our batch alignment strategy, we organize our experiments around the following research questions:

RQ1: Does the proposed batch alignment method improve multilingual consistency in reasoning and generation across languages?

RQ2: Can the alignment technique scale to unseen languages under the same task ?

RQ3: Does the proposed alignment approach generalize to out-of-domain tasks beyond RAG?

RQ4: Does batch-alignment enhance other language modeling aspects such as fluency and semantic understanding?

We conduct ablation studies to assess the influence of English as an anchor language, batch size, & the use of machine-translated data for alignment.

5 Results

We present empirical results for each research question. Baseline LLMs exhibit substantial multilingual inconsistency in the RAG setup, with an accuracy drop up to 29% in non-English languages (Table 5, Appendix B). The following subsections detail how our batch alignment strategy mitigates these gaps.

5.1 RQ1: Impact of Batch Alignment on Multilingual Consistency

We evaluated whether our proposed batch alignment strategy effectively improves multilingual consistency within RAG setups. Figure 3 compares the performance of three distinct setups: Baseline (pre-trained model), Default ORPO fine-tuning (uncoupled), and ORPO with batch alignment (ours): on English (left) and non-English languages (right).

English Results: ORPO fine-tuning significantly improves English accuracy over the baseline across all models. However, the incremental improvement from the default ORPO to ORPO with batch-alignment (ours) is minimal, indicating limited additional benefit for high-resource languages that

Model	English		Rest (avg.)	
	Baseline	ORPO (ours)	Baseline	ORPO (ours)
Mixtral-8x22B	74.2	80.2	41.7	44.3
Llama-3-1-70B	73.6	79.2	33.7	37.8
Llama-3-70B	65.8	74.1	32.6	35.9

Table 1: Accuracy (%) of baseline and ORPO with batch alignment (ours) on English and the average across four unseen languages (Rest (avg.)) demonstrating the cross-lingual generalization enabled by our alignment strategy.

are already well represented during pre-training.

Non-English Results: In contrast, our batch alignment strategy yields substantial gains in multilingual accuracy compared to the default ORPO fine-tuning. Models such as Llama-3.1, which initially exhibit large multilingual performance gaps, benefit the most from our alignment method. This trend is consistent across various models, emphasizing the universal advantage of explicit multilingual alignment in reducing inherent model biases. Table 6 and Table 8 in Appendix B provide detailed results across individual languages and models, further reinforcing these findings.

5.2 RQ2: Generalization to Unseen Languages

We evaluate generalization on four unseen languages (Thai, Vietnamese, Hungarian, Romanian). ORPO with batch-alignment (ours) consistently outperform baselines in these languages (Table 1); for example, Mixtral-8x22B and Llama-3-70B achieve accuracy gains of up to 4.1%. This demonstrates that our batch alignment strategy enhances intrinsic cross-lingual generalization beyond the languages seen during fine-tuning. Table 7 in Appendix B expands the per language break-up and confidence-interval results.

5.3 RQ3: Generalization to Out-of-Domain Tasks

Next, we validate our method's robustness by evaluating on more complex and diverse multilingual benchmarks such as Multilingual MMLU (MMMLU) and MGSM, which require extensive reasoning beyond retrieval-based QA. Table 2 clearly shows that our aligned models consistently outperform the pre-trained baselines on the MMMLU benchmark. We observe average performance gains of 3.9% for DPO (ours) and 4.2% for ORPO (ours) across multiple models, highlighting our method's robust generalization and its effectiveness in enhancing complex multilingual reasoning.

Similarly, on the challenging MGSM benchmark (Table 3), our alignment strategy significantly boosts exact-match accuracy across all models. Larger models like Llama-3.1-70B achieve striking improvements (from 57.2% to 74.9%), underscoring the scalability and effectiveness of our batch alignment approach for sophisticated multilingual reasoning tasks.

Models	**Baseline**	**DPO (Ours)**	**ORPO (Ours)**
Mistral-7B-v0.1	45.6	48.8	48.8
Mistral-7B-v0.2	47.7	49.3	50.0
Mistral-7B-v0.3	50.7	52.6	53.1
Mixtral-8x22B-v0.1	75.2	77.2	77.6
Llama-3-8B	49.0	59.4	59.5
Llama-3-70B	73.7	76.5	76.8
Llama-3.1-8B	61.8	68.4	68.4
Llama-3.1-70B	79.9	82.6	83.0

Table 2: Accuracy of baseline & batch-aligned models (Ours) on the MMMLU benchmark, highlighting consistent gains from our approach.

Model	Avg. Perf (Baseline)	Avg. Perf (ORPO Ours)
Mistral-7B-v0.1	14.4	21.3
Mistral-7B-v0.2	19.5	25.0
Mistral-7B-v0.3	15.3	27.0
Mixtral-8x22B-v0.1	26.7	33.8
Llama-3-8B	28.1	34.9
Llama-3-70B	67.9	71.8
Llama-3.1-8B	30.1	39.3
Llama-3.1-70B	57.2	74.9

Table 3: Average Exact Match Accuracy of each model on MGSM benchmark, baseline and batch-aligned (ORPO ours) model.

5.4 RQ4: Broader Language Modeling Improvements

Finally, we investigate whether batch alignment enhances broader aspects of language modeling beyond accuracy, focusing specifically on fluency and semantic comprehension. To this end, we evaluate on LAMBADA (multilingual), a close-style benchmark measuring contextual fluency and prediction, and PAWS-X, which tests cross-lingual paraphrase recognition and semantic understanding.

Table 4 reports performance on the LAMBADA (multilingual) and PAWS-X benchmarks. Our ORPO batch-aligned models consistently achieve lower perplexity scores and higher accuracy on LAMBADA, demonstrating improved fluency and coherence in multilingual text generation. Similarly, PAWS-X accuracy increases significantly

Model	LAMBADA Perplexity ↓		LAMBADA Accuracy (%) ↑		PAWS-X Accuracy (%) ↑	
	Baseline	ORPO (ours)	Baseline	ORPO (ours)	Baseline	ORPO (ours)
Mistral-7B-v0.1	54.9	49.8	30.6	37.3	61.8	63.9
Mistral-7B-v0.2	37.7	33.1	34.0	38.5	64.3	69.9
Mistral-7B-v0.3	29.2	24.8	36.2	39.9	63.2	69.7
Mixtral-8x22B-v0.1	11.0	8.2	44.8	49.7	61.0	64.1
Llama-3-8B	35.9	28.5	34.8	40.8	62.4	65.6
Llama-3-70B	27.2	19.5	37.1	42.6	59.4	64.9
Llama-3.1-8B	33.3	25.7	35.3	41.9	65.4	70.1
Llama-3.1-70B	21.1	15.0	38.7	43.4	60.7	66.1

Table 4: Performance comparison of Baseline and ORPO batch-aligned (ours) models on two benchmarks: LAMBADA for Perplexity (lower is better) & Accuracy (%), and PAWS-X for Accuracy (%).

across all tested models, indicating enhanced semantic understanding and better multilingual paraphrase recognition capabilities.

These broader language modeling improvements affirm the comprehensive benefits of our batch alignment method, extending beyond our controlled binary RAG probe: the same gains hold on MMMLU, MGSM, PAWS-X, and LAMBADA and are significant, enhancing the linguistic quality and semantic precision of multilingual models. We provide detailed results for each model and language in Appendix B.

5.5 Qualitative Error Analysis

Where errors arise. In our controlled binary RAG probe, retrieval, index, and negatives are held fixed across languages, and outputs are constrained to *Yes* / *I don't know*. Residual mistakes therefore reflect *generation-/alignment-/calibration*-side issues rather than retrieval or grading artifacts.

Patterns relative to English. Because the instruction/query template is fixed to English while documents are localized, we frequently observed *co-occurrence of errors*: items wrong in English remain wrong in non-English runs, indicating model-intrinsic reasoning limits rather than language-specific lexical issues. Conversely, we do not observe items correct in non-English but wrong in English, consistent with English remaining a strong anchor.

Data validation. The documents/queries are factual and human-translated; we found no systematic cultural or language bias. We also verified the translated data against each model's tokenizer and observed no unknown-token artifacts (e.g., `[unk]` token).

6 Ablation Studies

Key ablation findings are summarized below; full details are in Appendix C.

High-Resource Languages (C.1): Adding English as an anchor in training consistently boosts non-English accuracy, e.g., Mixtral-8x22B rises from 66.3% to 76.0%,with an average gain of 9.8% across models (Fig. 4, Table 9 & 10).

Machine-Translated Data (C.2): Training with machine-translated data yields strong gains, though slightly below manual translations, confirming scalability and practicality for batch-alignment (Fig. 5).

Batch Size (C.3): Larger batch sizes enhance multilingual accuracy by providing richer cross-lingual signals. Performance gains are consistent at higher batch sizes, underscoring the benefit of structured batch diversity in alignment (Fig. 6).

Model Size (C.4): We observe a weak positive correlation between model size and language improvement for both DPO and ORPO (Fig. 7), indicating that simply increasing model size does not fully close the multilingual performance gap.

7 Discussion

Our experiments demonstrate that batch-alignment is an effective and robust strategy for improving multilingual consistency in LLMs across varied tasks and benchmarks.

Impact on Multilingual Consistency. Batch alignment consistently improves non-English accuracy without affecting English performance (RQ1). We find that including high-resource languages, especially English, as anchors in each batch is crucial for reducing multilingual disparities, likely by facilitating internal cross-lingual alignment within the model, For instance, including English increases non-English accuracy by an average of 9.8% across models (see Fig. 4).

Generalization to Unseen Languages and Tasks. The ability to generalize to unseen languages (RQ2) further emphasizes the intrinsic multilingual capabilities enhanced by batch alignment. Our batch-aligned (ORPO) models not only outperform base-

line models on seen languages but also exhibit measurable improvements on languages never encountered during fine-tuning. This outcome underscores the broader applicability of our technique beyond the original training context.

Moreover, results on out-of-domain benchmarks such as MMMLU & MGSM (RQ3) provide strong evidence for the broader generalizability of our alignment approach. Our method notably enhances multilingual reasoning and knowledge-intensive tasks, showcasing its value beyond retrieval-based scenarios. Such generalization is particularly relevant for practical deployments where model adaptability across tasks and domains is crucial.

Broader Implications for Language Modeling. Beyond accuracy improvements, batch alignment substantially improves model fluency, coherence, and semantic understanding, as demonstrated by reductions in perplexity on LAMBADA (multilingual) & accuracy gains on the PAWS-X benchmark (RQ4). These broader improvements underscore that our alignment method fundamentally enhances multilingual language modeling rather than merely optimizing task-specific metrics. Thus, our technique contributes significantly towards building more reliable, coherent, and linguistically nuanced multilingual models.

Insights from Ablation Studies. Our detailed ablation studies provide critical insights as we observe that larger batch sizes consistently enhance multilingual accuracy, with performance gains persisting across all tested batch sizes & models, emphasizing the critical role of cross-lingual signal diversity within batches. Additionally, experiments with machine-translated data indicate that while manual translation remains superior, machine-translated data also delivers substantial performance improvements (up to 15.1%). This finding significantly enhances the practical feasibility of our method, especially in resource-limited scenarios.

8 Conclusion

We present a practical batch alignment strategy for fine-tuning large language models, substantially improving multilingual consistency in the generation component of RAG systems. Our approach, built on DPO and ORPO, yields up to a **23.9% gain in non-English task accuracy** with no trade-off in English performance, closing the multilingual gap critical to real-world deployments.

This method is simple to integrate into existing LLM pipelines, requires only parallel or machine-translated data, and delivers immediate benefits for industry applications, enabling robust, equitable customer support, virtual assistants, and information retrieval in any language. Our ablation studies offer actionable guidance: English anchoring is essential, machine translation is viable for scaling, and batch size can be tuned to resource constraints.

By making multilingual alignment more accessible and practical, we empower global enterprises to serve diverse linguistic markets, reduce operational risk, and increase user satisfaction. Our findings suggest that LLMs possess latent multilingual reasoning and generation capabilities, which can be surfaced and enhanced through targeted alignment, opening new directions for robust, equitable AI across languages.

9 Limitations

While our findings demonstrate robust improvements, several limitations offer directions for future research. First, our study primarily focuses on high-resource languages as alignment anchors, raising the question of how low-resource language alignment might benefit from leveraging intermediate-resource languages. Second, despite notable improvements, the reliance on preference-based optimization methods such as DPO and ORPO introduces complexity in training setups, warranting exploration of simpler and more scalable optimization techniques. Finally, evaluating alignment strategies in highly noisy or real-world user-generated or synthetic multilingual datasets remains an open challenge, essential for practical deployments.

Furthermore, the study does not explore the impact of long, noisy sequences that are often present in real-world multilingual data. Such sequences may adversely affect the alignment process and model performance. Future work should address these gaps by expanding language coverage, evaluating the effect of noisy sequences, and assessing the viability of synthetic data for training.

10 Ethical Considerations

Our work aims to advance equitable access to AI by reducing performance disparities between high-resource and low-resource languages, thereby supporting fairer and more inclusive multilingual technologies. By promoting consistency across languages, our approach can help mitigate language-based biases that have historically limited access

to high-quality AI services for underrepresented linguistic communities.

However, ethical deployment also requires careful attention to data quality & representation. Over-reliance on machine-translated/synthetic data could inadvertently amplify errors or encode unintended biases from dominant language corpora. We recommend ongoing benchmarking & transparent reporting of multilingual model behavior, especially for languages with limited resources, to minimize potential harms & maximize societal benefit.

References

Amit Agarwal. 2021. Evaluate generalisation & robustness of visual features from images to video. *ResearchGate*. Available at https://doi.org/10.13140/RG.2.2.33887.53928.

Amit Agarwal, Srikant Panda, Angeline Charles, Hitesh Laxmichand Patel, Bhargava Kumar, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, Hansa Meghwani, Karan Gupta, and Dong-Kyu Chae. 2025a. MVTamperBench: Evaluating robustness of vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18804–18828, Vienna, Austria. Association for Computational Linguistics.

Amit Agarwal, Srikant Panda, and Kulbhushan Pachauri. 2024. Synthetic document generation pipeline for training artificial intelligence models. US Patent App. 17/994,712.

Amit Agarwal, Srikant Panda, and Kulbhushan Pachauri. 2025b. FS-DAG: Few shot domain adapting graph networks for visually rich document understanding. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 100–114, Abu Dhabi, UAE. Association for Computational Linguistics.

Amit Agarwal, Hitesh Laxmichand Patel, Srikant Panda, Hansa Meghwani, Jyotika Singh, Karan Dua, Paul Li, Tao Sheng, Sujith Ravi, and Dan Roth. 2025c. Rci: A score for evaluating global and local reasoning in multimodal benchmarks. *Preprint*, arXiv:2509.23673.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, et al. 2024. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *Proceedings of NAACL 2024*, pages 1771–1800.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. Rlhf can speak many languages: Unlocking multilingual preference optimization for llms, 2024. *URL https://arxiv. org/abs/2407.02552*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Karan Dua, Puneet Mittal, Ranjeet Gupta, and Hitesh Laxmichand Patel. 2025. SpeechWeave: Diverse multilingual synthetic text & audio data generation pipeline for training text to speech models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 718–737, Vienna, Austria. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

EleutherAI. 2024. Lambada multilingual dataset. https://github.com/EleutherAI/lm-evaluation-harness/tree/main/lm_eval/tasks/lambada_multilingual. Accessed: 2025-06-07.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024a. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Jiwoo Hong, Noah Lee, and James Thorne. 2024b. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Bryan Li, Tamer Alkhouli, Daniele Bonadiman, Nikolaos Pappas, and Saab Mansour. 2024a. Eliciting better multilingual structured reasoning from llms through code. *arXiv preprint arXiv:2403.02567*.

Bryan Li, Samar Haider, Fiona Luo, Adwait Agashe, and Chris Callison-Burch. 2024b. Bordirlines: A dataset for evaluating cross-lingual retrieval-augmented generation. *arXiv preprint arXiv:2410.01171*.

Geyu Lin, Bin Wang, Zhengyuan Liu, and Nancy F Chen. 2024. Crossin: An efficient instruction tuning approach for cross-lingual knowledge alignment. *arXiv preprint arXiv:2404.11932*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Yinquan Lu, Wenhao Zhu, Lei Li, et al. 2024. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*.

Hansa Meghwani, Amit Agarwal, Priyaranjan Pattnayak, Hitesh Laxmichand Patel, and Srikant Panda. 2025. Hard negative mining for domain-specific retrieval in enterprise systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1013–1026, Vienna, Austria. Association for Computational Linguistics.

Sina Bagheri Nezhad and Ameeta Agrawal. 2024. What drives performance in multilingual language models? *arXiv preprint arXiv:2404.19159*.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2022. Cross-lingual retrieval augmented prompt for low-resource languages. *arXiv preprint arXiv:2212.09651*.

OpenAI. 2024. Multilingual massive multitask language understanding (mmmlu). https://huggingface.co/datasets/openai/MMMLU. Accessed: 2025-06-07.

Srikant Panda, Amit Agarwal, Gouttham Nambirajan, and Kulbhushan Pachauri. 2025a. Out of distribution element detection for information extraction. US Patent App. 18/347,983.

Srikant Panda, Amit Agarwal, and Hitesh Laxmichand Patel. 2025b. Accesseval: Benchmarking disability bias in large language models. *Preprint*, arXiv:2509.22703.

Srikant Panda, Vishnu Hari, Kalpana Panda, Amit Agarwal, and Hitesh Laxmichand Patel. 2025c. Who's asking? investigating bias through the lens of disability framed queries in llms. *Preprint*, arXiv:2508.15831.

Srikant Panda, Hitesh Laxmichand Patel, Shahad Al-Khalifa, Amit Agarwal, Hend Al-Khalifa, and Sharefah Al-Ghamdi. 2025d. Daiq: Auditing demographic attribute inference from question in llms. *Preprint*, arXiv:2508.15830.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Hitesh Laxmichand Patel, Amit Agarwal, Arion Das, Bhargava Kumar, Srikant Panda, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, and Dong-Kyu Chae. 2025a. Sweeval: Do llms really swear? a safety benchmark for testing limits for enterprise use. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 558–582.

Hitesh Laxmichand Patel, Amit Agarwal, Bhargava Kumar, Karan Gupta, and Priyaranjan Pattnayak. 2024. Llm for barcodes: Generating diverse synthetic data for identity documents. *arXiv preprint arXiv:2411.14962*.

Hitesh Laxmichand Patel, Amit Agarwal, Srikant Panda, Hansa Meghwani, Karan Dua, Paul Li, Tao Sheng, Sujith Ravi, and Dan Roth. 2025b. Pcri: Measuring context robustness in multimodal models for enterprise applications. *Preprint*, arXiv:2509.23879.

Priyaranjan Pattnayak, Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, and Srikant Panda. 2025a. Hybrid ai for responsive multi-turn online conversations with novel dynamic routing and feedback adaptation. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 215–229.

Priyaranjan Pattnayak, Hitesh Laxmichand Patel, and Amit Agarwal. 2025b. Tokenization matters: Improving zero-shot ner for indic languages. *Preprint*, arXiv:2504.16977.

Priyaranjan Pattnayak, Hitesh Laxmichand Patel, Amit Agarwal, Bhargava Kumar, Srikant Panda, and Tejaswini Kumar. 2025c. Clinical qa 2.0: Multi-task learning for answer extraction and categorization. *Preprint*, arXiv:2502.13108.

Priyaranjan Pattnayak, Hitesh Laxmichand Patel, Bhargava Kumar, Amit Agarwal, Ishan Banerjee, Srikant Panda, and Tejaswini Kumar. 2024. Survey of large multimodal model datasets, application categories and taxonomy. *arXiv preprint arXiv:2412.17759*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

T Pires. 2019. How multilingual is multilingual bert. *arXiv preprint arXiv:1906.01502*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *arXiv preprint arXiv:2401.01854*.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *arXiv preprint arXiv:2401.06838*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Jyotika Singh. 2021. Social media analysis using natural language processing techniques. In *Proceedings of the 20th Python in Science Conference*, SciPy, page 74–80. SciPy.

Jyotika Singh. 2022. pyaudioprocessing: Audio processing, feature extraction, and machine learning modeling. In *Proceedings of the 21st Python in Science Conference*, SciPy, page 152–158. SciPy.

Jyotika Singh. 2023. *Natural Language Processing in the Real World: Text Processing, Analytics, and Classification*. Chapman and Hall/CRC.

Ishan Tarunesh, Sushil Khyalia, Vishwajeet Kumar, Ganesh Ramakrishnan, and Preethi Jyothi. 2021. Meta-learning for effective multi-task and multilingual modelling. *arXiv preprint arXiv:2101.10368*.

Nandan Thakur, Luiz Bonifacio, Xinyu Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, et al. 2023. Nomiracl: Knowing when you don't know for robust multilingual retrieval-augmented generation. *arXiv preprint arXiv:2312.11361*.

Edwin Thomas, Amit Agarwal, Sandeep Jana, and Kulbhushan Pachauri. 2025. Model augmentation framework for domain assisted continual learning in deep learning. US Patent App. 18/406,905.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *arXiv preprint arXiv:2404.00929*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2024. Language imbalance driven rewarding for multilingual self-improving. *arXiv preprint arXiv:2410.08964*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Nan Yin, Mengzhu Wan, Li Shen, Hitesh Laxmichand Patel, Baopu Li, Bin Gu, and Huan Xiong. 2024. Continuous spiking graph neural networks. *arXiv preprint arXiv:2404.01897*.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617, Abu Dhabi, UAE. Association for Computational Linguistics.

Liang Zhang, Qin Jin, Haoyang Huang, Dongdong Zhang, and Furu Wei. 2024. Respond in my language: Mitigating language inconsistency in response generation based on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4177–4192.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

A Appendix

A.1 Extended Related Works

Multilingual Consistency in Large Language Models. Recent advances in multilingual LLMs aim to extend the capabilities of LLMs to a broad spectrum of languages, including low-resource and non-Latin scripts enabling systems like RAG. Nezhad and Agrawal (2024) analyzed factors influencing LLM performance across over 200 languages, showing that pretraining data size predominantly governs seen-language performance, while script type and language family are key for unseen languages. This underscores the importance of balanced and diverse pretraining data to achieve consistent multilingual performance. Zhao et al. (2024) studied the internal mechanisms of multilingual LLMs and revealed a prevalent "English-centric" reasoning phenomenon, where models often perform intermediate reasoning steps in English before generating outputs in the target language. This behavior contributes to cross-lingual inconsistencies across enterprise applications with global customer base, motivating methods targeting internal alignment.

Cross-Lingual Transfer and Preference Alignment. Cross-lingual knowledge transfer remains essential for robust multilingual LLMs. Zhu et al. (2024) assessed LLM translation capabilities, noting that although models like GPT-4 improve high-resource language performance, they still lag behind specialized supervised systems in low-resource scenarios. Fine-tuning has shown promise over in-context learning, particularly for low-resource languages (Asai et al., 2024). Additionally, Shaham et al. (2024) proposed minimal multilingual instruction tuning, demonstrating that even limited multilingual data can significantly enhance cross-lingual generalization and instruction following on unseen languages.

Recent research has expanded the landscape of multilingual alignment and consistency in LLMs. MAPO (She et al., 2024) employs multi-stage, translation-based preference optimization to align multilingual LLMs, while Language Imbalance Driven Rewarding (Yang et al., 2024) addresses training imbalances to improve fairness and accuracy. Zeng et al. (Zeng et al., 2025) show that LLMs naturally develop a shared internal representation during multilingual training, supporting the idea that intrinsic multilingual reasoning can be optimized through targeted alignment.

Several very recent works propose complementary strategies. Zhu et al., 2023 and Lin et al., 2024 explore cross-lingual instruction tuning, leveraging explicit translation tasks and mixed datasets to boost non-English and cross-lingual abilities, with a particular emphasis on data allocation and scaling. (Dang et al., 2024) and (She et al., 2024) advance preference-based optimization for multilingual LLMs, using reward models, synthetic or human preference data, and in some cases external translation models to align non-English outputs with English anchors.

In contrast, our work achieves multilingual alignment by directly leveraging semantically equivalent multilingual data within each training batch—**without requiring explicit translation supervision, reward models, or preference data**. This simple, batch-level alignment enables robust multilingual consistency and scalability for industry deployment. Our approach is orthogonal and complementary to recent advances, providing a more straightforward and resource-efficient alternative for practical multilingual LLM alignment.

Multilingual Consistency in RAG. Ensuring consistent multilingual reasoning and generation within RAG systems is a challenging problem, critical to enterprise applications which serve the same information across different languages to a global user-base (Pattnayak et al., 2025a; Meghwani et al., 2025; Singh, 2023, 2021). The NoMIRACL benchmark (Thakur et al., 2023) was introduced to rigorously evaluate RAG model's ability to maintain factual consistency across languages by pairing queries with relevant and irrelevant documents in multiple languages. (Zhang et al., 2024) focus on penalizing English responses to non-English prompts, addressing output language consistency without using multilingual data for fine-tuning, while effective, challenges remain due to differences in topic distributions and retrieval noise across languages and data types (Dua et al., 2025; Singh, 2022). Zhang et al. (2023) further demonstrated challenges faced by multilingual LLMs with code-switched inputs, emphasizing the need for specialized training to handle linguistically complex scenarios.

Recent benchmarks like BUFFET (Asai et al., 2024), MMMLU (OpenAI, 2024), MGSM (Cobbe et al., 2021; Shi et al., 2022), LAMBADA Multilingual (EleutherAI, 2024; Paperno et al., 2016), PAWS-X (Yang et al., 2019) and xSTREET (Li

et al., 2024a) reveal that LLMs still struggle to maintain consistent reasoning across languages, indicating that fundamental challenges remain; motivating the need for alignment-focused methods.

Our approach advances this by proposing a batch-wise alignment strategy using semantically equivalent multilingual data within each training batch, enabling direct alignment of reasoning and generation across languages during fine-tuning without dependence on external translation, rewards models or retrieval components.

Extensive Pre-training for Improved Translation. Lu et al., 2024 addressed gaps in multilingual capabilities by developing LLaMAX, a language model designed for enhanced translation across over 100 languages. Their approach involved extensive multilingual pretraining with data augmentation and vocabulary expansion, leading to significant improvements in both high-resource and low-resource translation tasks. This study demonstrates the importance of robust multilingual pretraining to boost LLMs' language generation capabilities.

Multilingual Pretraining and LLMs. Multilingual pretraining has been central to improving the performance of LLMs across different languages, providing a means to develop models that generalize effectively across diverse linguistic domains. Early models like mBERT (Devlin, 2018; Pires, 2019) laid the groundwork by showing that shared embeddings could facilitate cross-lingual transfer. However, models such as XLM-R (Conneau, 2019), with its large-scale pretraining on 2 terabytes of CommonCrawl data across 100 languages, demonstrated significant improvements in multilingual understanding, outperforming mBERT on benchmarks such as XNLI (Conneau et al., 2018) and MLQA (Lewis et al., 2019).

mT5 (Xue et al., 2021) further improved on these models by unifying the text-to-text framework for multilingual tasks, supporting a wide range of NLP tasks across 101 languages. LaBSE v2, released in 2023, builds on LaBSE (Feng et al., 2020) and employs enhanced language-agnostic training objectives to achieve superior cross-lingual sentence embeddings. This model is particularly well-suited for sentence-level tasks like retrieval and paraphrasing across 200 languages, including many low-resource languages, offering improvements over prior models by incorporating alignment techniques for better sentence representation. Furthermore, XGLM (Lin et al., 2021), an autoregressive language model trained on massive multilingual data, has become a recent state-of-the-art approach for multilingual tasks, using a probabilistic approach to model language sequences.

Multilingual Fine-tuning Strategies. Fine-tuning multilingual models for specific tasks like document understanding (Agarwal et al., 2025b; Yin et al., 2024; Patel et al., 2024; Panda et al., 2025a), video understanding (Agarwal, 2021; Agarwal et al., 2025a; Thomas et al., 2025), clinical trials (Pattnayak et al., 2025c), and accessible applications (Panda et al., 2025b,c,d) has evolved rapidly in recent years along with applications requiring both multi-modal and multilingual capabilities from models (Pattnayak et al., 2024). A key advancement is the use of adapter layers (Houlsby et al., 2019), which enable task-specific fine-tuning without modifying the entire model, resulting in reduced computational overhead. AdapterFusion (Pfeiffer et al., 2020) further extends this by fusing knowledge from multiple task-specific adapters, allowing for better transfer across tasks and languages.

A growing body of research has shown that fine-tuning on typologically similar languages can improve performance on low-resource languages (Pattnayak et al., 2025b). Lauscher et al., 2020 demonstrated the effectiveness of cross-lingual transfer learning through multilingual fine-tuning, showing that models fine-tuned on a high-resource language (e.g., Spanish) could improve performance on a related low-resource language (e.g., Catalan).

Recent studies have also explored multilingual multi-task learning (MTL) as a fine-tuning strategy. Tarunesh et al., 2021 proposed a new multi-task learning approach that fine-tunes multilingual models across several NLP tasks simultaneously, including machine translation, summarization, and question answering. Their results showed that training models on multiple tasks simultaneously can lead to performance improvements in individual tasks, particularly for low-resource languages.

Alignment Strategies for Multilingual Models. Beyond retrieval, aligning multilingual models internally has gained interest, which we further study in Appendix A.1. Recent methods explore multi-task and preference-based fine-tuning to reduce language biases and improve consistency (Patel et al., 2025a; Rafailov et al., 2024; Hong et al., 2024b; She et al., 2024; Yang et al., 2024; Zeng et al., 2025). Our approach advances this by

proposing a batch-wise alignment strategy using semantically equivalent multilingual data within each training batch, enabling direct alignment of generation across languages during fine-tuning without dependence on external translation or retrieval components.

A.2 Preliminaries

Direct Preference Optimization (DPO). DPO (Rafailov et al., 2024) is an implicit reward model for RLHF (Bai et al., 2022; Kirk et al., 2023) that ensures optimal policy to solve RLHF with a standard classification loss without needing RL to align human preference. Intuitively, it ensures a robust LLM fine-tuning without excessive hyperparameter tuning in a RL-free training environment, aligning LLMs with human preferences. Likewise, DPO is a straightforward method that can maximize the reward with KL-divergence constraint and find optimal policy based on a binary cross-entropy objective that enables simple training paradigm without additional RL overhead.

Odds-Ratio Preference Optimization (ORPO). OPRO (Hong et al., 2024a) is a supervised fine-tuning (SFT) based preference alignment optimization method. Unlike DPO and RLHF, which typically need a separate reference model with SFT, whereas ORPO does not require a reference model by weakly penalizing undesired generations and pass strong adaptation signal for selective responses with an odd-ratio term.

A.3 Extended Experimental Setup

Evaluation & Dataset. We first evaluate pre-trained LLM performance & instruction tuned LLM using standard DPO/ORPO training without batch modification as a control. We then apply our modified batch construction, where each batch contains different language versions of the same topic. The dataset is designed to provide consistent context across languages, and is split 70% for training, and 30% for testing, ensuring balanced representation across topics and business domains.

Evaluation Models. We evaluate eight open-sourced models,including three versions of the Mistral-7B (Jiang et al., 2023) model, specifically v0.1, v0.2, v0.3, Mixtral-8x22B (Jiang et al., 2024) model, and four models from the Llama family (Touvron et al., 2023; Dubey et al., 2024), specifically one 8B and one 70B model from versions 3 and 3.1, respectively. Unless otherwise stated, all LLMs referenced in this work are their instruct/tuned versions.

Training Pipeline. Figure 1 highlights the training paradigm for the proposed batch-alignment training. The LLMs are fine-tuned using both standard and modified DPO/ORPO training methods. We use a learning rate of 5e-6, batch size of 7, and a maximum sequence length of 4096 tokens. Training is performed for 3 epochs using mixed-precision on 4 NVIDIA A100 GPUs to reduce memory footprint and improve computational efficiency.

A.3.1 Metrics for Evaluation

We use Accuracy as the primary evaluation metric, as it effectively captures both true positives and true negatives, reflecting the model's ability to reason consistently across multiple languages. Since our setup ensures that the model is expected to generate predefined responses (i.e., "Yes, the answer is present" or "I don't know"), we use exact match accuracy to determine correctness against the ground truth. Unlike the NoMIRACL benchmark, which evaluates Hallucination Rate and Error Rate, we argue that Accuracy is a more comprehensive measure of reasoning consistency in multilingual settings. Hallucination Rate focuses solely on incorrect additions of information, while Error Rate does not differentiate between systematic generation errors and retrieval-induced inconsistencies. In contrast, Accuracy reflects the true end-to-end reasoning capability of the model across different languages, making it the most direct measure of multilingual consistency.

While accuracy is well-suited to our binary QA formulation, we acknowledge the value of complementary metrics (such as semantic similarity or human evaluation) for more nuanced assessments of multilingual alignment across diverse textual and visual tasks (Agarwal et al., 2025c; Patel et al., 2025b).

A.3.2 Prompt Design

Following NoMiracl, we use a structured prompt to query the LLM during training and evaluation. The prompt is designed to provide a consistent query format across all languages:

```
I will give you a question and
several contexts containing
information about the question.
Read the contexts carefully.
If any of the contexts answers
```

Models	Baseline							
	English	Arabic	Spanish	French	Japanese	Portuguese	Chinese	Rest (avg.)
Mistral-7B-Instruct-v0.1	56.0	7.0	51.4	50.6	37.6	51.4	49.8	41.3
Mistral-7B-Instruct-v0.2	64.0	57.2	63.0	61.0	61.2	63.6	61.6	61.3
Mistral-7B-Instruct-v0.3	70.4	55.6	65.8	66.8	62.8	70.6	62.6	64.0
Mixtral-8x22B-Instruct-v0.1	74.2	57.2	71.2	71.4	65.4	69.8	62.8	66.3
Llama-3-8B-Instruct	56.0	50.6	57.4	44.0	49.6	54.8	54.4	51.8
Llama-3-70B-Instruct	65.8	60.8	64.4	60.9	63.2	63.4	63.8	62.8
Llama-3.1-8B-Instruct	59.2	24.8	38.4	6.8	49.4	29.0	50.4	33.1
Llama-3.1-70B-Instruct	73.6	45.6	19.6	50.4	65.8	22.6	65.4	44.9

Table 5: Accuracy results of pre-trained LLMs for English and six non-English languages. The 'Rest (avg.)' column shows the average accuracy for non-English languages, highlighting a performance gap with English.

```
the question, respond as either
“Yes, answer is present”
or “I don’t know”:
QUESTION: {query}
CONTEXTS:
[1] {Passage title}: {Passage text}
[2] {Passage title}: {Passage text}
...
[10] {Passage title}: {Passage text}
OUTPUT:
```

The prompt in our dataset is designed to simulate a typical RAG system, where each passage is retrieved from a knowledge source. During training and evaluation, every passage addresses the same topic but is presented in a different language.

A.3.3 Design Rationale

We use a single relevant document per query, in line with the NoMIRACL framework, to ensure a controlled and interpretable evaluation of multilingual consistency, free from confounding effects of retrieval quality or relevance ranking. The binary classification setup enables straightforward exact-match evaluation, providing a robust measure of cross-lingual reasoning alignment.

We also standardize on English prompts for experimental clarity and reproducibility. However, our approach can be readily adapted to use other high-resource languages as alignment anchors, making it extensible to different multilingual scenarios.

B Extended Results

Baseline Performance: Table 5 shows the performance of pre-trained LLMs (Baseline) without the application of DPO/ORPO or batch alignment method. The performance gap between English and non-English languages is significant, with an average gap of (11.71%) across all models. The largest gap is observed in the Llama-3.1-8B model, where English outperforms the average of the other languages by (28.7%). Mistral-7B models show gradual improvements across versions, with Mistral-7B-v0.3 performing better than it's previous versions, particularly in English and Portuguese. The Mixtral-8x22B model performs consistently well across languages, particularly excelling in English, Spanish, and French.

In contrast, the Llama-3 family shows varying performance, with the Llama-3-70B model outperforming the 8B version across all languages. The Llama-3.1-8B model particularly struggles with Arabic and French, while its 70B counterpart underperforms in Spanish and Portuguese, highlighting inconsistent performance of the Llama-3.1 models across different sizes and languages.

Impact of Batch Alignment on DPO & ORPO: Table 6 compares the performance of models under default DPO settings and our proposed batch alignment method (with Batch Size = 7). Under default DPO, the performance gap between English and non-English languages increases to (13.9%), as the models become more optimized for English. However, with our batch alignment method applied to DPO, this gap is significantly reduced to an average of (4.1%), with a maximum improvement of (23.9%) observed in the Llama-3.1-8B model, which had initially struggled the most. This demonstrates that our method is particularly effective in boosting the performance of weaker models in non-English languages. Mistral-7B-Instruct v0.3 continues to show the best results in both English and non-English tasks, particularly excelling in English, Portuguese, and French, while Mixtral-8x22B also demonstrates substantial gains across all languages. Even for models that initially underperformed, such as Llama-3.1-8B, our batch alignment method leads to marked improvements

Models	DPO Finetuning (Default)							
	English	Arabic	Spanish	French	Japanese	Portuguese	Chinese	Rest (avg.)
Mistral-7B-Instruct-v0.1	60.3	27.2	54.6	54.6	44.4	57.7	52.5	48.5
Mistral-7B-Instruct-v0.2	73.5	61.2	66.4	65.8	64.2	66.2	67.6	65.2
Mistral-7B-Instruct-v0.3	76.3	62.3	66.8	67.4	65.3	71.5	68.9	67.0
Mixtral-8x22B-Instruct-v0.1	79.3	63.4	72.9	72.6	67.9	72.1	69.6	69.8
Llama-3-8B-Instruct	65.8	61.9	58.3	55.4	60.8	55.9	55.8	58.0
Llama-3-70B-Instruct	73.6	64.8	64.4	69.0	67.2	61.8	68.8	66.0
Llama-3.1-8B-Instruct	68.2	32.3	43.4	21.5	52.3	34.3	52.3	39.4
Llama-3.1-70B-Instruct	78.8	48.6	29.4	54.5	68.9	32.3	69.5	50.5
Models	**DPO + Ours (Batch Size = 7)**							
	English	**Arabic**	**Spanish**	**French**	**Japanese**	**Portuguese**	**Chinese**	**Rest (avg.)**
Mistral-7B-Instruct-v0.1	60.3	45.6(+18.4)	58.6(+4.0)	57.6(+3.0)	55.6(+11.2)	57.7(+0.0)	54.9(+2.4)	55.0(+6.5)
Mistral-7B-Instruct-v0.2	73.5	68.6(+7.4)	70.2(+3.8)	69.6(+3.8)	69.5(+5.3)	72.2(+6.0)	71.6(+4.0)	70.3(+5.1)
Mistral-7B-Instruct-v0.3	76.3	71.6(+9.3)	72.6(+5.8)	73.5(+6.1)	71.2(+5.9)	74.5(+3.0)	72.9(+4.0)	72.7(+5.7)
Mixtral-8x22B-Instruct-v0.1	79.3	72.5(+9.1)	75.6(+2.7)	74.6(+2.0)	76.6(+8.7)	74.6(+2.5)	75.6(+6.0)	74.9(+5.1)
Llama-3-8B-Instruct	65.8	63.8(+1.9)	61.2(+2.9)	62.3(+6.9)	64.3(+3.5)	62.4(+6.5)	61.3(+5.5)	62.6(+4.6)
Llama-3-70B-Instruct	73.6	71.2(+6.4)	73.0(+8.6)	73.2(+4.2)	72.1(+4.9)	69.5(+7.7)	71.0(+2.2)	71.7(+5.7)
Llama-3.1-8B-Instruct	68.2	65.5(+33.2)	62.3(+18.9)	62.8(+41.3)	64.5(+12.2)	63.1(+28.8)	61.8(+9.5)	63.3(+23.9)
Llama-3.1-70B-Instruct	78.8	71.9(+23.3)	73.6(+44.2)	74.2(+19.7)	73.3(+4.4)	71.1(+38.8)	71.9(+2.4)	72.7(+22.2)

Table 6: Accuracy comparison of LLMs across English and six non-English languages under default DPO and DPO with batch alignment. Improvements in non-English languages are shown in parentheses. "Rest (avg.)" is the average accuracy across non-English languages, showing enhanced cross-lingual consistency with batch alignment.

Model	English (95% CI)	Thai (95% CI)	Vietnamese (95% CI)	Hungarian (95% CI)	Romanian (95% CI)	Rest (avg.)
Mixtral-8×22B (Base)	74.20 [72.2–76.1]	34.80 [32.7–36.9]	41.50 [39.4–43.7]	42.70 [40.5–44.9]	47.60 [45.4–49.8]	41.7
Mixtral-8×22B (Ours)	80.20 [78.4–81.9]	37.90 [35.8–40.0]	43.50 [41.3–45.7]	45.90 [43.7–48.1]	49.90 [47.7–52.1]	44.3
Llama-3.1-70B (Base)	73.60 [71.6–75.5]	30.70 [28.7–32.8]	35.00 [32.9–37.1]	32.90 [30.9–35.0]	36.00 [33.9–38.1]	33.7
Llama-3.1-70B (Ours)	79.20 [77.4–80.9]	37.50 [35.4–39.6]	39.10 [37.0–41.3]	36.20 [34.1–38.3]	38.50 [36.4–40.7]	37.8
Llama-3-70B (Base)	65.80 [63.7–67.8]	28.60 [26.7–30.6]	34.60 [32.5–36.7]	31.50 [29.5–33.6]	35.80 [33.7–37.9]	32.6
Llama-3-70B (Ours)	74.10 [72.1–76.0]	34.10 [32.1–36.2]	36.20 [34.1–38.3]	36.10 [34.0–38.2]	37.30 [35.2–39.4]	35.9

Table 7: Expanded results for unseen languages. Accuracy (%) on English and four unseen languages. Brackets show 95% bootstrap CIs; *Rest (avg.)* is the macro-average over the non-English languages.
Notes: Values are accuracy (%); brackets show 95% bootstrap confidence intervals. Significance vs. baselines is validated with McNemar's test.

in Arabic and Portuguese, highlighting the robustness of our method in addressing performance gaps across languages.

Table 8 presents similar comparisons under the ORPO setting. With default ORPO, the performance gap between English and non-English languages increases to (13.6%), which is slightly smaller than the default DPO setting. However, with the application of our batch alignment technique, this gap is reduced to an average of (3.6%), further emphasizing the effectiveness of alignment training. Again, the Llama-3.1-8B model sees the greatest improvement, particularly in Arabic and French, where initial performance had been weakest.

Comparing DPO and ORPO: Tables 6 and 8 show that ORPO generally outperforms DPO in multilingual tasks, with an average improvement of (1.1%) for non-English languages. However, our batch alignment method significantly reduces the performance gap in both settings. The gap between English and non-English languages decreases from (13.9%) to (4.1%) in DPO, and from (13.6%) to (3.6%) in ORPO, underscoring the importance of optimizing internal reasoning consistency across languages, especially for weaker models and low-resource languages.

Generalization on Unseen Langauges (CI + significance). Table 7 expands the results of RQ2 5.2 (Table 1) on Thai, Vietnamese, Hungarian, and Romanian language with **95% bootstrap confidence intervals** (B=1000) per model and language, complementing the summary in the main text. Gains are **consistent across all four languages** and **statistically significant** under McNemar's test (two-sided, $p < 0.01$) for accuracy/EM. Across models, *Rest (avg.)* improves by **+2.6 to +4.1** points. For example, Mixtral-8×22B im-

Models	ORPO Finetuning (Default)							
	English	**Arabic**	**Spanish**	**French**	**Japanese**	**Portuguese**	**Chinese**	**Rest (avg.)**
Mistral-7B-Instruct-v0.1	60.8	28.3	55.1	55.3	45.2	58.6	54.5	49.5
Mistral-7B-Instruct-v0.2	74.2	62.3	67.2	66.4	65.2	67.3	68.3	66.1
Mistral-7B-Instruct-v0.3	76.6	63.4	67.3	68.2	66.4	72.5	69.6	67.9
Mixtral-8x22B-Instruct-v0.1	80.2	64.2	73.5	74.5	68.3	73.2	70.5	70.7
Llama-3-8B-Instruct	66.2	62.5	59.4	56.3	61.4	56.2	56.8	58.8
Llama-3-70B-Instruct	74.1	65.8	65.9	70.2	68.0	62.3	68.7	66.8
Llama-3.1-8B-Instruct	68.9	32.5	44.1	23.1	54.3	36.6	53.5	40.7
Llama-3.1-70B-Instruct	79.2	49.2	30.1	55.6	69.4	33.1	70.2	51.3
Models	**ORPO + Ours (Batch Size = 7)**							
	English	**Arabic**	**Spanish**	**French**	**Japanese**	**Portuguese**	**Chinese**	**Rest (avg.)**
Mistral-7B-Instruct-v0.1	60.8	59.6(+31.3)	59.0(+3.9)	58.2(+2.9)	56.4(+11.2)	59.3(+0.7)	56.1(+1.6)	58.1(+8.6)
Mistral-7B-Instruct-v0.2	74.2	69.1(+6.8)	71.1(+3.9)	70.9(+4.5)	70.1(+4.9)	73.3(+6.0)	72.3(+4.0)	71.1(+5.0)
Mistral-7B-Instruct-v0.3	76.6	72.9(+9.5)	73.2(+5.9)	74.1(+5.9)	72.1(+5.7)	74.9(+2.4)	73.5(+3.9)	73.5(+5.6)
Mixtral-8x22B-Instruct-v0.1	80.2	74.6(+10.4)	77.1(+3.6)	76.2(+1.7)	75.6(+7.3)	76.1(+2.9)	76.2(+5.7)	76.0(+5.3)
Llama-3-8B-Instruct	66.2	64.2(+1.7)	65.6(+6.2)	61.2(+4.9)	62.3(+0.9)	63.5(+7.3)	61.2(+4.4)	63.0(+4.2)
Llama-3-70B-Instruct	74.1	68.2(+2.4)	73.4(+7.5)	71.5(+1.3)	70.5(+2.5)	72.5(+10.2)	70.9(+2.2)	71.2(+4.4)
Llama-3.1-8B-Instruct	68.9	65.1(+32.6)	66.5(+22.4)	62.3(+39.2)	63.4(+9.1)	64.5(+27.9)	62.9(+9.4)	64.1(+23.4)
Llama-3.1-70B-Instruct	79.2	70.2(+21.0)	78.2(+48.1)	77.6(+22.0)	74.9(+5.5)	74.2(+41.1)	73.5(+3.3)	74.8(+23.5)

Table 8: Accuracy comparison of LLMs across English and six non-English languages under default ORPO and ORPO with batch alignment. Improvements in non-English languages are shown in parentheses. "Rest (avg.)" is the average accuracy across non-English languages, showing enhanced cross-lingual consistency with batch alignment.

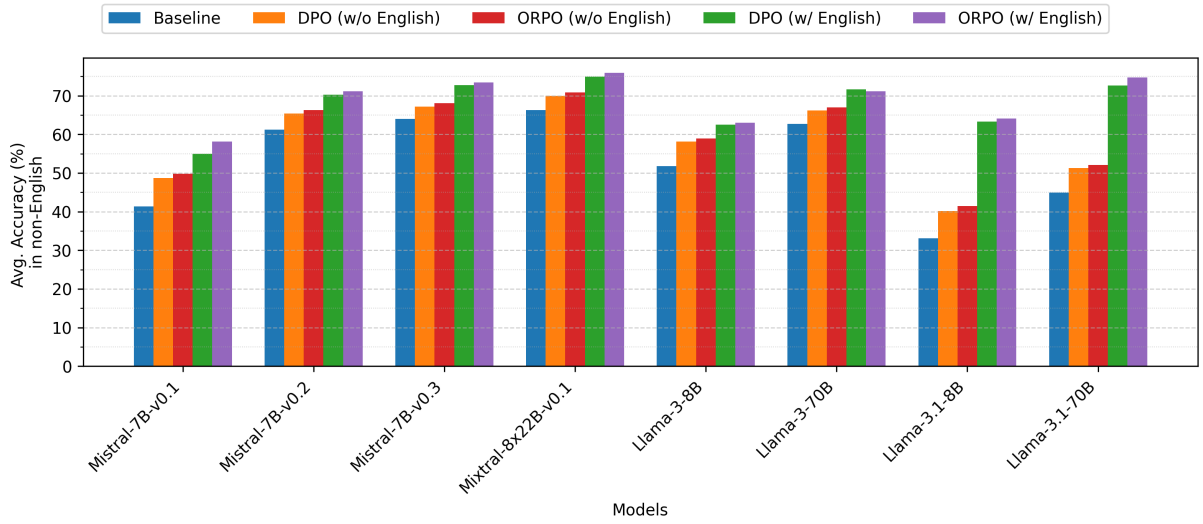

Figure 4: Impact of including vs. excluding English in training batches on non-English accuracy. Across all models, training with English (purple and green bars) consistently leads to higher non-English accuracy, demonstrating the role of high-resource languages in improving multilingual consistency. We use the *instruct* version of each model.

proves by **+3.1** (Thai), **+2.0** (Vietnamese), **+3.2** (Hungarian), **+2.3** (Romanian); Llama-3.1-70B improves by **+6.8**, **+4.1**, **+3.3**, **+2.5** on the same languages, while English remains stable. These results show that the multilingual consistency gains extend to **unseen** languages and are not confined to high-resource settings.

C Extended Ablation Studies

We conduct a series of ablation experiments to further dissect the components of our method. Specifically, we study the impact of batch size on alignment effectiveness and investigate the use of machine-translated data as an alternative to manual translations.

Models	Baseline		with English (DPO)		w/o English (DPO)	
	English	Rest (Avg.)	English	Rest (Avg.)	English	Rest (Avg.)
Mistral-7B-Instruct-v0.1	56.0	41.3	60.3	55.0	56.0	48.7
Mistral-7B-Instruct-v0.2	64.0	61.3	73.5	70.3	64.0	65.4
Mistral-7B-Instruct-v0.3	70.4	64.0	76.3	72.7	70.4	67.2
Mixtral-8x22B-Instruct-v0.1	74.2	66.3	79.3	74.9	74.2	69.9
Llama-3-8B-Instruct	56.0	51.8	65.8	62.6	56.0	58.2
Llama-3-70B-Instruct	65.8	62.8	73.6	71.7	65.8	66.2
Llama-3.1-8B-Instruct	59.2	33.1	68.2	63.3	59.2	40.2
Llama-3.1-70B-Instruct	73.6	44.9	78.8	72.7	73.6	51.3

Table 9: Comparison of LLM's performance with and without English in the training batches under our DPO settings. "Rest (avg.)" reflects the average accuracy across non-English languages, showing that including English significantly boosts cross-lingual performance.

Models	Baseline		with English (ORPO)		w/o English (ORPO)	
	English	Rest (Avg.)	English	Rest (Avg.)	English	Rest (Avg.)
Mistral-7B-Instruct-v0.1	56.0	41.3	60.8	58.1	56.0	49.8
Mistral-7B-Instruct-v0.2	64.0	61.3	74.2	71.1	64.0	66.3
Mistral-7B-Instruct-v0.3	70.4	64.0	76.6	73.5	70.4	68.1
Mixtral-8x22B-Instruct-v0.1	74.2	66.3	80.2	76.0	74.2	70.9
Llama-3-8B-Instruct	56.0	51.8	66.2	63.0	56.0	58.9
Llama-3-70B-Instruct	65.8	62.8	74.1	71.2	65.8	66.9
Llama-3.1-8B-Instruct	59.2	33.1	68.9	64.1	59.2	41.5
Llama-3.1-70B-Instruct	73.6	44.9	79.2	74.8	73.6	52.1

Table 10: Comparison of LLM's performance with and without English in the training batches under our ORPO settings. "Rest (avg.)" reflects the average accuracy across non-English languages, showing that including English significantly boosts cross-lingual performance.

C.1 Role of High-Resource Languages

Figure 4 investigates the critical role of English inclusion within training batches. Our results highlight that including English examples consistently enhances non-English accuracy by substantial margins: for example, Mixtral-8x22B improves from 66.3% (without English) to 76.0% (with English), and Llama-3.1-70B improves from 52.1% to 74.8%. This observation underscores the importance of leveraging high-resource languages like English as anchors to guide internal multilingual reasoning alignment.

As shown in Table 9, including English in the batch (BS = 7) significantly improved performance across most models in DPO settings. For example, the Mistral-7B-v0.1 model maintained an English score of (56.0%), but when English was excluded from the training, the average performance for other languages dropped by (6.3%), resulting in an average score of (48.7%). Similarly, the Llama-3-8B-Instruct model experienced a reduction of (4.4%), dropping to (58.2%) for non-English languages when English was not included. These results underscore the critical role of English in enhancing overall multilingual alignment and consistency.

In ORPO settings (Table 10), a similar trend was observed. Instruction-tuned models such as Mistral-7B and Llama-3.1-70B exhibited significant drops in performance when English was excluded, with reductions of (8.3%) and (22.7%), respectively. Notably, the Llama-3.1 family model showed a major decline of approximately (22%), further emphasizing the importance of high-resource languages like English for effective multilingual training. These findings highlight that including English in the training process plays a crucial role in stabilizing model performance across all languages, ensuring greater consistency and accuracy.

C.2 Effect of Machine Translated data

We expand our finding & evaluation beyond human curated datasets to machine translated data to ensure scalability. We translate the english queries and documents to Thai, Vietnamese, Hungarian & Romanian languages, and repeat the experiments with Llama-3-70B, Llama-3.1-70B & Mixtral-8x22B-v0.1 with ORPO only. Figure 5 shows the average improvement compared to the baseline & default (ORPO). We observe consistent

Models	Baseline (Pre-trained Performance)					
	English	Thai	Vietnamese	Hungarian	Romanian	Rest (avg.)
Mixtral-8x22B-v0.1	74.2	34.8	41.5	42.7	47.6	41.65
Llama-3-70B	65.8	28.6	34.6	31.5	35.8	32.63
Llama-3.1-70B	73.6	30.7	35.0	32.9	36.0	33.65
Models	**ORPO Finetuning (Default)**					
	English	**Thai**	**Vietnamese**	**Hungarian**	**Romanian**	**Rest (avg.)**
Mixtral-8x22B-Instruct-v0.1	80.2	39.5	45.8	48.2	52.6	46.53
Llama-3-70B-Instruct	74.1	37.8	39.7	36.1	39.1	38.18
Llama-3.1-70B-Instruct	79.2	39.6	43.5	39.9	42.6	41.40
Models	**ORPO + Ours (Batch Size = 7)**					
	English	**Thai**	**Vietnamese**	**Hungarian**	**Romanian**	**Rest (avg.)**
Mixtral-8x22B-Instruct-v0.1	80.2	43.5(+4.0)	49.9(+4.1)	57.8(+9.6)	58.6(+6.0)	52.45(+5.93)
Llama-3-70B-Instruct	74.1	41.5(+3.7)	45.4(+5.7)	41.5(+5.4)	46.5(+7.4)	43.73(+5.55)
Llama-3.1-70B-Instruct	79.2	44.5(+4.9)	49.2(+5.7)	44.5(+4.6)	51.5(+8.9)	47.43(+6.03)

Table 11: Accuracy comparison of LLMs across English and four non-English languages (machine-translated) under baseline, default ORPO and ORPO with batch alignment (ours). Improvements in non-English languages are shown in parentheses compared to default finetuning. "Rest (avg.)" is the average accuracy across non-English languages, showing enhanced cross-lingual consistency with batch alignment.

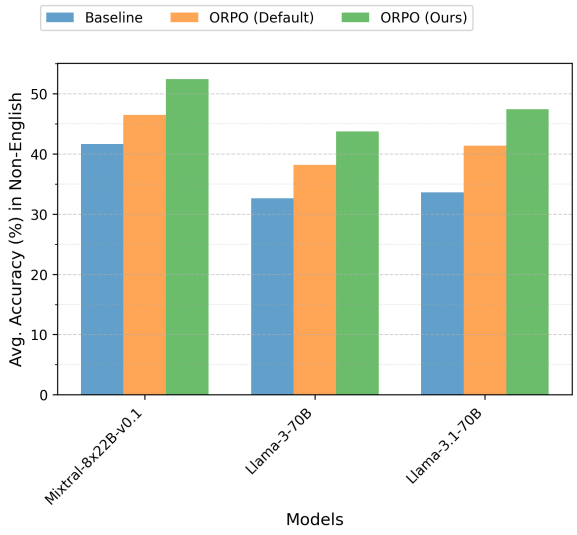

Figure 5: Illustrates avg. performance when machine translated data is used for batch-aligned finetuning of models across Thai, Vietnamese, Hungarian & Romanian langauge.

improvement across the three models, highlighting that machine translated data also helps to align the internal generation processess of LLM across languages, improving consistency and and making our proposed method more scalable. Though the improvement is slightly lower than the manual curated data, which could be due to the language or translation quality.

Table 11 presents the detailed results across three models—Mixtral-8x22B-v0.1, Llama-3-70B, and Llama-3.1-70B—comparing their baseline (pre-trained performance), ORPO fine-tuning, and our proposed method for each language separately.

Our findings indicate that machine-translated data contributes to consistent improvements across all models, validating its role in scalable multilingual adaptation. Hungarian and Romanian benefit the most, with Mixtral-8x22B-v0.1 achieving a +9.6% and +6.0% absolute gain compared to default finetuning, while Llama-3.1-70B improves by +4.6% and +8.9%, respectively. This demonstrates that machine translation effectively enhances performance for morphologically complex, lower-resource languages. In contrast, Thai and Vietnamese see relatively moderate but stable gains ($\leq$6%), potentially due to better inherent alignment with English.

Crucially, the improvements are observed across all models, demonstrating that our proposed strategy generalizes across different architectures. The results highlight that integrating machine-translated data in pretraining and fine-tuning pipelines is a viable alternative to human-curated datasets, offering a practical solution for improving model performance across languages where high-quality parallel data is scarce. Our findings demonstrate that this approach can be directly integrated into training large-scale multilingual LLMs, ensuring cross-lingual consistency without manual intervention.

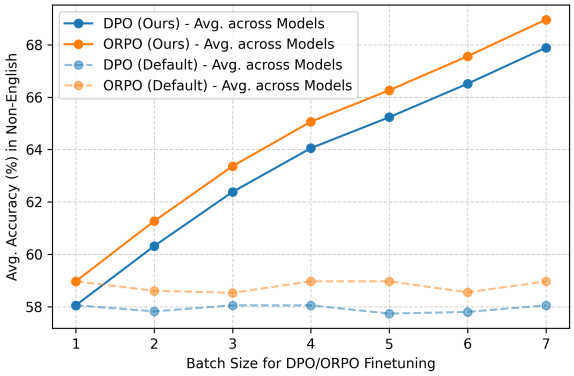

Figure 6: Illustrates how increasing batch size improves average accuracy across six non-English languages. Models with batch alignment show consistent gains, while default methods see minimal change.

C.3 Effect of Batch Size

Figure 6 shows the impact of increasing batch size on average accuracy across non-English languages, with results averaged across all the models. The results highlight how our proposed batch alignment methods improve performance as batch size increases in both DPO and ORPO alignment training, whereas the default techniques show little to no effect.

Larger batch sizes help models capture diverse languages representing the same concept enabling more cross-lingual training signals per update, improving their ability to generalize across multilingual contexts. In contrast, default DPO and ORPO methods, which randomly select samples, show no improvement as batch size increases, plateauing at 58-60% accuracy compared to ours which reaches upto 69%. This is due to the lack of structured language diversity in the default settings. Our findings highlight that increasing batch diversity is crucial in cross-lingual generalization for reducing performance gaps between languages, underscoring batch diversity's role in cross-lingual generalization.

We started with English as the default (batch_size=1), additional languages were progressively included in each batch (e.g., batch size 2 includes English and one random language, batch size 3 includes English and two random languages, and so on). Tables 12 and 13 summarize the impact of batch size under both DPO and ORPO settings for each model individually.

In DPO settings (Table 12), the Llama-3.1-8B model exhibited a notable performance increase of $(+5.5\%)$ with the inclusion of one additional language, while Llama-3-8B showed a modest gain of $(+1.1\%)$. As the batch size increased, performance improved across all models, with the Llama-3.1-8B model achieving a maximum boost of $(+23.9\%)$ at a batch size of 7. Similarly, in ORPO settings (Table 13), the Llama-3.1-70B model saw substantial gains of $(+23.5\%)$ at a batch size of 7, while Mixtral-8x22B-Instruct showed a rise of $(+5.3\%)$. These results indicate that larger, more diverse batches help models capture and generalize over languages with similar semantic content, particularly boosting consistent performance in non-English languages.

C.4 Models Size v/s Multlingual Alignment

Figure 7 shows the relationship between model size and language improvement for both DPO and ORPO under our alignment strategy. While both DPO & ORPO shows a positive corelation of 0.23 and 0.21 respectively, it is not particularly strong, demonstrating that larger models do tend to improve language performance, but the relationship isn't very pronounced. This could indicate that simply increasing model size may not completely close the language gap, and other factors (such as specific fine-tuning strategies) may play a role.

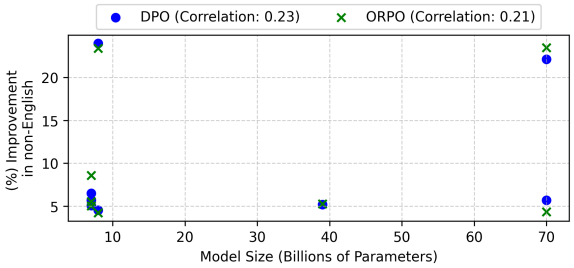

Figure 7: Shows the relationship between model size and average performance improvement in non-English languages under DPO and ORPO settings. The modest correlation observed demonstrates that factors beyond model size contribute to performance gains.

Models	DPO Finetuning							
	English	**Rest (Avg.)**	**BS = 2**	**BS = 3**	**BS = 4**	**BS = 5**	**BS = 6**	**BS = 7**
Mistral-7B-Instruct-v0.1	60.3	48.5	50.0	51.4	52.5	53.2	54.1	55.0
Mistral-7B-Instruct-v0.2	73.5	65.2	66.4	67.5	68.3	68.9	69.6	70.3
Mistral-7B-Instruct-v0.3	76.3	67.0	68.3	69.5	70.5	71.2	71.9	72.7
Mixtral-8x22B-Instruct-v0.1	79.3	69.8	70.9	72.0	72.9	73.5	74.2	74.9
Llama-3-8B-Instruct	65.8	58.0	59.1	60.0	60.8	61.3	61.9	62.6
Llama-3-70B-Instruct	73.6	66.0	67.3	68.5	69.5	70.1	70.9	71.7
Llama-3.1-8B-Instruct	68.2	39.4	44.9	49.9	54.0	56.9	60.0	63.3
Llama-3.1-70B-Instruct	78.8	50.5	55.6	60.3	64.0	66.7	69.6	72.7

Table 12: Average accuracy of LLMs in English and non-English languages as batch size increases under our DPO settings. "Rest (avg.)" indicates the average accuracy across non-English languages, showing improved cross-lingual performance with larger batch sizes. Here, BS = batch size.

Models	ORPO Finetuning							
	English	**Rest (Avg.)**	**BS = 2**	**BS = 3**	**BS = 4**	**BS = 5**	**BS = 6**	**BS = 7**
Mistral-7B-Instruct-v0.1	60.8	49.5	51.5	53.3	54.7	55.8	56.9	58.1
Mistral-7B-Instruct-v0.2	74.2	66.1	67.3	68.3	69.2	69.8	70.4	71.1
Mistral-7B-Instruct-v0.3	76.6	67.9	69.2	70.3	71.3	72.0	72.7	73.5
Mixtral-8x22B-Instruct-v0.1	80.2	70.7	71.9	73.0	73.9	74.5	75.2	76.0
Llama-3-8B-Instruct	66.2	59.7	58.8	60.6	61.3	61.9	62.4	63.0
Llama-3-70B-Instruct	74.1	66.8	67.8	68.7	69.5	70.0	70.6	71.2
Llama-3.1-8B-Instruct	68.9	40.7	46.1	51.0	55.0	57.8	60.8	64.1
Llama-3.1-70B-Instruct	79.2	51.3	56.7	61.6	65.6	68.4	71.5	74.8

Table 13: Average accuracy of LLMs in English and non-English languages as batch size increases under our ORPO settings. "Rest (avg.)" indicates the average accuracy across non-English languages, showing improved cross-lingual performance with larger batch sizes. Here, BS = batch size.