

Transporting Concepts Across Languages: Optimal Transport Toward English-Level Performance in Multilingual LLMs

Anh Trac Duc Dinh^{1,2} Anh Duc Le^{3*} Tho T. Quan²

¹Center for AI Research (CAIR), VinUniversity, Hanoi, Vietnam

²Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, Vietnam

³NTT Hi-Tech Institute, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam

*Corresponding author: anhld@ntt.edu.vn

Abstract

Large language models (LLMs) exhibit strong English performance due to its dominance in pretraining data, yet remain inconsistent across non-English languages. Existing cross-lingual alignment methods rely on contrastive objectives that symmetrically perturb English representations, inadvertently degrading the proficiency they use as a reference. We propose an Optimal Transport (OT) framework that maps non-English hidden-state distributions toward English at middle transformer layers via entropy-regularised Sinkhorn iterations, while an English consistency regularisation term keeps the alignment anchor stationary. This asymmetric formulation preserves English performance by construction. Experiments on XNLI, XQuAD, and MMMLU show that our method delivers strong cross-lingual transfer while successfully avoiding the English performance degradation observed in the contrastive baseline.

1 Introduction

LLMs achieve strong performance across NLP tasks, yet progress is unevenly distributed across languages. English dominates pretraining corpora: 92.65% of GPT-3’s tokens and 89.70% of LLaMA-2’s (Li et al., 2025), causing substantially weaker performance on non-English languages, with the gap widening further for low-resource languages such as Swahili, Urdu, or Yoruba (Ahuja et al., 2023). Cross-lingual alignment is therefore essential to bridge this gap, enabling knowledge acquired in English to transfer effectively to underrepresented languages without retraining from scratch.

One line of work addresses this by cross-lingual alignment: adapting a model so that semantically equivalent content across languages maps to nearby representations, enabling knowledge acquired in English to transfer to other languages. This approach has been extensively studied for encoder-only transformers such as XLM-R (Conneau et al.,

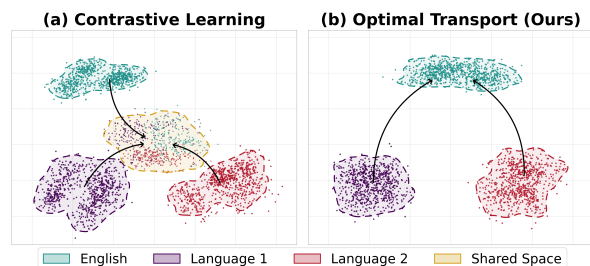


Figure 1: Conceptual Illustration of Contrastive Learning approach (a) and our Optimal Transport approach (b). Contrastive objectives pull all language representations into a shared space, perturbing the English anchor. Our asymmetric OT formulation transports non-English distributions toward a fixed English reference, preserving English proficiency by construction.

2020) and its variants (Hu et al., 2021), but the design choices for encoders do not transfer directly to autoregressive Large Language Models, whose causal attention and generative training objective pose different constraints.

Alignment methods designed for LLMs are comparatively scarce. Existing approaches rely on contrastive representation learning: Liu and Niehues (2025) apply a contrastive objective at a fixed middle layer during task-specific fine-tuning, while Bu et al. (2025) use a language-matching classifier in a two-stage training framework. Both improve non-English performance, but because contrastive objectives pull translation-pair representations together symmetrically, they inevitably perturb the English representations used as the alignment reference, reducing English performance as a side effect.

We approach cross-lingual alignment through Optimal Transport (Figure 1): rather than pulling paired representations together symmetrically, we frame alignment as a distributional transport problem that maps non-English hidden states toward English at middle transformer layers via entropy-regularised Sinkhorn iterations (Cuturi, 2013),

keeping the English distribution fixed by construction. An English consistency regularisation term stabilises the alignment anchor, and a language modelling loss preserves generative capacity.

To the best of our knowledge, this is the first work to formulate cross-lingual alignment specifically for autoregressive LLMs as an Optimal Transport problem, combining middle-layer targeting with learnable layer aggregation. Our main contributions are:

- A novel alignment framework that treats cross-lingual hidden-state transfer as a distributional transport problem (Section 3).
- A three-component training objective that jointly facilitates cross-lingual transfer while preserving English proficiency (Section 3).
- Consistent empirical gains over contrastive baseline on XNLI (classification), XQuAD (extractive QA), and competitive performance on MMMLU (knowledge recall) (Section 4.3).

2 Literature Review

2.1 English-Centric Bias in LLMs and the Case for Asymmetric Alignment

Modern LLMs are trained on corpora where English dominates by a large margin, and this imbalance leaves a structural imprint on how these models represent language internally. [Wendler et al. \(2024\)](#) and [Etxaniz et al. \(2024\)](#) show that when a multilingual LLM processes a non-English input, it does not reason in the input language; it first converts the input into English-like internal representations at intermediate layers, then generates from that English-centric space. In other words, English is not just the language the model performs best on; it is the internal language the model thinks in. [Wang et al. \(2024\)](#) further show that this English-pivot behaviour is a learned property that strengthens with pretraining scale, making it a deep-seated geometric feature of the model rather than a surface artifact.

This has a direct implication for alignment design. Since English representations serve as the model’s internal pivot for cross-lingual processing, methods that perturb them risk undermining the very foundation they align toward. Symmetric contrastive objectives, which pull both sides simultaneously, are structurally prone to this problem. An asymmetric formulation; transporting non-English representations toward a fixed English anchor is therefore not just preferable but necessary. Opti-

mal Transport provides exactly this property, and it motivates the core design of our framework.

2.2 Cross-lingual Alignment for Encoder-based Models

Explicit cross-lingual alignment has been most extensively studied for encoder-only architectures. Early frameworks fine-tune mBERT or XLM-R with explicit objectives on parallel data, ranging from word-level alignment ([Cao et al., 2020](#)) and multi-granularity objectives covering both word and sentence levels ([Hu et al., 2021](#)), to contrastive pretraining grounded in mutual information maximisation ([Chi et al., 2021](#)) and translation ranking over 6 billion sentence pairs ([Feng et al., 2022](#)). These works collectively establish that encoder representations can be meaningfully aligned across languages through explicit objectives on parallel data.

Optimal Transport has also appeared in this encoder setting. [Alqahtani et al. \(2021\)](#) fine-tune mBERT with Sinkhorn divergence as the alignment objective, learning soft token-level correspondences directly from parallel sentences without pre-collected word pairs. Importantly, they also introduce an English consistency loss that penalises deviation of the fine-tuned English representations from the original frozen ones; an early recognition that OT-based alignment risks drifting the very anchor it relies on. Our framework extends and formalises this intuition for autoregressive LLMs, where the problem is more acute: shared LoRA parameters mean non-English gradients can indirectly distort the English branch even when the transport direction is asymmetric by design. OT has also been applied to cross-lingual semantic parsing ([Sherborne et al., 2023](#)) and cross-lingual summarisation via Sinkhorn-based knowledge distillation ([Nguyen and Luu, 2022](#)). However, none of these works address the generative capacity and causal attention constraints of autoregressive LLMs.

2.3 Cross-lingual Alignment for Autoregressive LLMs: Limitations of Contrastive Approaches

Alignment methods for autoregressive LLMs remain scarce. AFP ([Li et al., 2024](#)) adds a multilingual contrastive objective on top of cross-lingual instruction tuning. [Liu and Niehues \(2025\)](#), which we refer to as MIDALIGN, alternates between task specific training and a layer-wise contrastive loss

applied at the middle layer. AlignX (Bu et al., 2025) performs contrastive learning at the intermediate layer alongside a language matching classifier at the output layer, followed by multilingual instruction fine-tuning to consolidate the alignment. These methods improve non-English performance but rely on symmetric contrastive objectives, which pull translation-pair representations together without an explicit mechanism to protect the English anchor. This risk perturbs the dominant-language representations that serve as the alignment reference. No existing work in this space formulates cross-lingual alignment for autoregressive LLMs as a distributional transport problem, leaving the asymmetric, anchor-preserving formulation unexplored.

2.4 Layer Specialisation in Multilingual Transformers

Not all transformer layers contribute equally to cross-lingual transfer. Early layers capture surface-level features that vary sharply across languages; middle layers encode semantic content in a largely language-agnostic way; and deep layers specialise for task-specific patterns tied to English surface form (Wendler et al., 2024; Wang et al., 2024; Tenney et al., 2019). MIDALIGN confirms this in the LLM setting, showing that alignment at the middle layer of Llama yields higher cross-lingual retrieval accuracy than any other depth (Liu and Niehues, 2025). Middle layers are therefore the natural target for distributional transport.

3 Methodology

3.1 Problem Setup

Given a parallel corpus of English–other language sentence pairs (x^{en}, x^{other}) , we fine-tune a pre-trained LLM using LoRA adapters (Hu et al., 2022). Each training step performs three forward passes through the same model: (1) English with LoRA *off* (frozen), producing $h_{en,frz}^{(l)}$ with no gradients; (2) English with LoRA *on*, producing $h_{en,lor}^{(l)}$; and (3) the other language with LoRA *on*, producing $h_{other}^{(l)}$ and output logits. The frozen English branch acts as a fixed distributional anchor: the model is pulled toward it without modifying it, thereby preserving English performance throughout training. The overall training objective combines three losses:

$$\mathcal{L} = \mathcal{L}_{LM} + \lambda_{OT} \mathcal{L}_{OT} + \lambda_{Reg} \mathcal{L}_{Reg} \quad (1)$$

where $\mathcal{L}_{LM}$ is the standard causal language modelling loss on the other language branch, $\mathcal{L}_{OT}$ is the cross-lingual optimal transport alignment loss, and $\mathcal{L}_{Reg}$ is an English consistency regularisation term.

3.2 Cross-lingual Alignment via Optimal Transport

Layer selection. We apply alignment exclusively to the *middle* layers $\mathcal{L}_{mid} = [\lfloor L/4 \rfloor, \lfloor 3L/4 \rfloor)$ of an L -layer transformer. Early layers encode surface-level token features while final layers are task-specific; intermediate layers carry semantic abstractions most amenable to cross-lingual transfer (Tenney et al., 2019). For Llama-3-8B-Instruct ($L=32$) this yields layers 8–24.

Token representations. Raw hidden states treat all positions equally, conflating content tokens with padding. Let $m^{other} \in \{0, 1\}^T$ and $m^{en} \in \{0, 1\}^S$ denote the attention masks for the other-language and English branches respectively, where T and S are the respective sequence lengths. For both branches, non-padding tokens are ℓ_2 -normalised independently:

$$\tilde{h}_{other,i}^{(l)} = \frac{m_i^{other} \cdot h_{other,i}^{(l)}}{\|h_{other,i}^{(l)}\|_2}, \quad (2)$$

and similarly for the frozen English branch:

$$\tilde{h}_{en,j}^{(l)} = \frac{m_j^{en} \cdot h_{en,j}^{(l)}}{\|h_{en,j}^{(l)}\|_2}. \quad (3)$$

Padding positions are excluded by setting their cost to 2 (the maximum cosine distance) in the cost matrix. Both branches use uniform marginals $a_i = m_i^{other} / \sum_k m_k^{other}$ and $b_j = m_j^{en} / \sum_k m_k^{en}$ in the Sinkhorn solver.

Optimal transport alignment. With normalised representations in hand, we measure the distributional discrepancy between other language and English hidden states at layer l via a cosine cost matrix (see Optimal Transport background in Appendix A.1):

$$C_{ij}^{(l)} = 1 - \langle \tilde{h}_{other,i}^{(l)}, \tilde{h}_{en,j}^{(l)} \rangle, \quad C_{ij}^{(l)} \in [0, 2]. \quad (4)$$

Padding positions are excluded by setting their cost to 2. Rather than solving the exact Kantorovich linear programme, whose cubic complexity $\mathcal{O}((T+S)^3)$ is prohibitive for token-level sequences computed across multiple layers per step.

We adopt the entropy-regularised relaxation of [Curi \(2013\)](#):

$$T^{(l)\star} = \arg \min_{T \in \Pi(\mathbf{a}, \mathbf{b})} \langle C^{(l)}, T \rangle_F - \varepsilon H(T), \quad (5)$$

solved via log-domain Sinkhorn–Knopp iterations (see Appendix A.2) at cost $\mathcal{O}(TS/\varepsilon^2)$ per layer. The layer-wise OT loss is then:

$$\mathcal{L}_{\text{OT}}^{(l)} = \sum_{i,j} T_{ij}^{(l)\star} \cdot C_{ij}^{(l)}. \quad (6)$$

Layer aggregation. Rather than averaging uniformly across layers, we introduce a set of learnable scalar weights $\{w_l\}_{l \in \mathcal{L}_{\text{mid}}}$ and aggregate via a softmax:

$$\mathcal{L}_{\text{OT}} = \sum_{l \in \mathcal{L}_{\text{mid}}} \underbrace{\frac{e^{w_l}}{\sum_k e^{w_k}}}_{\alpha_l} \mathcal{L}_{\text{OT}}^{(l)}. \quad (7)$$

This allows the model to down-weight layers that are already well-aligned and focus capacity on those that benefit most from transport.

3.3 English Consistency Regularisation

To prevent the LoRA update from distorting the English representations that serve as the alignment target, we penalise the per-token ℓ_2 deviation between the trainable and frozen English branches at each middle layer:

$$\mathcal{L}_{\text{Reg}} = \sum_{l \in \mathcal{L}_{\text{mid}}} \alpha_l \cdot \frac{1}{B} \sum_{i=1}^B \frac{1}{|M_i|} \sum_{t \in M_i} d_{i,t}^{(l)},$$

$$d_{i,t}^{(l)} = \|h_{\text{en,lor}a}^{(l)}(i, t) - h_{\text{en,frz}}^{(l)}(i, t)\|_2^2, \quad (8)$$

where M_i is the set of non-padding token positions in sample i and α_l are the same learned layer weights used in $\mathcal{L}_{\text{OT}}$. Intuitively, $\mathcal{L}_{\text{OT}}$ pulls the other language distribution toward English, while $\mathcal{L}_{\text{Reg}}$ keeps the English anchor stationary; together enforcing asymmetric, one-directional transfer.

3.4 Two-stage Alignment Finetuning Setup

We adopt a two-stage training strategy that decouples task acquisition from cross-lingual transfer, motivated by the observation that a model must first possess strong task-solving ability before that ability can be transferred across languages.

Stage 1 – English Task Fine-tuning. Since OT alignment anchors other-language representations to a *frozen* English branch, the quality of that anchor is critical. We fine-tune the top N layers on English task data, concentrating task-specific knowledge in the upper layers ([Tenney et al., 2019](#)) while preserving the middle-layer semantics that $\mathcal{L}_{\text{OT}}$ relies on in Stage 2.

Stage 2 – Cross-lingual OT Alignment. With Stage 1 weights frozen, we train LoRA adapters on bilingual parallel corpora using the combined objective: $\mathcal{L}_{\text{OT}}$ pulls other-language middle-layer representations toward the fixed English anchor, while $\mathcal{L}_{\text{Reg}}$ keeps that anchor stationary, enforcing asymmetric one-directional transfer.

4 Experiments

4.1 Datasets

English Multi-Task Fine-tuning. We fine-tune the base LLM using Causal Language Modelling on three English benchmarks: SQuAD 2.0 ([Rajpurkar et al., 2018](#)) (QA), SNLI ([Bowman et al., 2015](#)) (NLI), and MMLU ([Hendrycks et al., 2021](#)) (MCQ). Training splits are balanced by truncating larger datasets to match the smallest one (seed 42).

Cross-lingual Alignment. We align multilingual representations to English using two parallel corpora: FLORES-200 ([Team et al., 2022](#)) (200 languages, 100% of data) and OPUS-100 ([Tiedemann, 2012](#)) (100 languages, 10% per language). This intentionally *low-resource* setup ($\approx 26\text{K}$ samples/language on average) stress-tests alignment under data-scarce conditions. Table 1 summarises the corpus statistics.

Table 1: Alignment corpus statistics.

Statistics	FLORES-200	OPUS-100 (10%)	Total
Languages	200	100	216
Total samples	215,556	5,329,085	5,564,641
Mean / lang	1,012	60,102	25,762
Median	1,012	97,911	1,012
Min	1,012	63	1,012
Max	1,012	126,722	127,734

Downstream Evaluation. Cross-lingual transfer is assessed on multilingual counterparts of the English finetuning 1 tasks: XQuAD ([Artetxe et al., 2019](#)) (QA, 12 languages), XNLI ([Conneau et al., 2018](#)) (NLI, 15 languages), and MMMLU ([Hendrycks et al., 2021](#)) (MCQ, multiple languages).

Table 2: Results on Downstream benchmarks (XNLI, XQuAD, MMMLU). Vanilla and MIDALIGN are finetuned directly from Llama3-8B-Instruct. Finetuning serves as the Stage 1 checkpoint for Optimal Transport (OT (Ours)), which applies cross-lingual alignment in Stage 2 following our two-stage training setup. Bold indicates best, underline indicates second best; yellow shaded cells denote English and blue shaded cells denote overall performance. MMMLU provides only 14 non-English locale configurations by design and includes no English split, since the benchmark measures transfer out of English rather than English proficiency itself.

	XNLI (Acc)															
Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Overall
Vanilla	55.03	57.21	61.60	57.68	67.94	60.50	61.14	53.37	58.54	45.83	53.65	54.81	50.42	57.05	58.54	56.89
Finetuning	65.39	68.68	72.18	67.29	83.25	73.51	73.15	65.63	71.42	55.01	65.15	65.19	60.10	69.78	70.56	68.42
MIDALIGN	68.78	71.44	71.98	71.42	78.10	73.17	72.48	66.23	71.24	58.54	67.39	66.79	60.62	69.46	69.52	69.14
OT (Ours)	67.25	69.88	72.34	67.90	83.41	74.25	73.41	66.63	72.30	56.61	66.11	65.41	62.38	69.56	70.84	69.36
	XQuAD (F1/EM)															
Model	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh				Overall
Vanilla	66.1/41.9	69.1/41.8	65.0/32.3	78.6/58.8	72.8/43.5	54.2/23.6	68.4/43.5	66.1/41.8	67.0/48.9	60.3/29.6	71.6/45.0	43.1/35.8		–		65.2/40.5
Finetuning	44.8/31.2	69.6/53.6	59.1/40.4	83.6/72.5	73.6/55.5	43.1/29.7	74.5/58.7	52.2/35.8	51.4/42.0	64.0/44.2	70.9/50.9	60.0/53.2		–		62.2/47.3
MIDALIGN	11.1/8.5	46.2/37.1	20.0/16.2	73.6/62.2	49.1/36.6	13.5/10.8	40.1/32.1	15.7/10.3	15.3/12.4	31.3/23.0	34.6/31.3	15.3/12.9		–		30.5/23.9
OT (Ours)	61.9/43.9	70.1/54.1	66.9/46.9	83.4/72.1	75.3/56.9	59.1/42.3	75.5/59.8	59.2/40.9	61.6/53.2	62.9/40.8	72.1/51.0	62.9/58.0		–		67.6/51.5
	MMMLU (Acc)															
Model	AR	BN	DE	ES	FR	HI	ID	IT	JA	KO	PT	SW	YO	ZH	Def	Overall
Vanilla	43.18	25.61	51.35	54.58	53.93	37.20	48.62	52.41	46.10	45.23	53.73	35.19	28.69	50.50	44.74	44.74
Finetuning	43.12	25.39	51.98	55.01	54.44	37.12	49.38	52.59	46.42	45.81	53.87	35.37	28.89	50.95	45.04	45.03
MIDALIGN	46.23	26.27	52.76	55.20	54.60	38.50	51.15	53.87	48.50	48.15	54.99	37.19	27.74	52.27	46.24	46.24
OT (Ours)	44.26	25.59	52.82	55.74	54.56	37.94	50.38	53.66	47.12	46.68	54.99	36.75	28.52	51.62	45.76	45.76

4.2 Baselines

All models use Llama-3-8B-Instruct as the backbone; as a standard decoder-only transformer with architecture-agnostic OT objectives operating at the hidden-state level, findings generalise to other decoder-only LLMs. **Vanilla** is the unmodified lower-bound. **Finetuning** trains the top N layers ($N = 8$) on English multi-task data (SQuAD 2.0, SNLI, MMLU), serving as our Stage 1 backbone. **MIDALIGN** (Liu and Niehues, 2025), our primary baseline, alternates task training with Layer 16 contrastive alignment. Our **Optimal Transport (OT)** comprises Stage 1 (Finetuning checkpoint) and Stage 2: LoRA-based OT alignment across middle layers 8–24 ($r = 16$, $\alpha = 32$, dropout=0.05) on bilingual corpora. It optimizes $\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_{\text{OT}}\mathcal{L}_{\text{OT}} + \lambda_{\text{Reg}}\mathcal{L}_{\text{Reg}}$ ($\lambda_{\text{OT}} = 10$, $\lambda_{\text{Reg}} = 50$) using Sinkhorn ($\epsilon = 0.1$, 100 iterations). Lastly, **AlignX** (Bu et al., 2025) employs continued pre-training alignment; however, it is omitted from direct comparison because its zero-shot setting lacks English task fine-tuning, preventing a fair evaluation across benchmarks. All finetuned with 3 epochs, $lr = 2e^{-5}$, warmup step = 0.1, linear scheduling.

4.3 Main Results

Table 2 summarises performance across three benchmarks.

Optimal Transport achieves strong and consistent cross-lingual transfer. Our method achieves the best average on XNLI (69.36) and

XQuAD (67.6/51.5), outperforming all baselines. On MMMLU, it ranks second (45.76), slightly behind MIDALIGN (46.24) by 0.48 points. We attribute this gap to MIDALIGN’s symmetric contrastive alignment, which can inadvertently pull upper-layer representations closer together, benefiting MCQ settings where knowledge retrieval relies on matching between question and option tokens, consistent with (Liu and Niehues, 2025) on the role of intermediate alignment. In contrast, our asymmetric OT objective restricts alignment to middle layers while largely avoiding direct perturbation of upper layers, maintaining strong English and QA performance at the cost of a minor MCQ trade-off. Importantly, others either underperform on specific tasks, show regression on English (MIDALIGN), or fail to improve cross-lingual transfer beyond the task-finetuned baseline.

MIDALIGN collapses on extractive QA and degrades English. Despite competitive NLI and MCQ scores, MIDALIGN degrades on XQuAD (30.5/23.9), falling well below Vanilla (65.2/40.5)—a drop of over 37 F1 points that eliminates it as a practical solution. Two limitations may compound this failure: (1) alternate training isolates alignment updates from $\mathcal{L}_{\text{LM}}$, so the model never learns to align representations while preserving the span generative supervision required for token-level span extraction; classification-style tasks remain relatively unaffected because they do not depend on jointly optimized generation, whereas QA collapses; and (2) symmetric contrastive alignment pulls English

representations toward other languages, directly degrading English proficiency (78.10 vs. 83.25 on XNLI; 73.6/62.2 vs. 83.6/72.5 on XQuAD). This trade-off limits the applicability of MIDALIGN as a general-purpose alignment method. Our asymmetric OT objective and $\mathcal{L}_{LM}$ together prevent both failure modes.

English performance is preserved throughout alignment. On the English test portions of XNLI and XQuAD, OT matches or exceeds the Finetuning checkpoint (83.41 vs. 83.25; 83.4/72.1 vs. 83.6/72.5), confirming that $\mathcal{L}_{Reg}$ successfully stabilises the English anchor during Stage 2. This stands in sharp contrast to MIDALIGN, where cross-lingual gains are purchased at a measurable cost to English. Among the methods evaluated, OT is the method that consistently improves cross-lingual transfer without regressing on English.

Robustness under low-resource conditions. Operating at $\approx 26K$ samples per language on average, Our OT method consistently improves over the Finetuning baseline across low-resource languages (sw, ur, yo), further reinforcing that the gains are not limited to well-resourced languages. Taken together, the results across all three benchmarks establish OT as the only method that achieves strong performance on all task types without sacrificing English proficiency, a property that neither MIDALIGN nor the Finetuning baseline can claim.

4.4 Generalization Beyond Llama Backbone

Backbone Generalization Results. To test whether our framework generalises beyond Llama-3-8B-Instruct, we apply the identical two-stage pipeline, with hyperparameters unchanged ($\lambda_{OT} = 10$, $\lambda_{Reg} = 50$, $\varepsilon = 0.1$), to Qwen3.5-0.8B*¹ (non-thinking mode), a smaller model from a different family with a strong multilingual pretraining prior. Unlike Llama-3-8B-Instruct’s uniform transformer stack, Qwen3.5-0.8B*’s $L = 24$ layers follow a hybrid pattern: six blocks of three Gated DeltaNet layers followed by one full-attention layer (full-attention interval of 4). For Stage 1, we fine-tune the top $N = 4$ layers, matching the model’s native block boundary. For Stage 2, we apply the same rule $L_{mid} = \lfloor L/4 \rfloor, \lfloor 3L/4 \rfloor$, yield-

¹Model Card: <https://huggingface.co/Qwen/Qwen3.5-0.8B>. Unlike Qwen2.5/Qwen3, where the unsuffixed checkpoint denotes the base model, Qwen3.5’s unsuffixed checkpoint is the instruction-tuned model, released alongside a separate Qwen3.5-0.8B-Base pretrain-only checkpoint.

ing layers [6, 18]; within this range only layers 8, 12, and 16 are full-attention, the rest being Gated DeltaNet. We restrict $\mathcal{L}_{OT}$ and $\mathcal{L}_{Reg}$ to these three full-attention layers, as they are architecturally analogous to Llama-3-8B-Instruct’s self-attention blocks; the Gated DeltaNet layers carry a compressed recurrent state rather than token-level hidden states, and we leave OT alignment over linear-attention states to future work. Table 3 shows consistent gains despite no backbone-specific tuning: XNLI improves from 52.71 to 55.78, XQuAD from 53.93/43.46 to 54.78/44.56 F1/EM, and MMMLU from 35.47 to 36.91, while English stays stable (XNLI en: 50.15→51.32). This confirms asymmetric OT alignment transfers across backbone families and model scales, with smaller margins than Llama-3-8B-Instruct likely reflecting Qwen3.5-0.8B*’s already stronger multilingual prior and correspondingly less headroom for improvement.

When English Is Not the Strongest Anchor: A Cultural-Localization Case Study on Chinese. Chinese is the sole language that regresses under alignment (XNLI zh: 49.12→42.37; XQuAD zh F1: 60.35→58.62; MMMLU ZH: 47.22→43.66), even though English itself stays essentially flat. Unlike Llama-3-8B-Instruct, Qwen3.5-0.8B* is already highly competitive in Chinese—on MMMLU its Chinese accuracy (47.22)—so Chinese does not sit in a weak, under-specified region of representation space the way it does for our main backbone. When a non-English language already occupies a semantically rich region, transporting it toward the English anchor can disrupt its existing structure rather than supply missing information, yielding a net regression. This suggests the fixed English-anchor assumption is best suited to languages genuinely under-represented relative to English, and motivates future work on adaptive or language-specific anchors when a model’s dominant-language prior is not uniformly English.

5 Ablation Study

The three losses operate at different scales: $\mathcal{L}_{LM} \in [3, 8]$, $\mathcal{L}_{OT} \in [0.2, 0.4]$, and $\mathcal{L}_{Reg} \in [0.008, 0.01]$. The coefficients $\lambda_{OT} = 10$ and $\lambda_{Reg} = 50$ normalise each term to a comparable magnitude, where $\lambda_{OT}\mathcal{L}_{OT}$ and $\lambda_{Reg}\mathcal{L}_{Reg}$ contribute ≈ 25 –40% and ≈ 5 –10% of $\mathcal{L}_{LM}$, respectively. They are thus treated as fixed loss-scale normalizers rather than

Table 3: Generalisation of our asymmetric OT framework across backbones: results on XNLI, XQuAD, and MMMLU for Qwen3.5-0.8B before and after OT alignment, compared against Llama3-8B-OT. * denotes the post-trained (instruction-tuned) checkpoint of each backbone family. Llama3-8B*-OT is the main OT method reported in Table 2. yellow shaded cells denote English, red shaded cells denote Chinese and blue shaded cells denote overall performance.

XNLI (Acc)																
Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Overall
Qwen3.5-0.8B*	42.33	54.35	64.37	48.93	50.15	54.35	59.65	52.49	66.33	43.27	49.23	60.97	50.12	41.39	49.12	52.71
Qwen3.5-0.8B*-OT	52.33	59.32	65.24	50.78	51.32	69.54	64.57	57.68	68.29	45.82	50.41	63.59	50.72	44.77	42.37	55.78
Llama3-8B*-OT	67.25	69.88	72.34	67.90	83.41	74.25	73.41	66.63	72.30	56.61	66.11	65.41	62.38	69.56	70.84	69.36

XQuAD (F1/EM)																
Model	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh				Overall
Qwen3.5-0.8B*	49.87/38.72	52.68/46.22	42.02/33.98	66.98/64.44	63.73/42.69	47.82/36.04	61.05/40.47	34.17/30.92	51.87/40.23	44.71/39.86	51.03/41.65	60.35/51.26	—	—	—	53.93/43.46
Qwen3.5-0.8B*-OT	52.12/39.43	54.35/45.67	43.12/35.65	67.89/65.93	65.43/44.23	48.93/35.12	64.28/43.63	35.49/31.42	55.65/42.18	45.36/41.05	56.76/44.81	58.62/49.87	—	—	—	54.78/44.56
Llama3-8B*-OT	61.93/43.87	70.05/54.11	66.88/46.89	83.43/72.10	75.32/56.89	59.05/42.27	75.46/59.83	59.19/40.92	61.64/53.19	62.93/40.76	72.11/51.01	62.90/57.98	—	—	—	67.60/51.51

MMMLU (Acc)																
Model	AR	BN	DE	ES	FR	HI	ID	IT	JA	KO	PT	SW	YO	ZH	Def	Overall
Qwen3.5-0.8B*	34.52	16.34	32.45	41.66	45.12	34.54	40.11	47.36	31.65	35.91	42.94	28.95	17.88	47.22	35.43	35.47
Qwen3.5-0.8B*-OT	35.69	15.38	34.26	46.32	47.89	35.47	41.54	48.73	35.87	39.55	43.33	30.44	18.72	43.66	36.78	36.91
Llama3-8B*-OT	44.26	25.59	52.82	55.74	54.56	37.94	50.38	53.66	47.12	46.68	54.99	36.75	28.52	51.62	45.76	45.76

tuned hyperparameters, and held constant across all experiments. We instead ablate factors that directly affect performance: Sinkhorn parameters (ϵ , iterations), each loss component, and the choice of aligned transformer layers. We use Llama3-8B-Instruct to ablate the significant of components.

5.1 Architectural Ablations

Figures 2 and Figure 5 reveal two complementary insights about the loss components (see Appendix A.3 for per-language details).

OT alignment is powerful but requires $\mathcal{L}_{\text{Reg}}$ to anchor English. The *OT only* variant achieves the strongest peaks on several European and Asian languages in XNLI (e.g. de 73.07, es 75.27, vi 70.66), yet its English XNLI drops to 78.18—over five points below the Finetuning checkpoint—because shared LoRA weights accumulate updates that indirectly distort the English branch. Adding $\mathcal{L}_{\text{Reg}}$ (*OT + Reg*) fully restores English fidelity (83.47) and lifts the XNLI average to **70.28**, outperforming *MIDALIGN* (69.14) without any generative supervision.

$\mathcal{L}_{\text{LM}}$ is indispensable for extractive QA. Every variant lacking $\mathcal{L}_{\text{LM}}$ collapses on XQuAD (*OT only*: 28.4 F1; *OT + Reg*: 48.3 F1), as span extraction requires token-level generative capacity that distributional alignment alone cannot provide.

The full objective achieves the most consistent balance across all evaluated tasks. *LM + OT + Reg* is the only variant that sustains competitive performance simultaneously across all tasks: XNLI 69.4, XQuAD 67.6/51.5, and MMMLU 45.8, with English scores matching or exceeding the Finetuning checkpoint throughout.

5.2 Layer Selection Ablations

Figures 3 and Figure 6 compare OT alignment applied to Early (1–8), Middle (8–24), and Deep (24–32) layers of Llama-3-8B (see Appendix A.4 for per-language details).

Middle layers are the optimal alignment target. Middle layers dominate on three of four metrics (XNLI 69.4, XQuAD 67.6 F1/51.5 EM, MMMLU 45.8), outperforming both early and deep configurations. Early layers encode surface-level features that differ sharply across language families, forcing the transport plan to bridge typological rather than semantic dissimilarities. Deep layers have committed to task-specific English-conditioned patterns that OT cannot undo. Middle layers escape both failure modes by carrying language-agnostic semantic abstractions amenable to distributional transport. The gains are sharpest for morphologically distant languages: Arabic improves by +15.1 F1 and Hindi by +12.9 F1 over early layers, gaps that dwarf improvements on typologically closer languages. Deep layers show a more nuanced failure — competitive English scores but sharp degradation on Arabic (44.9 F1) and Hindi (43.1 F1) extraction, a skewed profile visible in Figure 6.

Learned weights independently corroborate this finding. Without any explicit supervision over layer identity, the learnable aggregation weights $\{\alpha_l\}$ concentrate over 52% of total alignment budget on layers 12–14 ($\alpha_{13}=0.181$, $\alpha_{12}=0.177$, $\alpha_{14}=0.171$), decaying to near-zero beyond layer 15. This convergent signal from two independent sources jointly validates middle layers as the semantic core of cross-lingual transferability,

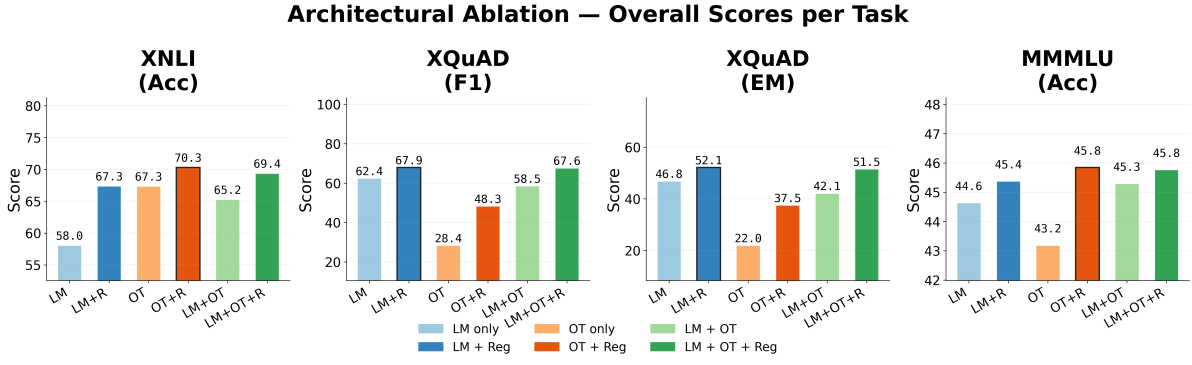


Figure 2: Overall scores per task for each loss-component variant. OT+Reg achieves the highest XNLI (70.3) while LM+Reg leads on XQuAD F1 (67.9); the full objective LM+OT+Reg delivers the most balanced performance across all four metrics.

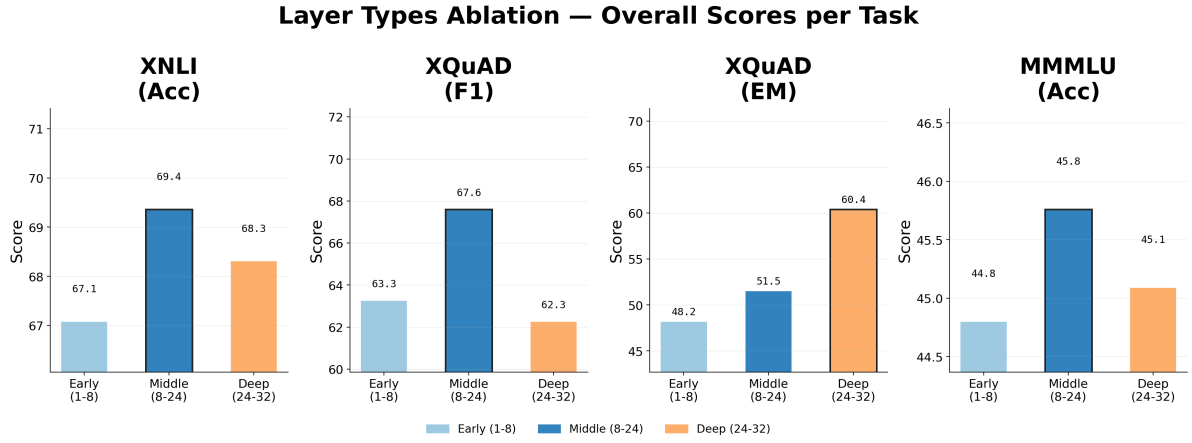


Figure 3: Overall scores per task across three layer groups (see Appendix 5 for per-language details). Middle layers (8–24) achieve the best XNLI (69.4), XQuAD F1 (67.6), and MMMLU (45.8), while Deep layers (24–32) lead only on XQuAD EM (60.4) at the cost of weaker cross-lingual transfer on all other metrics.

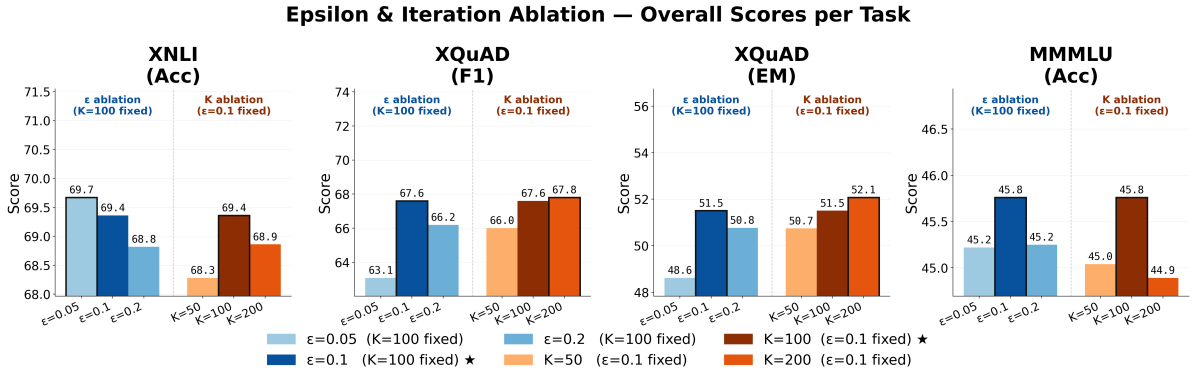


Figure 4: Overall scores per task for Sinkhorn ϵ (left, $K=100$ fixed) and iteration count (right, $\epsilon=0.1$ fixed) ablations. $\epsilon=0.1$ with $K=100$ achieves the best aggregate balance across all four metrics, while smaller ϵ improves XNLI at the cost of XQuAD and larger K yields diminishing returns with increased wall-clock overhead.

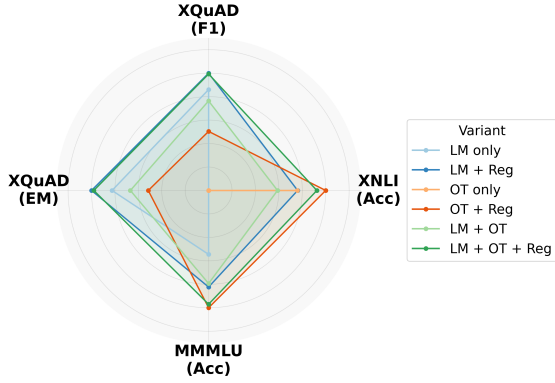


Figure 5: Radar chart comparing all six loss variants across four metrics. LM+OT+Reg (dark green) covers the largest balanced area, whereas OT-only variants (orange) collapse on XQuAD despite strong XNLI scores.

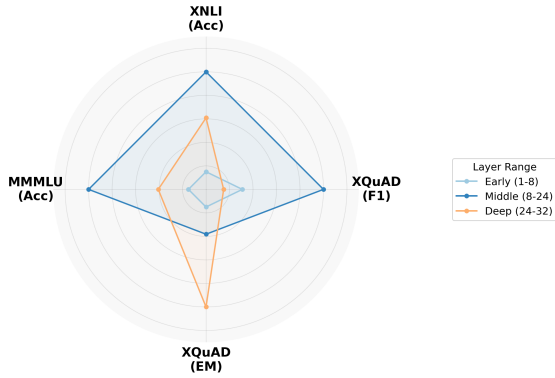


Figure 6: Radar chart comparing alignment across Early, Middle, and Deep layer groups. Middle layers (dark blue) cover all four metrics most uniformly; Deep layers (orange) show an asymmetric profile, strong on XNLI but collapsing on XQuAD EM.

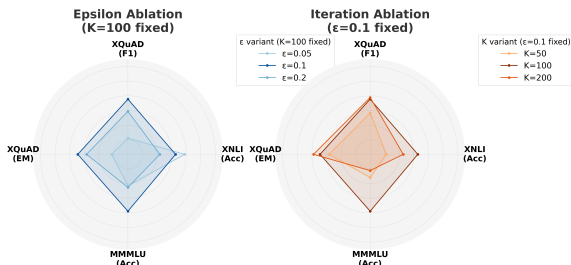


Figure 7: Radar charts for ϵ ablation ($K=100$ fixed, left) and iteration ablation ($\epsilon=0.1$ fixed, right). $\epsilon=0.1$ covers the largest balanced area across all metrics; for K , performance saturates at $K=100$, with $K=50$ visibly weaker and $K=200$ offering no consistent gain.

and provides principled justification for our layer range of 8–24. See Appendix A.6 (Figure 8).

5.3 Sinkhorn’s Parameters Ablations

Figures 4 and Figure 7 examine two orthogonal axes of the Sinkhorn solver: ϵ controls transport plan sharpness, while K governs convergence quality (see Appendix A.5 for per-language details).

ϵ governs the sharpness–generation trade-off. Fixing $K=100$ and varying $\epsilon \in \{0.05, 0.1, 0.2\}$ reveals a clear tension. A sharp plan ($\epsilon=0.05$) maximises XNLI (69.7) but damages generative capacity: XQuAD collapses to 63.1 F1/48.6 EM, with Arabic dropping nearly 14 F1 points versus $\epsilon=0.1$. An overly diffuse plan ($\epsilon=0.2$) weakens the alignment signal: XNLI falls to 68.8 and XQuAD to 66.2 F1. The moderate setting $\epsilon=0.1$ achieves the best aggregate balance (XNLI 69.4, XQuAD 67.6/51.5, MMMLU 45.8) and is adopted as the default.

K governs convergence; quality saturates at $K=100$. Fixing $\epsilon=0.1$, $K=50$ yields the weakest results (XNLI 68.3, MMMLU 45.0) due to an under-converged transport gradient. $K=200$ marginally improves XQuAD (67.8 F1) but regresses on XNLI (68.9) and MMMLU (44.9). Although quadrupling K from 50 to 200 nearly triples wall-clock training time (23.98 h $\rightarrow$ 56.79 h), the analytic FLOP cost remains virtually unchanged (Table 7), since Sinkhorn iterations contribute negligibly relative to the dominant LLM forward/backward passes. $K=100$ is therefore the optimal efficiency–quality trade-off.

6 Conclusions

We proposed an Optimal Transport framework for cross-lingual alignment in autoregressive LLMs that transports non-English hidden-state distributions toward a frozen English anchor at middle transformer layers, paired with an English consistency regularisation term. This asymmetric formulation preserves English proficiency by construction, mitigating a structural weakness of symmetric contrastive methods. Experiments on XNLI, XQuAD, and MMMLU confirm consistent cross-lingual gains without regressing on English. Ablation studies confirm that all three objective components are necessary, with learnable layer aggregation self-localising alignment to the lower-middle layers where transfer is most effective.

Limitations

Although our comparison includes a limited number of baselines, this is intentional: our core contribution targets the structural weakness of symmetric contrastive alignment, and MIDALIGN serves as the most direct and suitable representative of this paradigm, sharing our middle-layer motivation and alignment design. AlignX is excluded due to its fundamentally different zero-shot, continued pre-training setup, which prevents fair comparison across our task-specific benchmarks. Additionally, the loss coefficients λ_{OT} and λ_{Reg} are treated as fixed loss-scale normalizers rather than tuned hyperparameters; their behaviour under different model scales, data regimes, or training dynamics remains unexplored, and alternative weightings may yield further gains in settings where the three objectives compete more sharply for the same LoRA parameters. Furthermore, while our method consistently improves cross-lingual transfer on average, per-language gains remain uneven, as the OT framework can only transport representations within the geometric space the base model has already learned from pretraining data, leaving a residual performance gap that distributional alignment alone cannot fully close. Our main experiments use a single, untuned backbone (Llama-3-8B-Instruct); a supplementary Qwen3.5-0.8B* run confirms generalisation but also reveals a boundary case where transporting an already-strong non-English representation (Chinese) toward the English anchor is counterproductive, motivating adaptive, language-specific anchors as future work.

Acknowledgments

We acknowledge the Center for AI Research (CAIR), VinUniversity, Hanoi, Vietnam, Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, and Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam, for supporting this study. We also acknowledge Mr. Tan-Ha Mai of National Taiwan University for providing computational support.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. [MEGA: Multilingual evaluation of generative AI](#).

In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4232–4267, Singapore. Association for Computational Linguistics.

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. 2021. [Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3904–3919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. [On the cross-lingual transferability of monolingual representations](#). *CoRR*, abs/1910.11856.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Mengyu Bu, Shaolei Zhang, Zhongjun He, Hua Wu, and Yang Feng. 2025. [AlignX: Advancing multilingual large language models with multilingual representation alignment](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6460–6489, Suzhou, China. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. [Multilingual alignment of contextual word representations](#). In *International Conference on Learning Representations*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. [InfoXLM: An information-theoretic framework for cross-lingual language model pre-training](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Rutu Rinott, Guillaume Lample, 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, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.
- Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- 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*.
- Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.
- L. Kantorovitch. 1958. On the translocation of masses. *Management Science*, 5(1):1–4.
- Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8058–8076, Mexico City, Mexico. Association for Computational Linguistics.
- Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: a metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’25/IAAI’25/EAAI’25*. AAAI Press.
- Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15979–15996, Vienna, Austria. Association for Computational Linguistics.
- Thong Nguyen and Anh Tuan Luu. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *AAAI Conference on Artificial Intelligence*.
- G. Peyré and M. Cuturi. 2019. Computational optimal transport. *Foundations and Trends in Machine Learning*, 51(1):1–44.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- Bernhard Schmitzer. 2019. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481.
- Tom Sherborne, Tom Hosking, and Mirella Lapata. 2023. Optimal transport posterior alignment for cross-lingual semantic parsing. *Transactions of the Association for Computational Linguistics*, 11:1432–1450.
- Richard Sinkhorn and Paul Knopp. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21:343–348.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- Cédric Villani. 2003. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island.
- Cédric Villani. 2009. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg.
- Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. [Probing the emergence of cross-lingual alignment during LLM training](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. 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.

A Appendices

A.1 Preliminaries: Optimal Transport

Kantorovich Formulation. Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be two probability measures defined over metric spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. The *Kantorovich problem* seeks a *transport plan* (Kantorovitch, 1958)—a joint probability measure $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ —that couples μ and ν at minimum expected cost:

$$\mathcal{L}_{\text{OT}}(\mu, \nu) = \min_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y), \quad (9)$$

where $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$ is a ground cost function and $\Pi(\mu, \nu)$ is the set of all couplings with marginals μ and ν :

$$\Pi(\mu, \nu) = \left\{ \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid \pi(\cdot, \mathcal{Y}) = \mu, \pi(\mathcal{X}, \cdot) = \nu \right\}. \quad (10)$$

When $c(x, y) = d(x, y)^p$ for a metric d and $p \geq 1$, the p -th root of (9) defines the p -Wasserstein distance:

$$\mathcal{W}_p(\mu, \nu) = \left(\min_{\pi \in \Pi(\mu, \nu)} \int d(x, y)^p d\pi(x, y) \right)^{1/p}. \quad (11)$$

The Wasserstein distance is a proper metric on the space of probability measures (Villani, 2003, 2009), making it geometrically meaningful for comparing distributions over continuous embedding spaces.

Computational Bottleneck of the Kantorovich Problem. In the discrete setting—where $\mu = \sum_{i=1}^m a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n b_j \delta_{y_j}$ —the Kantorovich problem becomes a *linear programme* (LP) over the $(m \times n)$ -dimensional transportation polytope:

$$\min_{\mathbf{T} \geq 0} \langle \mathbf{C}, \mathbf{T} \rangle_F \quad \text{s.t.} \quad \mathbf{T} \mathbf{1}_n = \mathbf{a}, \quad \mathbf{T}^\top \mathbf{1}_m = \mathbf{b}, \quad (12)$$

where $\mathbf{C} \in \mathbb{R}^{m \times n}$ is the cost matrix, $\mathbf{T} \in \mathbb{R}_{\geq 0}^{m \times n}$ is the discrete transport plan, and $\mathbf{a} \in \Delta^m$, $\mathbf{b} \in \Delta^n$ are the probability weight vectors. Although this LP can be solved exactly via interior-point or simplex methods, the complexity scales as $\mathcal{O}((m+n)^3 \log(m+n))$ (Peyré and Cuturi, 2019), which is prohibitively expensive for sequences of hundreds of tokens computed across multiple layers per training step.

A.2 Preliminaries: Entropy-Regularised Optimal Transport (Sinkhorn Algorithm)

Entropy Regularisation. To overcome the computational barrier of the Kantorovich LP, Cuturi

(2013) introduced an entropy-regularised relaxation. The discrete entropy of a transport plan is defined as:

$$H(\mathbf{T}) = - \sum_{i,j} T_{ij} (\log T_{ij} - 1). \quad (13)$$

Adding a negated entropy term as a regulariser yields the *Sinkhorn problem*:

$$\mathbf{T}_\varepsilon^* = \arg \min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{T} \rangle_F - \varepsilon H(\mathbf{T}), \quad (14)$$

where $\varepsilon > 0$ is the temperature hyperparameter controlling the trade-off between transport cost fidelity and plan smoothness. By strong duality, the unique solution to (14) takes the form:

$$T_{\varepsilon, ij}^* = u_i \cdot K_{ij} \cdot v_j, \quad \mathbf{K} = \exp(-\mathbf{C}/\varepsilon), \quad (15)$$

where $\mathbf{u} \in \mathbb{R}_{>0}^m$ and $\mathbf{v} \in \mathbb{R}_{>0}^n$ are dual scaling vectors recovered by *Sinkhorn–Knopp iterations* (Sinkhorn and Knopp, 1967):

$$\mathbf{u}^{(t+1)} = \frac{\mathbf{a}}{\mathbf{K} \mathbf{v}^{(t)}}, \quad \mathbf{v}^{(t+1)} = \frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}^{(t+1)}}, \quad (16)$$

initialised with $\mathbf{v}^{(0)} = \mathbf{1}_n$. These updates alternate projections onto the two marginal constraints and converge linearly in $\mathcal{O}(mn)$ per iteration, reducing overall complexity to $\mathcal{O}(mn/\varepsilon^2)$ —a dramatic improvement over the cubic LP solver.

Log-Domain Stabilisation. For small ε , the Gibbs kernel $K_{ij} = e^{-C_{ij}/\varepsilon}$ can underflow to zero in floating-point arithmetic. We therefore maintain log-scaling vectors $\mathbf{f} = \varepsilon \log \mathbf{u}$ and $\mathbf{g} = \varepsilon \log \mathbf{v}$ (the Kantorovich dual potentials) and perform the iterations in log-domain (Schmitzer, 2019):

$$f_i^{(t+1)} = \varepsilon \log a_i - \varepsilon \log \sum_j \exp\left(\frac{g_j^{(t)} - C_{ij}}{\varepsilon}\right), \quad (17)$$

$$g_j^{(t+1)} = \varepsilon \log b_j - \varepsilon \log \sum_i \exp\left(\frac{f_i^{(t+1)} - C_{ij}}{\varepsilon}\right). \quad (18)$$

The log-sum-exp operations are numerically stable and amenable to automatic differentiation, enabling gradient flow through the OT objective.

A.3 Detail Architectural Ablations Study

Table 4 reports performance of six objective variants obtained by selectively enabling or disabling each loss component: $\mathcal{L}_{\text{LM}}$, $\mathcal{L}_{\text{OT}}$, and $\mathcal{L}_{\text{Reg}}$. We

Table 4: Architectural Ablations on downstream benchmarks. Bold indicates best, underline indicates second best. Parameters $\lambda_{\text{Reg}} = 50$, $\lambda_{\text{OT}} = 10$ and Sinkhorn parameters $\epsilon = 0.1$, $\text{iter} = 100$ are kept fixed across all variants, where the main model is $\mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{OT}} + \mathcal{L}_{\text{Reg}}$; yellow shaded cells denote English and blue shaded cells denote overall performance.

	XNLI (Acc)															
Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Overall
LM only	54.77	58.62	63.03	55.93	76.49	61.94	61.00	55.29	59.88	48.66	54.05	55.63	49.18	55.63	60.68	58.05
LM + Reg	63.19	67.90	72.12	65.77	83.15	72.79	72.48	64.51	70.84	54.83	63.05	63.47	59.14	67.43	69.60	67.35
OT only	65.93	69.72	73.07	70.00	78.18	75.27	74.47	59.56	72.20	53.27	62.79	61.72	53.17	70.66	69.70	67.31
OT + Reg	68.78	72.51	72.53	69.34	83.47	74.01	73.75	67.76	72.48	59.64	66.85	67.62	62.73	71.36	71.32	70.28
LM + OT	65.25	70.14	72.02	68.86	79.84	73.89	72.67	66.49	71.28	60.00	65.83	65.37	60.46	69.74	70.22	65.25
LM + OT + Reg	67.25	69.88	72.34	67.90	83.41	74.25	73.41	66.63	72.30	56.61	66.11	65.41	62.38	69.56	70.84	69.36
	XQuAD (F1/EM)															
Model	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh				Overall
LM only	55.13/36.89	65.69/49.50	60.39/41.01	81.18/67.31	68.21/49.66	56.59/40.25	68.52/53.70	54.79/38.24	54.57/46.39	59.94/40.50	66.48/45.63	56.77/52.77	-	-	-	62.36/46.82
LM + Reg	61.33/44.03	70.45/54.20	66.70/47.65	83.41/72.10	74.16/55.63	60.42/43.61	75.55/59.75	58.88/41.85	63.65/54.96	63.98/42.10	72.84/51.01	63.16/58.32	-	-	-	67.88/52.10
OT only	12.81/9.50	38.38/30.34	22.17/16.97	79.92/68.15	35.29/26.64	14.34/11.51	36.23/28.40	15.77/10.67	16.86/12.02	24.26/17.73	33.02/23.28	11.11/8.32	-	-	-	28.35/21.96
OT + Reg	15.95/11.85	69.34/53.95	41.11/29.33	83.14/72.02	72.20/54.54	15.22/12.44	73.02/58.57	42.74/29.50	18.53/14.29	49.88/35.88	50.38/34.96	47.66/42.52	-	-	-	48.26/37.49
LM + OT	51.61/34.12	64.83/46.72	59.54/38.74	82.96/69.50	67.68/47.98	44.89/30.17	67.62/50.67	49.94/31.34	47.45/36.89	49.64/31.68	65.20/42.86	50.96/44.54	-	-	-	58.53/42.10
LM + OT + Reg	61.93/43.87	70.05/54.11	66.88/46.89	83.43/72.10	75.32/56.89	59.05/42.27	75.46/59.83	59.19/40.92	61.64/53.19	62.93/40.76	72.11/51.01	62.90/57.98	-	-	-	67.60/51.51
	MMMLU (Acc)															
Model	AR	BN	DE	ES	FR	HI	ID	IT	JA	KO	PT	SW	YO	ZH	Def	Overall
LM only	43.35	24.80	51.37	53.92	53.60	36.90	49.21	52.63	45.48	45.09	53.47	36.39	28.04	50.56	44.63	44.63
LM + Reg	44.11	24.71	52.68	55.93	53.82	38.06	50.20	52.35	47.51	46.70	54.86	35.58	28.23	50.53	45.24	45.37
OT only	41.98	28.80	49.84	52.79	51.47	37.16	46.31	50.09	43.06	42.31	51.30	33.11	27.35	48.83	43.19	43.18
OT + Reg	44.07	32.50	52.02	55.25	54.52	39.96	49.32	52.56	46.52	46.11	53.76	35.02	29.40	50.88	45.83	45.84
LM + OT	43.69	25.14	52.38	54.92	54.31	38.35	49.96	53.28	45.75	45.66	54.74	36.63	28.20	51.15	45.27	45.29
LM + OT + Reg	44.26	25.59	52.82	55.74	54.56	37.94	50.38	53.66	47.12	46.68	54.99	36.75	28.52	51.62	45.76	45.76

discuss the findings in terms of three inter-related themes: the alignment strength of $\mathcal{L}_{\text{OT}}$, the stabilising role of $\mathcal{L}_{\text{Reg}}$, the generative necessity of $\mathcal{L}_{\text{LM}}$, and the balancing act achieved by the full objective. We use Llama3-8B-Instruct to ablate the significant of components.

OT is a powerful alignment signal but cannot anchor English on its own. The *OT only* variant demonstrates that $\mathcal{L}_{\text{OT}}$ alone is a remarkably effective cross-lingual alignment mechanism: it achieves the highest single-language peaks on XNLI for several European languages (e.g. de 73.07, es 75.27, fr 74.47), especially vi 70.66, surpassing every other variant at those positions. However, this alignment strength comes at a significant cost to English fidelity. The *OT only* variant scores only 78.18 on the English portion of XNLI—a regression of more than five points relative to the *Finetuning* checkpoint (83.25) and to the full model (83.41). This degradation is noteworthy because the OT objective is *asymmetric* by design: only the other-language representations are pulled toward the frozen English anchor, not the reverse. Nevertheless, the LoRA parameters trained solely under $\mathcal{L}_{\text{OT}}$ accumulate updates that indirectly distort the English branch, since the trainable LoRA weights are shared across all forward passes. The result confirms that architectural asymmetry alone is insufficient to preserve English performance; an explicit regularisation term is necessary.

$\mathcal{L}_{\text{Reg}}$ is the critical stabiliser. Adding $\mathcal{L}_{\text{Reg}}$ to the OT signal (*OT + Reg*) restores English XNLI accuracy to 83.47—exceeding even the *Finetuning*

ing baseline—while sustaining strong cross-lingual transfer (average 70.28, the highest across all variants on XNLI). Critically, *OT + Reg* outperforms the MIDALIGN baseline from the main results (69.14 average), demonstrating that the combination of asymmetric OT alignment with an explicit English anchor is a more principled cross-lingual strategy than symmetric contrastive alignment. The mechanism is intuitive: $\mathcal{L}_{\text{OT}}$ pulls other-language distributions toward the English anchor in representation space, while $\mathcal{L}_{\text{Reg}}$ penalises deviations of the trainable English branch from the frozen one, keeping that anchor stationary throughout training. Together they enforce a one-directional transport without allowing the anchor to drift. The complementarity is also reflected in the layer-weight interaction: both losses share the same learned aggregation weights $\{\alpha_l\}$, focusing regularisation pressure on exactly the layers where transport is most active (layers 12–14; see Figure 8).

The OT objective is structurally incompatible with extractive QA absent a language modelling loss. Despite its strong classification performance, the *OT only* variant produces catastrophic scores on XQuAD (28.35/21.96), a collapse even more severe than that of MIDALIGN (30.5/23.9) in the main comparison. *OT + Reg* partially recovers (48.26/37.49) but remains far below the *Finetuning* baseline (62.2/47.3). The root cause is the absence of $\mathcal{L}_{\text{LM}}$: without autoregressive supervision, the model loses the token-level generative capacity required for span extraction. Classification heads can exploit distributional

alignment in the final hidden state even under representation distortion, but span extraction requires the model to generate precise character sequences conditioned on contextual positions—a capability that degrades rapidly when generative gradients are withheld. This asymmetry between classification and generation tasks is a fundamental tension in any representation-level alignment scheme, and it directly motivates the inclusion of $\mathcal{L}_{\text{LM}}$ in the full objective.

$\mathcal{L}_{\text{LM}}$ restores generation but dilutes alignment.

Incorporating $\mathcal{L}_{\text{LM}}$ alongside the OT objective ($LM + OT$) raises XQuAD to 58.53/42.10, a significant recovery, but notably weakens XNLI average (65.25) relative to *OT only* (67.31). This reveals a tension: the language modelling gradient updates the model toward next-token prediction, partially counteracting the distributional alignment pressure of $\mathcal{L}_{\text{OT}}$. In effect, the two objectives compete for the same LoRA parameters, and without $\mathcal{L}_{\text{Reg}}$ to anchor English, the three-way tug results in inconsistent performance across task types.

$\mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{Reg}}$ achieves the best extractive QA.

Interestingly, the variant that omits the OT loss entirely but combines language modelling with regularisation ($LM + \text{Reg}$) attains the highest XQuAD score across all ablation variants (67.88/52.10), surpassing even the full model ($LM + OT + \text{Reg}$, 67.60/51.51). $\mathcal{L}_{\text{Reg}}$ acts here not as an alignment signal but as a strong weight-space constraint that prevents the LoRA fine-tuning from overfitting to the other language and forgetting English generative patterns. This result exposes a further trade-off: maximising XQuAD performance appears to require sacrificing some of the cross-lingual alignment pressure provided by $\mathcal{L}_{\text{OT}}$, since the OT gradient competes with the generative objective when both are active. Conversely, on XNLI the $LM + \text{Reg}$ variant scores only 67.35, substantially below *OT + Reg* (70.28) and the full model (69.36), confirming that $\mathcal{L}_{\text{OT}}$ is indispensable for strong sentence-level cross-lingual transfer.

The full objective $\mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{OT}} + \mathcal{L}_{\text{Reg}}$ is the balanced solution. No single-loss or two-loss variant achieves consistently competitive performance across all three benchmarks simultaneously. *OT + Reg* leads on XNLI but collapses on XQuAD; *LM + Reg* leads on XQuAD but lags on XNLI; *LM + OT* partially compromises both. The full three-component objective $LM + OT + \text{Reg}$ is the

only variant that sustains competitive performance across classification (XNLI 69.36, second only to *OT + Reg*), extractive QA (XQuAD 67.60/51.51, second only to *LM + Reg*), and knowledge recall (MMMLU 45.76). Crucially, it is also the only variant that maintains English XNLI (83.41) and English XQuAD (83.43/72.10) at or above the *Fine-tuning* checkpoint, validating the combined effect of $\mathcal{L}_{\text{Reg}}$ and $\mathcal{L}_{\text{LM}}$ as dual safeguards of English proficiency. The ablation thus provides principled justification for each component: OT delivers alignment, regularisation preserves the anchor, and language modelling sustains generative capacity—each component contributes meaningfully, and removing any one leads to degradation on at least one benchmark.

A.4 Detail Layer Selection Ablation Study

We partition the 32 transformer layers of Llama-3-8B into three non-overlapping groups to examine where cross-lingual alignment is most effective: *Early* (layers 1–8), *Middle* (layers 8–24), and *Deep* (layers 24–32). This partition is motivated by the layer specialisation hypothesis: early layers encode surface-level syntactic and typological features that are highly language-specific; middle layers carry semantic abstractions that are largely language-agnostic and therefore most amenable to cross-lingual transfer; and deep layers encode task-specific reasoning patterns that alignment should not disturb. Table 5 reports results across all three benchmarks.

Early layers encode surface form, not transferable semantics.

Applying $\mathcal{L}_{\text{OT}}$ to layers 1–8 yields the weakest overall performance on both XNLI (67.08) and MMMLU (44.80), and produces notably lower cross-lingual transfer on XQuAD (63.26/48.19) compared to Middle (67.60/51.51). The degradation is most pronounced for morphologically distant languages: Arabic drops to 46.79/32.61 on XQuAD versus 61.93/43.87 for Middle, and Hindi to 46.20/32.18 versus 59.05/42.27. This is consistent with the view that early-layer representations are dominated by language-surface statistics—script, morphology, tokenisation artefacts—which differ sharply across language families, making them poor targets for a geometry-based transport objective. Aligning at this level forces the transport plan to bridge surface dissimilarities rather than semantic ones, wasting alignment capacity on typological noise.

Table 5: Layer Selection Ablations on XNLI, XQuAD, and MMMLU. Bold indicates best, underline indicates second best; yellow shaded cells denote English and blue shaded cells denote overall performance.

XNLI (Acc)																
Layer Type	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Overall
Early (1-8)	64.41	67.58	70.69	64.26	81.31	72.07	72.19	64.35	69.04	55.19	62.07	65.71	60.30	68.16	68.80	67.08
Middle (8-24)	67.25	69.88	72.34	67.90	83.41	74.25	73.41	66.63	72.30	56.61	66.11	65.41	62.38	69.56	70.84	69.36
Deep (24-32)	65.79	67.76	72.26	67.64	80.49	73.77	73.21	65.89	71.66	55.11	65.49	65.39	60.48	69.94	69.72	68.31
XQuAD (F1/EM)																
Layer Type	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	Overall			
Early (1-8)	46.79/32.61	70.57/53.36	59.58/41.34	83.34/72.02	73.19/55.97	46.20/32.18	74.44/58.91	53.75/36.55	55.11/45.63	63.99/44.71	71.34/51.43	60.84/71.34	—	—	—	63.26/48.19
Middle (8-24)	61.93/43.87	70.05/54.11	66.88/46.89	83.43/72.10	75.32/56.89	59.05/42.27	75.46/59.83	59.19/40.92	61.64/53.19	62.93/40.76	72.11/51.01	62.90/57.98	—	—	—	67.60/51.51
Deep (24-32)	44.91/30.84	69.63/53.70	58.64/40.17	83.37/72.27	73.61/55.63	43.14/29.83	44.12/30.46	74.13/58.57	52.54/35.97	52.29/42.94	70.81/63.65	60.39/53.61	—	—	—	62.26/60.39
MMMLU (Acc)																
Layer Type	AR	BN	DE	ES	FR	HI	ID	IT	JA	KO	PT	SW	YO	ZH	Def	Overall
Early (1-8)	42.95	25.29	51.92	54.95	54.21	36.93	48.98	52.62	46.45	45.58	53.95	35.32	28.34	50.65	44.90	44.80
Middle (8-24)	44.26	25.59	52.82	55.74	54.56	37.94	50.38	53.66	47.12	46.68	54.99	36.75	28.52	51.62	45.76	45.76
Deep (24-32)	43.39	25.40	51.92	54.99	54.49	37.19	49.37	52.71	46.59	45.91	53.77	35.46	28.96	51.17	45.09	45.09

Middle layers are the optimal alignment target.

The Middle variant (layers 8–24) achieves the best overall performance across all three benchmarks: XNLI 69.36, XQuAD 67.60/51.51, and MMMLU 45.76—consistently outperforming both Early and Deep configurations. The gains are most salient on semantically demanding tasks and morphologically distant languages: on XQuAD, Arabic improves by +15.14 F1 and +11.26 EM over Early, and Hindi improves by +12.85 F1 over Early. Crucially, the Middle variant also best preserves English fidelity (XNLI en: 83.41; XQuAD en: 83.43/72.10), confirming that middle layers carry language-agnostic semantic content that transfers without disrupting surface-level or task-specific representations encoded elsewhere. This finding is independently corroborated by the learned layer weights $\{\alpha_l\}$ in Figure 8, where the model concentrates over 52% of alignment weight on layers 12–14 entirely without supervision over layer identity—a convergent signal from two independent sources that layers 8–24 are the semantic core of cross-lingual transferability.

Deep layers are task-specialised and resistant to alignment. Targeting layers 24–32 yields intermediate performance: XNLI 68.31 and MMMLU 45.09, both below Middle but above Early on most metrics. However, XQuAD reveals a more nuanced failure: while the Deep variant achieves competitive English scores (en: 83.37/72.27), its cross-lingual extraction degrades sharply for morphologically distant languages (ar: 44.91/30.84, hi: 43.14/29.83), falling closer to the Early pattern than to Middle. This suggests that deep layers have already specialised toward task-specific token prediction conditioned on English surface patterns, making OT-based redistribution of their representations counterproductive: the transport plan can shift distributions but cannot undo the task-specific

encoding that these layers have committed to. A further observation is that the Deep variant shows unusually high EM on Vietnamese (vi: 63.65) relative to its F1 (70.81), suggesting that when deep-layer alignment does succeed, it tends to produce more precise but less recall-oriented span extraction.

A.5 Detail Sinkhorn’s Parameters Ablations

We ablate two Sinkhorn hyperparameters: the entropy regularisation coefficient ε , which controls the sharpness of the transport plan, and the iteration count K , which governs convergence quality. Results are reported in Table 6.

Moderate entropy yields the best cross-task balance. Fixing $K=100$ and varying $\varepsilon \in \{0.05, 0.1, 0.2\}$, we observe a clear tension between alignment sharpness and generative preservation. At $\varepsilon=0.05$, the transport plan is highly concentrated, producing the strongest sentence-level classification signal and the best XNLI average (69.67); however, the sharp plan damages the generative capacity required for span extraction, causing XQuAD to collapse to 63.08/48.61—with Arabic dropping to 48.21/33.53 F1/EM, nearly 14 points below the $\varepsilon=0.1$ setting (61.93/43.87). At $\varepsilon=0.2$, the plan becomes overly diffuse, providing insufficient distributional pull: XNLI falls to 68.82 and XQuAD to 66.19/50.76. Setting $\varepsilon=0.1$ achieves the best aggregate performance across all three benchmarks—XNLI 69.36, XQuAD 67.60/51.51, MMMLU 45.76—and is therefore adopted as the default.

Convergence quality saturates rapidly beyond 100 iterations. Fixing $\varepsilon=0.1$ and varying $K \in \{50, 100, 200\}$, we find that plan quality improves meaningfully from $K=50$ to $K=100$ but plateaus (or even regresses) beyond that point. At $K=50$,

Table 6: Epsilon and Iteration Ablations on XNLI, XQuAD, and MMMLU. Bold indicates best, underline indicates second best; yellow shaded cells denote English and blue shaded cells denote overall performance.

	XNLI (Acc)															
Eps-Iter	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Overall
$\epsilon=0.05, K=100$	66.35	70.58	72.46	68.76	83.27	74.25	73.77	67.13	72.63	57.41	67.23	66.61	63.03	70.24	71.38	69.67
$\epsilon=0.1, K=100$	67.25	69.88	72.34	67.90	83.41	74.25	73.41	66.63	72.30	56.61	66.11	65.41	62.38	69.56	70.84	69.36
$\epsilon=0.2, K=100$	65.39	69.48	72.46	67.76	83.29	73.75	73.39	66.61	72.38	55.61	65.59	65.25	61.40	68.90	71.08	68.82
	XQuAD (F1/EM)															
Eps-Iter	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh				Overall
$\epsilon=0.05, K=100$	48.21/33.53	69.46/53.45	61.14/43.87	83.37/72.52	73.67/56.55	42.25/29.58	74.30/58.91	52.11/35.80	58.02/49.33	62.83/44.37	71.18/50.92	60.39/54.54	-	-	-	63.08/48.61
$\epsilon=0.1, K=100$	61.93/43.87	70.05/54.11	66.88/46.89	83.43/72.10	75.32/56.89	59.05/42.27	75.46/59.83	59.19/40.92	61.64/53.19	62.93/40.76	72.11/51.01	62.90/57.98	-	-	-	67.60/51.51
$\epsilon=0.2, K=100$	57.11/40.42	69.74/53.03	65.38/46.89	83.19/72.02	73.93/56.55	53.20/37.90	75.43/60.00	57.41/39.41	59.75/50.42	65.11/45.63	71.92/50.59	62.06/56.30	-	-	-	66.19/50.76
	MMMLU (Acc)															
Eps-Iter	AR	BN	DE	ES	FR	HI	ID	IT	JA	KO	PT	SW	YO	ZH	Def	Overall
$\epsilon=0.05, K=100$	43.13	25.23	51.99	55.30	54.56	37.51	49.57	52.65	46.80	46.41	53.95	35.65	28.31	50.95	45.22	45.22
$\epsilon=0.1, K=100$	44.26	25.59	52.82	55.74	54.56	37.94	50.38	53.66	47.12	46.68	54.99	36.75	28.52	51.62	45.76	45.76
$\epsilon=0.2, K=100$	43.69	25.53	52.22	55.38	54.59	37.26	49.86	52.91	46.80	46.39	54.14	36.00	27.94	50.84	45.25	45.25
	XNLI (Acc)															
Eps-Iter	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Overall
$\epsilon=0.1, K=50$	65.67	67.48	71.38	67.96	83.43	73.79	73.02	67.17	71.38	56.09	66.01	63.23	61.88	67.86	67.88	68.28
$\epsilon=0.1, K=100$	67.25	69.88	72.34	67.90	83.41	74.25	73.41	66.63	72.30	56.61	66.11	65.41	62.38	69.56	70.84	69.36
$\epsilon=0.1, K=200$	65.53	69.78	72.40	67.98	83.31	73.75	73.35	66.69	72.18	55.45	65.61	65.75	61.74	68.84	70.60	68.86
	XQuAD (F1/EM)															
Eps-Iter	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh				Overall
$\epsilon=0.1, K=50$	56.47/40.08	69.70/53.61	65.08/46.72	83.20/72.18	73.90/56.55	52.27/36.97	75.70/60.42	56.46/38.49	59.85/50.76	64.75/45.63	72.64/51.34	62.01/56.13	-	-	-	66.00/50.74
$\epsilon=0.1, K=100$	61.93/43.87	70.05/54.11	66.88/46.89	83.43/72.10	75.32/56.89	59.05/42.27	75.46/59.83	59.19/40.92	61.64/53.19	62.93/40.76	72.11/51.01	62.90/57.98	-	-	-	67.60/51.51
$\epsilon=0.1, K=200$	62.44/44.33	70.58/54.28	67.42/47.06	83.65/72.18	76.26/56.72	61.01/46.72	76.43/59.75	58.82/38.99	57.83/49.16	65.55/46.97	72.62/52.50	61.03/56.21	-	-	-	67.80/52.07
	MMMLU (Acc)															
Eps-Iter	AR	BN	DE	ES	FR	HI	ID	IT	JA	KO	PT	SW	YO	ZH	Def	Overall
$\epsilon=0.1, K=50$	43.79	25.59	52.34	54.66	52.66	37.53	49.79	52.96	46.97	45.33	54.15	36.08	28.29	50.08	45.36	45.04
$\epsilon=0.1, K=100$	44.26	25.59	52.82	55.74	54.56	37.94	50.38	53.66	47.12	46.68	54.99	36.75	28.52	51.62	45.76	45.76
$\epsilon=0.1, K=200$	42.79	24.29	52.65	53.66	53.13	36.39	48.52	53.78	47.71	46.83	54.54	36.38	28.29	49.48	44.84	44.89

the under-converged plan yields the weakest overall results (XNLI 68.28, MMMLU 45.04), as the noisy gradient signal from an insufficiently iterated transport plan attenuates the alignment pressure on non-English representations. At $K=200$, XQuAD marginally improves over $K=100$ (67.80/52.07 vs. 67.60/51.51), but XNLI and MMMLU regress to 68.86 and 44.89 respectively, suggesting that an overly precise plan over-commits to pairwise token matching at the expense of distributional flexibility needed for classification and knowledge recall. $K=100$ leads or ties on the majority of per-language scores and is the only setting that achieves competitive performance simultaneously across all task types.

Wall-clock cost scales with K while FLOP cost does not. Although increasing K from 50 to 200 quadruples the number of Sinkhorn iterations, the analytically computed FLOP per sample remains virtually unchanged (594.781 $\rightarrow$ 594.782 GFLOP), since the iterative transport computation contributes negligibly relative to the three LLM forward passes and LoRA backward pass that dominate the training step. The practical cost, however, scales directly with K : throughput drops from 193.4 to 81.7 samples/s and estimated training time nearly triples from 23.98 h to 56.79 h over three epochs (Table 7). This discrepancy arises from the sequential nature of Sinkhorn iterations on GPU, which increases wall-clock time with-

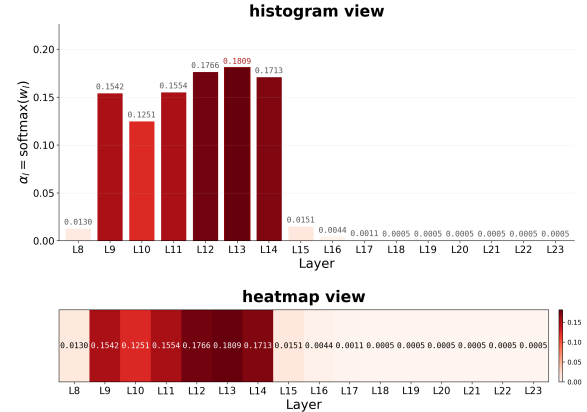


Figure 8: Learned layer weights $\alpha_l = \text{softmax}(w_l)$ across middle layers 8–23. Alignment concentrates in layers 12–14 ($>52\%$ of total weight), with near-zero contribution beyond layer 15.

out proportionally increasing arithmetic intensity. The $K=100$ configuration therefore represents the optimal efficiency–quality trade-off: it closes the performance gap over $K=50$ at a moderate cost increase (3.755 s vs. 2.648 s per batch), while $K=200$ offers diminishing returns at nearly double the wall-clock overhead.

A.6 Where Does Alignment Occur? Localising the Semantic Core

Figure 8 shows the learned softmax weights $\{\alpha_l\}$ across layers 8–23 after Stage 2 training of Llama3-8B-Instruct-OT. The model concentrates over 52%

Setting	Time / batch	Samples / s	Est. time / epoch	Est. time / 3 epochs	FLOP / sample
$\varepsilon=0.1, K=50$	2.648 s	193.4	7.99 h	23.98 h	594.781 GFLOP
$\varepsilon=0.1, K=100$	3.755 s	136.3	11.34 h	34.01 h	594.781 GFLOP
$\varepsilon=0.1, K=200$	6.270 s	81.7	18.93 h	56.79 h	594.782 GFLOP

Table 7: Estimated training cost for different Sinkhorn iteration counts K on $4 \times \text{H100 SXM5 GPUs}$ (global batch size 128, sequence length 256, bfloat16). Timing is reported as the median of three measured runs after two warm-up passes. FLOP per sample is computed analytically and includes three forward passes (frozen EN branch, LoRA-enabled EN branch, and LoRA-enabled target branch) together with LoRA-only backward propagation; each forward pass accounts for both the LoRA adapter and the language modelling head ($\mathcal{L}_{\text{LM}}$), while the frozen backbone contributes no trainable FLOPs. Although the Sinkhorn iteration count increases from $K=50$ to $K=200$, the overall FLOP per sample remains nearly unchanged since the OT computation contributes negligibly compared to the dominant LM-head and LoRA forward/backward cost.

of total weight on layers 12–14 ($\alpha_{12}=0.177$, $\alpha_{13}=0.181$, $\alpha_{14}=0.171$), with weights decaying sharply beyond layer 15 (<0.005 for layers 16–23). This peaked distribution confirms that OT self-localises to the lower-middle layers identified as the semantic core for language-agnostic representations (Tenney et al., 2019; Liu and Niehues, 2025), without any explicit supervision over layer identity. Near-zero weights on upper layers further suggest these encode task-specific features that alignment should not disturb. This validates our learnable aggregation as a core design contribution: rather than applying transport uniformly, the model automatically focuses alignment where cross-lingual transfer is most effective.

A.7 Stage 1 Anchor Quality: English Task Fine-tuning Performance

Since $\mathcal{L}_{\text{Reg}}$ anchors Stage 2 alignment to the frozen English branch produced by Stage 1 (Section 3.4), the quality of this checkpoint directly determines the reliability of the anchor that non-English representations are transported toward. We report the Stage 1 training trajectory here as explicit evidence that the frozen English branch entering Stage 2 is stable and competent, separately from the held-out cross-lingual comparison in Table 2.

Setup. We fine-tune the top $N = 8$ layers of Llama-3-8B-Instruct via Causal Language Modelling on the three balanced English task datasets described in Section 4.1: SQuAD 2.0 (QA), SNLI (NLI), and MMLU (MCQ), for 3 epochs with $\text{lr} = 2e-5$, linear scheduling, and 0.1 warmup. We report end-of-epoch accuracy on MMLU and SNLI, and F1/EM on SQuAD, all computed on their respective in-domain validation splits.

The English anchor stabilises well before Stage 2 begins. All three tasks converge to a consis-

Table 8: Stage 1 (English Multi-Task Fine-tuning) performance of Llama-3-8B-Instruct across training epochs. Scores are computed on the in-domain validation splits of MMLU, SNLI, and SQuAD 2.0, and serve as anchor-quality evidence for the frozen English branch used by $\mathcal{L}_{\text{Reg}}$ in Stage 2.

Epoch	MMLU Acc.	SNLI Acc.	SQuAD F1	SQuAD EM
1	60.5%	90.3%	85.0%	81%
2	63.3%	91.9%	86.0%	83%
3	61.2%	92.0%	85.5%	82%

tent performance band within the first epoch and remain stable thereafter: MMLU accuracy fluctuates narrowly around ~ 61 – 63% , SNLI accuracy rises monotonically to 92.0% , and SQuAD F1/EM stay within a one-point band across epochs (85.0 – 86.0 F1, 81 – 83 EM). No task shows the sharp epoch-over-epoch degradation that would indicate an unstable or overfit checkpoint. We select the epoch-2 checkpoint, which peaks jointly on MMLU (63.3%) and SQuAD ($86.0/83$), as the frozen Stage 1 anchor for all Stage 2 experiments reported in Table 2.

Implication for $\mathcal{L}_{\text{Reg}}$. Because $\mathcal{L}_{\text{Reg}}$ penalises deviation of the trainable English branch from this frozen checkpoint (Eq. 8), its effectiveness presupposes that the checkpoint itself is a competent, non-degenerate representation of English task performance. The trajectory in Table 8 confirms this precondition: the anchor is not an undertrained or unstable snapshot but a converged multi-task checkpoint, which is consistent with the strong English preservation we observe throughout Stage 2 (English XNLI 83.41 , English XQuAD $83.43/72.10$, both at or above this Stage 1 baseline).

A.8 Case Study: Visualizing Optimal Transport Plans Across Language Pairs

To qualitatively inspect what the learned transport plans capture, we visualise the Sinkhorn solution T^* for the sentence “The cat sat on the mat” and its translations into Vietnamese, Chinese, and Arabic.

English–Vietnamese. The BOS token aligns exclusively with its Vietnamese counterpart, preserving positional structure. Semantically, *cat* concentrates the majority of its mass on the sub-token pair $m + \text{èo}$, the two-piece tokenisation of *mèo* (cat), demonstrating that the plan correctly reassembles morphologically split targets. *sat* peaks at *ngồi* (sit/sat), *on* at *trên* (on/above), and *mat* at *thảm* (mat/rug), with *The* transporting to *Con*, the Vietnamese noun classifier. The overall diagonal tendency indicates well-ordered, semantically faithful alignment despite the absence of explicit word-level supervision.

English–Chinese. The Chinese plan is the sharpest of the three (lowest OT loss = 0.449), reflecting the compact, morpheme-dense tokenisation of Chinese. *cat* concentrates 0.106 of its mass almost exclusively on 猫 (cat), the highest single-cell value outside BOS across all three language pairs. *sat* and *on* both peak at 坐在 (sat/on), a compound token that jointly encodes the verb and its locative particle—a many-to-one alignment that the transport plan accommodates naturally by distributing mass from two English tokens to a single Chinese one. *The* spreads its mass more diffusely, consistent with the absence of a direct Chinese article.

English–Arabic. Arabic yields the highest OT loss (0.573), consistent with its rich morphology, right-to-left script, and pro-drop syntax, all of which increase the surface distance between the two distributions. Mass is notably more diffuse: no cell outside BOS exceeds 0.072, compared to 0.099 and 0.106 for Vietnamese and Chinese respectively. Nevertheless, semantic coherence persists: *The* peaks at the Arabic definite-article clitic, *cat* concentrates on the Arabic token for *qitt*, and *mat* aligns most strongly with the final token encoding the mat noun. The higher loss and increased diffuseness directly corroborate the quantitative finding in Table 2 that Arabic remains among the harder transfer targets.

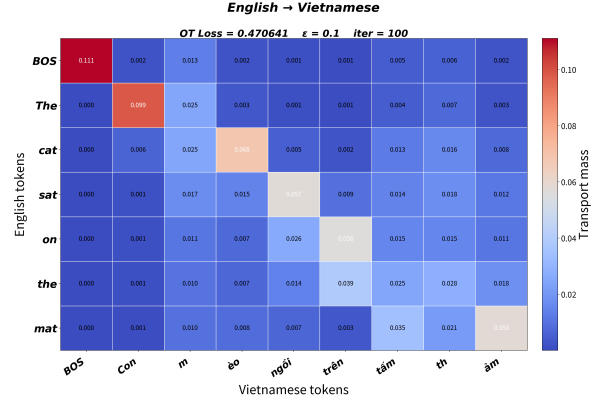


Figure 9: Optimal transport plan for English → Vietnamese (“The cat sat on the mat” → “Con mèo ngồi trên thảm”). OT Loss = 0.471, $\epsilon=0.1$, $K=100$.

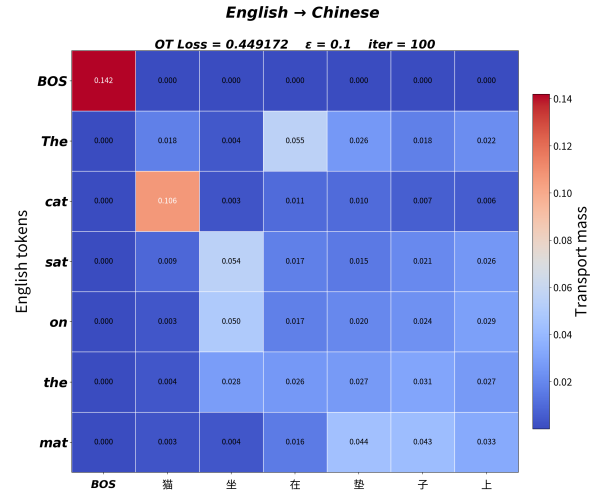


Figure 10: Optimal transport plan for English → Chinese (“The cat sat on the mat” → “猫坐在垫子上”). OT Loss = 0.449, $\epsilon=0.1$, $K=100$.

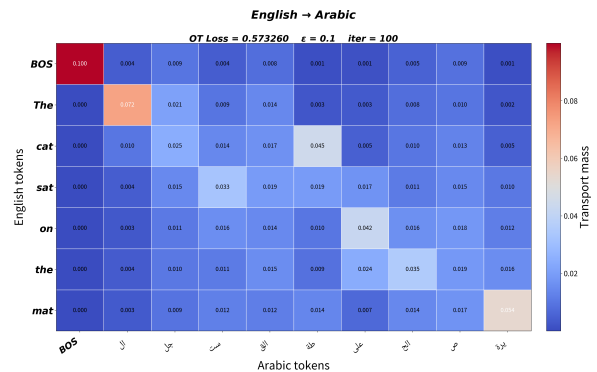


Figure 11: Optimal transport plan for English → Arabic (“The cat sat on the mat” → “al-qitta jalsat ‘alā al-ḥaṣīra”). OT Loss = 0.573, $\epsilon=0.1$, $K=100$.

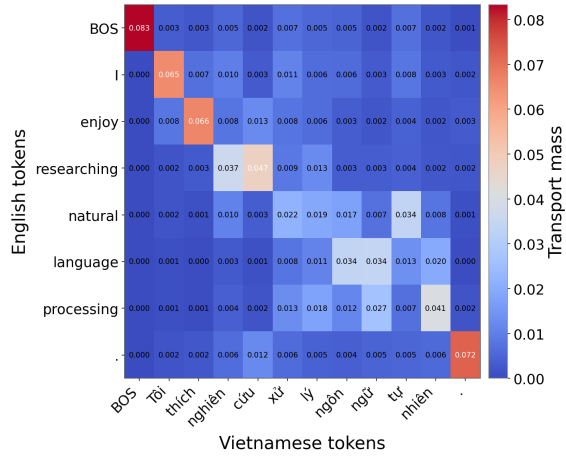


Figure 12: Optimal transport plan for Vietnamese $\rightarrow$ English (“Tôi thích nghiên cứu xử lý ngôn ngữ tự nhiên” $\rightarrow$ “I enjoy researching natural language processing”). $\varepsilon=0.1$, $K=100$. The plan accommodates token-count asymmetry naturally: the two-piece Vietnamese compound *tự nhiên* jointly transports mass to the single English token *natural*, and similarly *ngôn ngữ* distributes mass to *language*, demonstrating that the entropy-regularised transport plan captures semantic correspondence without requiring one-to-one token alignment.

A.9 Case Study: Epsilon-Iteration Sinkhorn Parameters Visualization

Figure 12 visualises the plan for the sentence pair “Tôi thích nghiên cứu xử lý ngôn ngữ tự nhiên” $\rightarrow$ “I enjoy researching natural language processing” at $\varepsilon=0.1$, $K=100$. The plan reveals two characteristic behaviours. First, it handles token-count asymmetry gracefully: the compound *tự nhiên* (two Vietnamese tokens) distributes mass collectively onto the single English token *natural*, and *ngôn ngữ* (two tokens) concentrates on *language*—many-to-one alignments that the continuous transport plan accommodates without any explicit word-alignment supervision. Second, the near-diagonal structure confirms that positional and semantic ordering is largely preserved across the two languages despite their typological differences.

Figure 13 sweeps $\varepsilon \in \{0.05, 0.1, 0.2, 0.4\}$ and $K \in \{50, 100, 200, 400\}$ on the same sentence pair. Across columns, decreasing ε sharpens the plan toward a near-permutation matrix, concentrating mass on the closest token pairs but at the cost of reduced flexibility for diffuse multi-token correspondences. Increasing ε spreads mass uniformly, eventually approaching the trivial doubly-stochastic matrix with no discriminative

Design axis	Alqahtani et al. (2021)	Ours
Backbone	mBERT (enc.)	Causal LLM
Objective	$\mathcal{L}_{OT}, \mathcal{L}_{L2}$	$\mathcal{L}_{LM}, \mathcal{L}_{OT}, \mathcal{L}_{Reg}$
Layer target	Last layer only	Mid 8–24
Anchor scheme	Reg. vs. orig. pre-Frozen FT reps	per-step branch
Fine-tuning	Full-model	LoRA (PEFT)
Training regime	Single-stage	Two-stage
<i>Component checklist</i>		
Causal LM loss $\mathcal{L}_{LM}$	✗	✓
Middle-layer targeting	✗	✓
Learnable layer weights	✗	✓
Parameter-efficient	✗	✓
Two-stage pipeline	✗	✓
Span QA eval.	✓	✓
Knowledge-intensive (MCQ) eval.	✗	✓
Anchor/target consistency reg.	✓	✓
Sinkhorn-based OT	✓	✓

Table 9: Comparison of our asymmetric OT framework with Alqahtani et al. (2021). Both use entropy-regularised Sinkhorn OT with a consistency term that penalises drift of the anchor/target representations, and both evaluate on XNLI and XQuAD. Our method targets autoregressive decoders: it adds a language-modelling loss, aligns learnable-weighted middle layers rather than only the final layer, uses parameter-efficient LoRA adapters instead of full-model fine-tuning, and protects the anchor at every training step against shared-parameter drift via a two-stage pipeline. [†] Alqahtani et al. (2021) evaluate XQuAD via an extractive span-prediction head on a bidirectional encoder; our XQuAD evaluation instead requires the decoder to autoregressively generate the answer span, a substantially harder setting for the same benchmark.

signal. Across rows, $K=50$ produces visibly under-converged plans with residual off-diagonal noise; convergence stabilises by $K=100$ and changes negligibly beyond that, corroborating the efficiency-quality trade-off reported in Table 6 and Section A.5.

A.10 Comparison with Alqahtani et al. (2021)

Table 9 summarises how our framework differs from the OT-based alignment method of Alqahtani et al. (2021), the closest prior work applying Sinkhorn-regularised optimal transport to cross-lingual representation alignment.

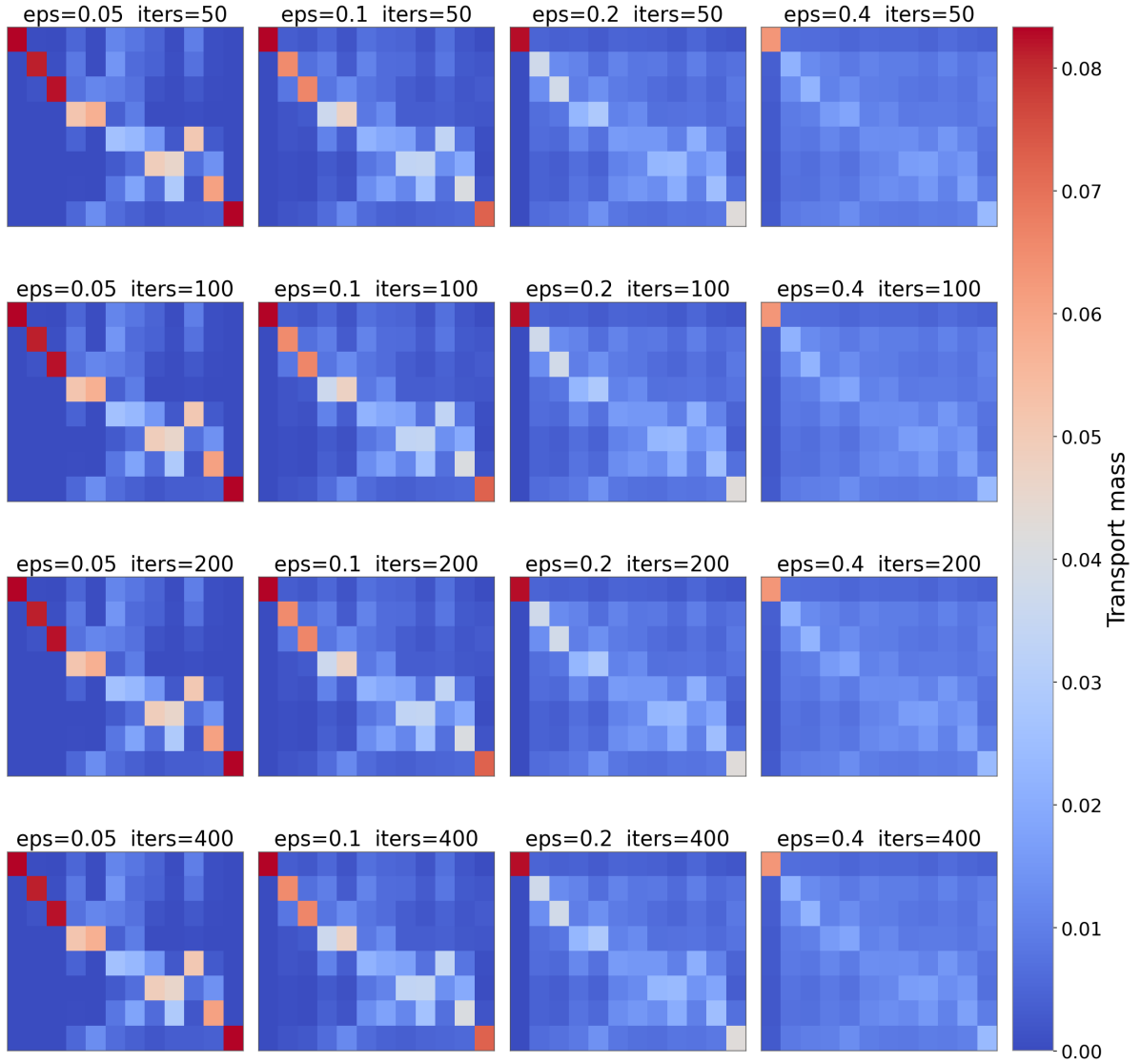


Figure 13: Effect of entropy regularisation $\varepsilon \in \{0.05, 0.1, 0.2, 0.4\}$ (columns) and iteration count $K \in \{50, 100, 200, 400\}$ (rows) on the Sinkhorn transport plan for the Vietnamese–English pair. Smaller ε produces sharper, more diagonal plans that concentrate mass on semantically closest token pairs; larger ε diffuses mass uniformly, weakening the alignment signal. Increasing K improves convergence toward the entropy-regularised optimum, with plans at $K=50$ visibly under-converged relative to $K \geq 100$; beyond $K=100$, the plan geometry stabilises, consistent with the quantitative results in Table 6.