Middle-Layer Representation Alignment for Cross-Lingual Transfer in Fine-Tuned LLMs

Danni Liu **Jan Niehues**
Karlsruhe Institute of Technology, Germany
{danni.liu, jan.niehues}@kit.edu

Abstract

While large language models demonstrate remarkable capabilities at task-specific applications through fine-tuning, extending these benefits across diverse languages is essential for broad accessibility. However, effective cross-lingual transfer is hindered by LLM performance gaps across languages and the scarcity of fine-tuning data in many languages. Through analysis of LLM internal representations from over 1,000+ language pairs, we discover that middle layers exhibit the strongest potential for cross-lingual alignment. Building on this finding, we propose a middle-layer alignment objective integrated into task-specific training. Our experiments on slot filling, machine translation, and structured text generation show consistent improvements in cross-lingual transfer, especially to lower-resource languages. The method is robust to the choice of alignment languages and generalizes to languages unseen during alignment. Furthermore, we show that separately trained alignment modules can be merged with existing task-specific modules, improving cross-lingual capabilities without full re-training. Our code is publicly available[1].

1 Introduction

Decoder-only large language models (LLMs) have emerged as the dominant paradigm in NLP. While these models exhibit promising zero-shot capabilities (Wei et al., 2022; Chowdhery et al., 2023), further task-specific fine-tuning remains crucial for optimal performance in many applications (Shen et al., 2024; Xu et al., 2024; Alves et al., 2024). During fine-tuning, a practical challenge is that the available training data rarely covers all languages supported by LLMs. This highlights the importance of cross-lingual transfer to extend task-specific performance gains across languages.

While cross-lingual transfer has been extensively studied (Wang and Zheng, 2015; Ruder et al., 2019; Artetxe and Schwenk, 2019b; Pfeiffer et al., 2020), achieving it on generative tasks with variable-length outputs remains challenging (Vu et al., 2022; Li and Murray, 2023) compared to classification tasks. This challenge is especially relevant for LLMs, which formulate all tasks as next-token prediction problems.

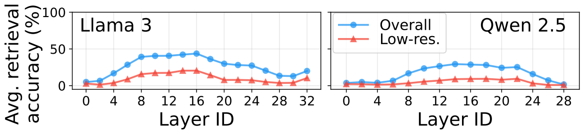
The theoretical foundation of cross-lingual transfer lies in the analogous relationships between concepts across languages. This intuition was first demonstrated in cross-lingual word embeddings (Mikolov et al., 2013; Lample et al., 2018; Xu and Koehn, 2021), where these vector representations exhibit isometric relationships, i.e., the geometric structure of semantically equivalent items is preserved across different languages. This isometry property has proven crucial for transferring learned models across languages (Schuster et al., 2019; Wang et al., 2024b). Subsequent encoder-decoder models (Ha et al., 2016) and decoder-only models (Wu et al., 2024a) also exhibit similar properties in their internal representations.

While pretrained multilingual models naturally develop some degree of unified multilingual representations (Pires et al., 2019; Conneau et al., 2020; Muller et al., 2021), explicitly strengthening the relationships between semantically equivalent content has shown benefits in various downstream tasks: cross-lingual retrieval (Yu et al., 2018), parallel text mining (Schwenk et al., 2021), zero-shot classification (Hu et al., 2021; Gritta and Iacobacci, 2021) and translation (Arivazhagan et al., 2019; Pham et al., 2019; Duquenne et al., 2022). Despite different approaches, these works share a common objective: *aligning* representations of semantically equivalent content across languages while preserving overall expressiveness.

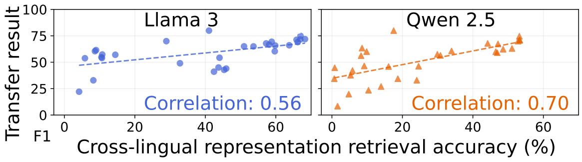
Cross-lingual alignment approaches have been successfully applied to models preceding LLMs. For *encoder-only* models, outputs can be aligned by e.g., minimizing distances between parallel

[1] https://github.com/dannigt/mid-align

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15979–15996
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics



(a) Cross-lingual semantic alignment (measured by average retrieval accuracy over 35 languages and 1190 language directions) varies by layer, with the middle layer showing the highest score. Lower-resource languages are poorly aligned.



(b) Positive correlation between base model cross-lingual semantic alignment and downstream transfer performance.

Figure 1: Two observations (§2) motivating our approach of aligning multilingual representations (§3).

sentence representations (Feng et al., 2022) or cross-lingual masked language modeling objectives (Conneau and Lample, 2019). These techniques are largely applicable to *encoder-decoder* models, where alignment is typically enforced to the encoder outputs (Duquenne et al., 2023). In contrast, *decoder-only* models lack such clear separation between input processing and output generation. This makes it less obvious where and how to optimize for cross-lingual alignment, as also highlighted in the survey by Hämmerl et al. (2024).

In this work, we start by quantifying the degree of cross-lingual alignment present in two prominent LLMs, Llama 3 (AI @ Meta et al., 2024) and Qwen 2.5 (Qwen Team et al., 2025). We then apply these insights to improve cross-lingual transfer in task-specific fine-tuning. By alternatively training on alignment and task-specific data, we aim to improve the cross-lingual generalization to languages without fine-tuning data. We demonstrate transfer improvements across diverse tasks: slot filling, machine translation, and structured text generation. Our main findings include:

- Applying alignment objectives to middle layers during LLM task-specific fine-tuning improves cross-lingual transfer (§5.1) and enhances alignment across all network depths (§5.2).
- The transfer improvements extend beyond those languages seen in alignment (§5.1).
- Our approach is robust to the choice of languages used for alignment training (§6.2, 6.3).
- Task-specific and alignment modules trained separately can be combined post-hoc to improve transfer performance (§6.4).

2 Analyzing Cross-Lingual Alignment

To understand how well LLM representations capture semantic equivalence across languages, we use translation retrieval as a diagnostic task. We choose this retrieval task over other metrics like cosine similarity or SVCCA score (Raghu et al., 2017) because it better captures *relative* semantic relationships. That is, if a model's representations enable us to identify a sentence's translation from a set of candidates, the exact numerical distance between the query and the retrieved translation is less important than the ability to rank translations as the most semantically similar.

Specifically, we first extract model activations at each network layer for all language variants of the input text. To handle variable-length sequences, we create fixed-size sentence embeddings by mean-pooling the activations over the sequence length dimension. For translation retrieval, given a query sentence in one language, we compare its embedding to the embeddings of candidate sentences in the target language using ratio-based margin similarity (Artetxe and Schwenk, 2019a)[2]. For N languages, we evaluate retrieval accuracy across all $N(N-1)$ possible language pairs. We use the FLoRes-200 dataset (NLLB Team, 2024), which provides high-quality multiway parallel texts across diverse languages (detailed setup in §4.2).

Our investigation of LLama 3 and and Qwen 2.5 models[3] reveals three key findings:

Overall weak semantic alignment, with peak in middle layers: As shown in Figure 1a, the average translation retrieval accuracy across 1,190 language pairs remains below 50%, with Llama 3 outperforming Qwen 2.5. Low-resource languages[4] show especially weak alignment, achieving less than half of the overall average accuracy. In particular, the *middle* layers of both models demonstrate the strongest retrieval performance. This suggests stronger potential for cross-lingual transfer at these intermediate representations.

Strong correlation between base LLM semantic alignment and downstream task transfer: To what extent can the semantic alignment present in the base LLM predict cross-lingual transfer performance after supervised fine-tuning? Using multilingual slot filling as a case study, we train models

[2] shown to outperform cosine similarity for cross-lingual retrieval tasks (Artetxe and Schwenk, 2019a)

[3] specifically the `8B-Instruct` and `7B-Instruct` variants

[4] resource levels as defined by NLLB Team (2024)

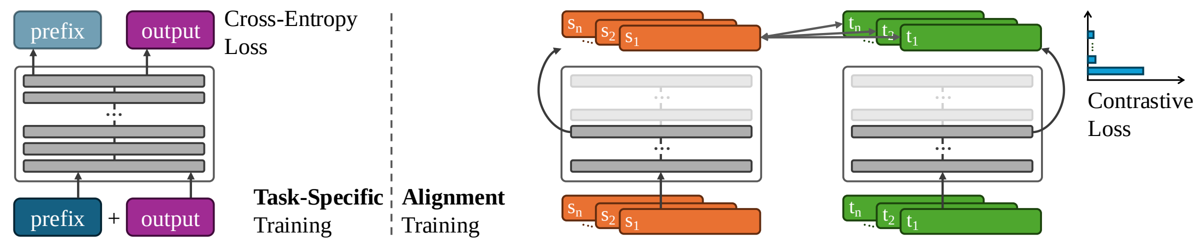


Figure 2: Illustration of our approach, alternating training between task-specific (left) and alignment (right) objectives. The alignment objective operates on middle-layer representations.

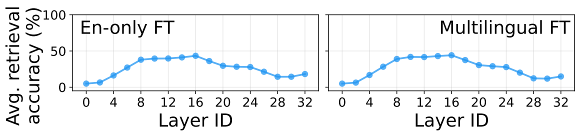


Figure 3: Task-specific fine-tuning shows minimal impact on semantic alignment.

on 5 high-resource languages jointly and evaluate transfer performance on 25 additional languages (detailed setup in §4.1). As shown in Figure 1b, for both Llama 3 and Qwen 2.5, we observe strong positive correlations ($p < 0.01$) between middle-layer retrieval accuracy and downstream task performance. This correlation suggests that increasing cross-lingual alignment in LLM intermediate representations may improve cross-lingual transfer.

Task-specific fine-tuning preserves but does not enhance semantic alignment: After analyzing the base LLMs, we examine how supervised fine-tuning affects the models' internal semantic alignment. Using the same multilingual slot filling task as before, we study both English-only and multilingual fine-tuning. Despite multilingual fine-tuning being an established method for improving cross-lingual transfer (Li and Murray, 2023; Chirkova and Nikoulina, 2024), we observe that neither training configuration alters the models' cross-lingual semantic alignment (Figure 3). This preservation of baseline alignment patterns, even under multilingual training, indicates that pure fine-tuning does not sufficiently strengthen cross-lingual alignment. This further motivates us towards explicit cross-lingual alignment during fine-tuning.

3 Explicit Alignment in fine-tuning

Alternate Training As shown in Figure 2, we optimize either the task-specific objective or the alignment objective in each training step. Compared to joint optimization that computes a combined loss for both objectives and performs a single backward pass, this approach does not involve manually balancing objective weights and mitigates potential gradient conflicts between objectives. It also showed stronger task performance empirically.

Task Objective We follow standard causal language modeling, using a cross-entropy loss over the predicted text conditioned on the input prefix.

Alignment Objective We use a contrastive loss motivated by its successful applications in sentence embedding (Feng et al., 2022), dense retrieval (Karpukhin et al., 2020) and modality alignment (Ye et al., 2022; Girdhar et al., 2023). The loss maximizes the similarity between translations while minimizing similarity between non-translations. Given a batch $\mathcal{B}$ of n pairs of parallel sentences, the alignment loss for a sentence pair (s, t) is:

$$\mathcal{L}_{\text{align}} = -\log \frac{\exp(\text{sim}(\mathbf{h}_s^i, \mathbf{h}_t^i))}{\sum_{v \in \mathcal{B}} \exp(\text{sim}(\mathbf{h}_s^i, \mathbf{h}_v^i))} \quad (1)$$

where $\mathbf{h}_s^i$ is the mean-pooled[5] hidden states at the i^{th} LLM layer for input s and $\text{sim}(\cdot, \cdot)$ is a similarity function. Motivated our finding that middle layers have the strongest cross-lingual alignment potential, we select i as the middle layer and compare its performance to other layer positions. We use cosine similarity following prior works (Gao et al., 2021; Ye et al., 2022). The similarity score is optionally scaled by a temperature parameter τ, which controls the peakiness of the softmax distribution and in turn determines the relative importance of non-translation pairs. This temperature parameter is tuned on the development sets.

Activating Individual Objectives Note that the task and alignment losses can be activated separately. Deactivating the alignment loss degenerates to standard task-only training. Conversely, deactivating the task loss trains the model only for alignment. The modularity allows combining separately-trained task and alignment models.

Data Requirement Our approach requires minimal parallel data. Later experiments show that

[5]Initial experiments with attention pooling degraded performance. We also tried a stop-gradient operator on English representations to align non-English representations towards English, but it did not give consistent gains.

	Dataset	Languages
Slot Filling		
Task - train	MASSIVE	{ar, en, es, ru, zh}
Task - test	MASSIVE	supervised + {af, az, cy, de, el, fr, hi, is, ja, jv, sw, th, tl, tr, ur}
Alignment	Tatoeba	low-res.: {cy, jv, jp, sw, tl}-en mid-res.: {el, hi, th, tr}-en high-res.: {ar, es, ru, zh}-en
Machine Translation		
Task - train	ALMA	{cs, de, is, ru, zh} ↔ en
Task - test	WMT 23	supervised + {he, ja, uk} ↔ en
Alignment		(same as "Task - train")
JSON Generation (challenge task)		
Task - train	UNER	{en, pt, zh}
Task - test	UNER	supervised + {da, hr, sk, sr, sv}
Alignment	Tatoeba	{da, sv}-en
Semantic Alignment Evaluation (diagnostic task)		
Alignment	FLoRes-200	$N(N-1)$ pairs for N lang.

Table 1: Dataset statistics for three downstream tasks and one diagnostic task. "Train" refers to languages involved in SFT, and "test" includes SFT languages and additional transfer languages unseen during training. See Appendix B for more details.

for lower-resource languages, a few hundreds of sentences of parallel data is sufficient to improve transfer. Our approach also offers a practical advantage over alternatives that require monolingual language modeling training for each transfer target language (Ansell et al., 2022; Vu et al., 2022; Chronopoulou et al., 2024).

4 Experimental Setup

4.1 Data

In general, we fine-tune on several high-resource languages and then evaluate transfer performance on additional languages. We do not focus on English-only fine-tuning, since our initial experiments demonstrated that multilingual fine-tuning substantially outperforms English-only fine-tuning[6], thus establishing it as a stronger baseline. Table 1 presents a dataset overview. Descriptions of the language codes are in Appendix C.

Main Task Data: We evaluate our approach on slot filling and machine translation, both modeled as generative tasks with templates shown in Appendix D.2. For slot filling, we use the MASSIVE dataset (FitzGerald et al., 2023). We train on 5 high-resource languages, and evaluate transfer performance on 15 additional diverse languages, 5 of which have non-Latin writing systems. This task presents a challenge due to the 60 possible slots, requiring strictly following the output format for correct parsing. For machine translation, we use ALMA (Xu et al., 2024)'s training and test data, and additionally test on 6 zero-shot directions from WMT 23 (Kocmi et al., 2023).

Challenge Task Data: To assess performance on long-sequence processing and structured text generation, we include JSON generation as a challenge task. We use the UNER dataset (Mayhew et al., 2024) from the Aya collection (Singh et al., 2024), which requires following example instructions and extracting named entities into JSON format. A challenge not present in the previous tasks is the longer inputs, with an average input length exceeding 150 tokens in English. For this task, we train on 3 high-resource languages (en, pt, zh) and transfer to the 5 remaining languages.

Alignment Data: For alignment, we mainly use parallel data to English from Tatoeba (Tiedemann, 2020), except for machine translation, where the training sentences are inherently parallel. For slot filling, our main experiments align the five languages with the weakest baseline[7] transfer performance (cy, jv, jp, sw, tl) reported by the dataset creators (FitzGerald et al., 2023). We choose them because their weak baseline performance suggests a lack of effective transfer, providing a strong testbed for evaluating the potential benefits of our alignment approach. For ablation, we alter the following factors of the alignment data:

- Resource level (low, medium, high-resource)
- Language coverage
- Domain (oracle data, different, very distant)

For machine translation, given the inherent semantic equivalence of translation pairs, we directly leverage the translation data for alignment. For JSON generation, we align the two lowest-resourced in UNER (da and sv)[8] to English. For lower-resource languages, the alignment data are a few hundreds as detailed in Appendix B.

4.2 Evaluation

Semantic Alignment Evaluation: As described in §2, we evaluate cross-lingual semantic alignment by retrieval accuracy. Given N languages, we perform many-to-many retrieval and average the

[6]These English-only FT results are in Appendix A.

[7]their baseline is an XLM-R model trained on English

[8]While Serbian (sr) is also low-resourced in UNER, we exclude it from alignment due to data quality. Running language identification reveals that many sentences in the Serbian alignment data are not actually in Serbian.

accuracy over the $N(N-1)$ language pairs. For the initial analyses (§2), the 35 languages are listed in Appendix C. We use the FLoRes-200 (NLLB Team, 2024) development set with 997 parallel sentences. While FLoRes partially overlaps with ALMA's training data, it remains the only reliable massively multilingual multiway corpus to the best of our knowledge. Alternative such as Tatoeba have been advised against due to data imbalance and noise (Heffernan et al., 2022; Janeiro et al., 2024). We also demonstrate that this overlap does not result in memorization effects (§6.3). When reporting an aggregated retrieval accuracy for a model, we average over all language pairs at even-numbered layers' retrieval accuracy, excluding the input embedding layer.

Task Performance Evaluation: For slot filling, we report F_1 scores using the original evaluation script by FitzGerald et al. (2023). For machine translation, we report BLEU[9] (Papineni et al., 2002) and COMET-22 (Rei et al., 2022) scores. For JSON generation, we parse the generated outputs back to named entity tuples and then evaluate F_1 scores.

4.3 Model, Training, and Inference

We build upon Llama (AI @ Meta et al., 2024) and Qwen (Qwen Team et al., 2025), specifically `Meta-Llama-3-8B-Instruct`[10] and `Qwen2.5-7B-Instruct`. We use LoRA (Hu et al., 2022) adapters with a rank of 8 for all attention components and linear projections. The effective batch size is 128 for both objectives, with mini-batches of 32 examples considered for the contrastive objective[11]. Alignment data from different languages are re-sampled to an approximately uniform distribution. More details are in Appendix D.

5 Main Results

The main results are summarized in Table 2. Before assessing our proposed approach, we first establish the necessity of supervised FT by comparing it with zero-shot usage of the LLMs (rows $(2,5)$ vs. $(1,4)$). On slot filling, the zero-shot performance of Llama 3 is very poor, achieving only 6.6% F_1 on English due to difficulties in adhering to task-specific formats. We therefore do not evaluate its zero-shot performance on all languages. In machine translation, supervised fine-tuning shows substantial gains of 4-6 COMET over zero-shot.

5.1 Overall Performance Comparison

Gains in cross-lingual transfer with supervised performance preserved: Our approach improves cross-lingual transfer across different tasks and models. For slot filling, we observe gains in both supervised and transfer (F_1 $+0.4$ and $+1.5$ respectively) settings on Llama fine-tuning, with similar improvements on Qwen (F_1 $+0.7$ supervised, $+1.8$ transfer). In machine translation with Llama in row (3), our approach brings substantial gains when transferring to out-of-English directions ($+1.5$ BLEU, $+1.1$ COMET).[12] For into-English directions, there is a modest improvement in $+0.5$ BLEU and $+0.2$ COMET. The larger gains on out-of-English directions suggest the approach is more beneficial for non-English generation in this case. For Qwen in row (6), our approach shows minor gains in into-English translation ($+1.1$ BLEU but no change in COMET), and does not influence out-of-English scores. It also leads to a degradation (-0.8 COMET) on supervised directions. This is potentially due to Qwen's non-English-centric pretraining combined with our English-centric alignment data. With this exception, our approach maintains or improves supervised performance while enhancing transfer.

Aligned languages improve the most, but gains extend to other languages: The diverse language coverage in the slot filling dataset allows us to compare how the alignment objective benefits transfer to both aligned and non-aligned languages. While aligned languages show the strongest improvements (F_1 $+4.2$ and $+4.9$ for Llama and Qwen respectively), the benefits extend to other languages. Over the remaining 10 non-aligned languages, there is an average F_1 improvement of 0.4 (per-language results in Appendix E). This suggest that the alignment step enhances the model's general cross-lingual transfer capabilities rather than optimizing for specific language pairs.

Smaller gains on non-Latin script languages: Beyond overall performance improvements, we ob-

[9] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.2 sacreBLEU (Post, 2018) signature, with "tok:ja-mecab-0.996-IPA" for Japanese and "tok:zh" for Chinese.

[10] chosen over more recent versions to limit test set contamination, as its knowledge cutoff (March 2023) predates our translation test set (WMT 23).

[11] While contrastive learning typically benefits from larger batch sizes (Chen et al., 2022), our initial experiments with increased batch sizes did not give consistent improvements.

[12] The observation of alignment not improving supervised directions is in line with Pan et al. (2021), where the purely contrastive learning setup also does not improve supervised scores over their baseline ("m-Transformer").

ID Model	Slot Filling (MASSIVE)				Machine Translation (WMT23)						
	Supervised (5 lang.)	Transfer (15 lang.)	Transfer (aligned)	Retrieval (all 20 lang.)	Supervised (5 lang.↔En)		Transfer (3 lang.→En)		Transfer (En→3 lang.)		Retrieval (all 9 lang.)
	F_1	F_1	F_1	Acc.	BLEU	COMET	BLEU	COMET	BLEU	COMET	Acc.
(1) LLAMA 3	–	–	–	39.1	25.8	75.5	27.8	75.8	14.8	71.3	51.5
(2) + SFT	76.6	60.2	51.7	39.4	30.0	81.5	31.8	82.8	15.5	79.6	(55.3)
(3) + alignment	**77.0**	**61.7**	**55.5**	73.2	29.9	81.5	**32.3**	83.0	**17.0**	**80.7**	(84.5)
(4) QWEN 2.5	–	–	–	21.4	23.0	74.5	28.5	81.3	12.6	71.2	36.5
(5) + SFT	76.3	53.5	41.6	20.9	27.4	78.4	29.7	82.7	14.6	76.9	(38.8)
(6) + alignment	**77.0**	**55.3**	**46.5**	20.5	27.2	77.6	**30.8**	82.7	14.7	76.9	(75.6)

Table 2: Overall supervised and transfer results. Retrieval accuracy are averaged over all language pairs and layers. **Bold**: highest task scores which outperforms the other setups. (Results in brackets): potentially inflated scores due to partial overlap between retrieval and translation data. Language-specific results are in Appendix E.

serve smaller gains on languages with diverse writing systems. Specifically, for the non-Latin script transfer languages in the slot filling task (Greek, Hindi, Japanese, Thai, Urdu), the average improvement is only 0.5 F_1 in contrast to the overall average gain of 1.5. This reduced gain is likely related to suboptimal tokenization for these languages in multilingual models (Rust et al., 2021; Petrov et al., 2023; Hong et al., 2024). When tokens poorly align with linguistic units, the mean-pooled sentence representations may poorly capture semantics, thereby impacting our alignment objective.

	Supervised↑	Transfer↑	Transfer SD↓
Slot filling (MASSIVE): F_1			
SFT baseline	76.6	60.2	–
Middle (layer 16)	77.0	61.7	2.6
Top (layer 32)	76.6	62.0	3.3
Bottom (layer 8)	76.8	58.0	2.9
Machine translation (WMT23): COMET			
SFT baseline	81.5	79.6	–
Middle (layer 16)	81.5	80.7	3.7
Top (layer 32)	82.0	80.2	4.2
Bottom (layer 8)	81.2	80.1	5.6

Table 3: Impact of alignment loss placement on Llama 3. Last column: standard deviation of gains on transfer languages compared to baseline. "Top" leads to more uneven gains across languages, while "bottom" degrades both supervised and transfer performance.

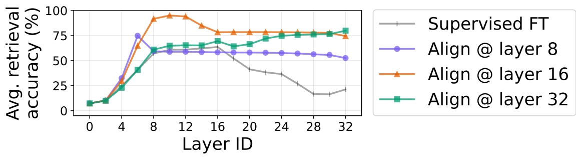


Figure 4: Retrieval accuracy over model depths when aligning different layers of Llama 3. Middle layer placement leads to overall better alignment.

5.2 Alignment Loss Placement

To validate our choice of middle-layer alignment motivated by the analysis in §2, we compare performance when applying the alignment loss at different network depths: bottom (8^{th}), middle (16^{th}), and top (32^{nd}) layers of Llama.

Middle layer placement achieves more balanced improvements in transfer languages: As shown in Table 3, compared to the "middle" configuration, the "bottom" configuration clearly leads to poor overall performance in both supervised and transfer settings, with a particularly strong degradation on the slot filling task. While top-layer alignment maintains overall strong performance, it shows more unbalanced gains across transfer languages, as evidenced by the higher standard deviation of performance changes on transfer languages.

Middle layer placement achieves better alignment across network depths: To better understand the effects of different loss placements, we run the translation retrieval task over model activations at from different intermediate layers. As shown in Figure 4, When the alignment loss is applied at the middle (16^{th}) layer, semantic alignment is enhanced not only at that layer but also in multiple preceding layers. In contrast, top-layer alignment primarily affects only the final layer, and bottom-layer alignment shows limited improvement in alignment quality across all layers. This is likely because the lower layers are occupied with processing more fundamental text features (Belinkov et al., 2017; Peters et al., 2018) rather than abstract semantic meanings.

Aligning several layers does not show consistent gains: Results in Table 3 suggest that aligning at multiple layers may be complementary. In Appendix F, we show that adding alignment losses at both middle and top layers brings further improvements on slot filling, but does not on machine

Alignment Setup	Avg supervised F1↑ (5 lang.)	Avg transfer F1↑ (15 lang.)
5 lang↔en (Table 2 row 3)	77.0	61.7
All↔English (38 pairs)	77.5	63.6 (+1.9)
All↔all (238 pairs)	77.7	63.6 (+1.9)

Table 4: Effects of larger-scale alignment configurations on slot filling. Aligning all non-English languages to English (+1.9 F1) outperforms our base configuration, while fully connecting all 20 languages offers no additional gain beyond English-only alignment.

Alignment Language Resource Level	Super. (5 lang.)	Transfer (15 lang.)	Gain on Aligned (4/5 lang.)
SFT (row (2) Table 2)	76.6	60.2	–
Low (row (3) Table 2)	77.0	61.7	+3.8
Medium	77.8	61.4	+1.1
High	77.6	60.4	+0.7

Table 5: Effect of alignment language resource levels on slot filling F1↑. In three groups of alignment languages: low ({cy, jv, jp, sw, tl}), medium ({el, hi, th, tr}), and high-resource ({ar, es, ru, zh}), languages involved in alignment consistently show improvements, with the strongest gains (+3.8 F1) in the low-resource group.

translation. This task-dependent behavior indicates that how to best align multiple layers still requires further investigation.

5.3 Impact on Representation Retrieval

To assess the impact of the alignment loss on the learned model representations, we also report the retrieval accuracy for all languages involved in each task (20 for slot filling and 9 for machine translation) after fine-tuning in Table 2. For Llama on the slot filling task, the alignment loss substantially improves retrieval accuracy from 39.4% to 73.2%. For Qwen, the alignment loss does not improve retrieval among the 20 slot filling languages, possibly due to the lower accuracy of the base model with many low-resource languages with 0% accuracy, making improvement more challenging. For machine translation, as noted earlier §4.2, the retrieval test data overlaps with part of the task training data, potentially inflating accuracy (marked in brackets in Table 2). However, we verify that this overlap does *not* lead to perfect retrieval accuracy: Specifically, at the 16th layer where the alignment loss is applied, English-source retrieval accuracies for supervised languages show varying accuracy: cs (98.1%), de (96.5%), is (66.9%), ru (90.6%), and zh (94.8%). This suggests that the overlap does not make the retrieval diagnostic task trivial.

6 Analyses

6.1 Larger-Scale Alignment

While our main configuration for slot filling (Table 1) allows studying performance on languages not involved in alignment, we also explore larger-scale alignment scenarios as oracle setups where all languages have parallel data. We conduct additional experiments on slot filling with two expanded configurations:

- All 19 non-English languages aligned to English (38 directional pairs)
- All 20 languages aligned to each other (238 pairs with alignment data from all 380 possible pairs).

The results in Table 4 show that expanding alignment to all languages further improves transfer performance (F1 +1.9) in an oracle setup where every transfer language has alignment data. However, multiway alignment data does not further improve transfer, suggesting that aligning to English implicitly creates multiway alignment effects.

6.2 Resource Level of Alignment Languages

In our main experiments, we selected the 5 languages with the weakest performance from the MASSIVE baseline (FitzGerald et al., 2023) for alignment. We now vary the resource level of the alignment languages using a medium-resource group with {el, hi, th, tr}−en and a high-resource group with {ar, es, ru, zh}−en, which also have supervised task training data. As shown in Table 5, all three configurations improve F_1 scores for the languages involved in alignment. However, the low-resource group exhibit the largest gains (+3.8 F_1), indicating that our approach is most beneficial to languages with weaker initial performance. Moreover, overall transfer gains relative to the SFT baseline diminish when using high-resource languages for alignment, likely because these languages already have well-aligned representations and aligning them provides little benefit to lower-resource languages in the transfer set. Overall, the results show that our approach is robust to the choice of alignment languages, but selecting initially poorly aligned languages could provide broader benefits across different languages.

6.3 Generalization of Learned Alignment

Table 6 examines the language and domain generalization of our alignment component. To isolate the effects of task-specific joint training, we train

Alignment Data	**Overall** (20 lang.)
Multi {ar,es,ru,zh,sw}-en	80.2
Only de-en	71.9
Only es-en	72.9
Only zh-en	72.7
de-en FLoRes (oracle)	77.7
Tatoeba (different)	71.9
IWSLT (very distant)	68.5

Table 6: Alignment generalization across languages and domains. *Upper*: Multilingual training improves overall alignment. *Lower*: Impacts of alignment transfer reasonably across domains, with performance drops when training data differs from test domain.

Setup	Supervised	Transfer
Slot filling (MASSIVE): $F_1\uparrow$		
SFT (row (2) Table 2)	76.6	60.2
Joint (row (3) Table 2)	77.0 (+0.4)	61.7 (+1.5)
Merge	76.9 (+0.3)	61.3 (+1.1)
Machine translation (WMT23): COMET$\uparrow$		
SFT (row (4) Table 2)	81.5	79.6
Joint (row (5) Table 2)	81.5 (+0.0)	80.7 (+1.1)
Merge	82.0 (+0.5)	80.2 (+0.6)

Table 7: Performance comparison of merged alignment and task modules versus joint training. Post-hoc merging of separately-trained LoRA adapters achieves comparable improvements to joint training.

	Supervised (en, pt, zh)	**Transfer** (da, sv)	**Transfer** (5 lang.)
Llama SFT	83.4	82.1	79.3
+ alignment	82.4	83.1	79.8

Table 8: Results on JSON generation evaluated with F_1, showing modest gains for aligned languages but decreased performance for supervised languages.

the models using only the alignment loss, following the same setup as our previous experiments but without optimizing on task-specific data. We then evaluate retrieval accuracy as described in §4.2.

Language Generalization: While our main experiments align multiple language pairs, we now use single languages for alignment. As shown in Table 6 (upper portion), that single-language alignment training leads to diminished performance compared to multilingual training. Interestingly, we see comparable accuracy drops regardless of which individual language is used for alignment, suggesting that the gains of multilingual alignment come from the diversity of the training data rather than characteristics of individual languages.

Domain Generalization: To isolate the effects of multilinguality, we focus on alignment between a single language pair (English-German). In Table 6 (lower portion), we first establish an oracle setup using models trained on FLoRes data (Wikipedia domain, overlapping with retrieval data). We then compare to two setups where the alignment data come from other domains: Tatoeba (short sentences for language learning; different) and IWSLT 2017 (public speaking transcriptions; very distant). While we observe a decrease in retrieval accuracy compared to the oracle setup, the results suggest that, to enforce alignment into the model, it is not strictly necessary to source alignment data from the same domain as the task-specific data.

6.4 Merging Alignment and Task Modules

Our previous experiments focused on models jointly trained on both task and alignment objectives. However, in practice, it may be necessary to enhance existing task-specific models with cross-lingual capabilities, where joint re-training is infeasible due to computational constraints or unavailability of the original task training data. Inspired by recent advances in model merging (Matena and Raffel, 2022; Ilharco et al., 2023), we explore the feasibility of combining separately-trained task and alignment modules. We merge two sets of trained LoRA adapters by averaging their weights[13]: the alignment module trained in isolation (§6.3), and task-specific modules (rows (2) and (5) in Table 2).

Table 7 shows that this post-hoc merging brings comparable improvements comparable to joint training. Moreover, the improvements are more evenly distributed across languages compared to the larger gains observed on languages used directly in alignment. These results demonstrate that our alignment approach is modular and can be combined with existing task-specific models.

6.5 Long Sequence Processing

We investigate a more challenge task requiring longer input and output generation using UNER (§4.1). As shown in Table 8, while aligned languages still show improvements, the gains are more modest compared to previous experiments, with an F_1 increase of 1.0 on aligned languages and 0.5 across all transfer languages. Moreover, there is an average degradation of 1.0 F_1 on supervised languages, mainly due to the decline in Chinese (-2.2 F_1). We suspect that this is due to our sentence-

[13]We use a weighted average tuned on the development set (details in Appendix D.3)

level alignment objective operates on fixed-length representations, which creates conflicts with processing longer sequences. As Chinese is the only character-based language in the JSON generation dataset, which has roughly twice the number of tokens compared to English of equivalent content, the conflict could be more influential for Chinese.

7 Related Works

Multilingual Capabilities of LLMs LLM performance varies across languages due to imbalanced pre-training data volume. However, even predominantly English-centric models (Touvron et al., 2023) exhibit some degree of multilingual capability (Aycock and Bawden, 2024; Yuan et al., 2024), potentially due to the unintentional ingestion of multilingual data during pretraining (Briakou et al., 2023). Meanwhile, many recent LLMs have expanded their language coverage (AI @ Meta et al., 2024; Qwen Team et al., 2025). Despite these inherent multilingual capabilities, extending them to downstream tasks in low-resource settings (Adelani et al., 2024; Iyer et al., 2024) remains challenging.

Multilingual Representation Alignment Enhancing meaningful cross-lingual relationships between model representations has been a well-studied area in the context of many tasks, including intermediate tasks such as bilingual lexicon induction (Zhang et al., 2017) and sentence embeddings (Feng et al., 2022; Li et al., 2023), as well as more direct applications like information retrieval (Izacard et al., 2022) and translation (Pham et al., 2019).

Multilingual representation alignment can be achieved by various mechanisms, such as similarity losses that push translations toward each other (Pham et al., 2019), contrastive losses (Hadsell et al., 2006) that additionally incorporate non-translation pairs, and adversarial losses (Ganin and Lempitsky, 2015) that remove language-specific signals. The cross-lingual transfer capabilities of these approaches is extensively documented in the literature. In particular, contrastive learning methods have shown promising results (Pan et al., 2021; Chi et al., 2021; Qi et al., 2022). Our contribution is not applying contrastive learning itself, but rather investigating how to effectively align multilingual spaces specifically in decoder-only models.

In the context of LLMs, Wang et al. (2024b) use linear projections learned offline to align non-English representations with English ones during decoding. Our work differs in that our alignment objective is parameterized by the same weights as task-specific fine-tuning, and is directly applicable to multilingual fine-tuning. Wu et al. (2024a) align LLM top-layer representations specifically for the task of semantic textual similarity (STS). Different from this work, they do not consider cross-lingual transfer in downstream tasks or explore intermediate LLM layers for alignment.

LLM Representation Analysis Several recent works have analyzed LLM internal representations with geometric analysis of representation spaces (Razzhigaev et al., 2024; Lee et al., 2024), probing classifiers (Wang et al., 2024a; Li et al., 2025), or logit lens analysis (Wu et al., 2024b). Multiple studies (Wu et al., 2024b; Mao and Yu, 2024; Zhong et al., 2024) reported higher representational similarity in middle layers in various evaluation settings, complementing our findings. Wu et al. (2024b) identify "semantic hubs" in LLM middle layers that integrate information from various data types, while we focus specifically on cross-lingual representations rather than multi-modality. Mao and Yu (2024) show that SFT on machine translation increases similarity between parallel sentences from the same MT corpus, while we show that SFT on a non-translation task does *not* increase representation similarity, thereby motivating explicit alignment during SFT. Zhong et al. (2024) measure pairwise similarity to English representations ("latent language") on high resource languages, while we focus on pairwise similarity across a more diverse set of language pairs. Kargaran et al. (2024) use similarity between parallel sentences to estimate cross-lingual transfer capabilities. Our analysis in §2 shares the same motivation, and we additionally show that actively enforcing alignment can improve transfer performance.

8 Conclusion

We presented a simple yet effective approach for enhancing cross-lingual transfer in LLMs through middle-layer representation alignment during fine-tuning. Our experimental results lead to several practical recommendations: 1) Aligning a few weakly-performing languages yields broad transfer benefits. A few hundreds of parallel sentences as alignment data are sufficient. 2) Alignment data can be sourced from different domains as the task. 3) Existing task-specific models can be enhanced with our approach via parameter merging without the need of full re-training.

Limitations

Performance on languages with diverse scripts: As discussed in §5.1, our approach shows smaller gains on languages non-Latin scripts. This limitation is likely related to fundamental tokenization challenges, where suboptimal token segmentation negatively impacts the quality of mean-pooled representations. While our initial experiments on attention pooling did not lead to improvements, exploring more sophisticated pooling mechanisms could potentially address this challenge in future work.

Computational overhead during training: The alternating optimization between task and alignment objectives doubles the computational cost during training compared to standard fine-tuning. In computationally constrained settings, our merging approach (§6.4), which separates task-specific and alignment training, should be prioritized. Given that alignment can be effectively performed using only a small number of parallel sentences (a few hundred per language), this modular approach can significantly reduce the overall computational cost.

Trade-offs between supervised and transfer performance in challenging scenarios: While our approach generally maintains or improves supervised task performance while improving transfer, we observe degradation in supervised performance in two specific scenarios. First, in structured text generation (§6.5), the method shows reduced effectiveness and can impair supervised performance (-1.0 F_1), suggesting that our sentence-level alignment may interfere with the processing of longer, structured sequences. Second, when applying the method to models with weak initial cross-lingual alignment (§5.1), there could be a trade-off between improved transfer and supervised performance.

Acknowledgments

We thank the reviewers for their feedback, as well as Felix Stahlberg and Google Research. Part of this work was funded by the KiKIT (The Pilot Program for Core-Informatics at the KIT) of the Helmholtz Association. The authors gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program. HoreKa is partly funded by the German Research Foundation (DFG).

References

David Ifeoluwa Adelani, A. Seza Doğruöz, André Coneglian, and Atul Kr. Ojha. 2024. Comparing LLM prompting with cross-lingual transfer performance on indigenous and low-resource Brazilian languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 34–41, Mexico City, Mexico. Association for Computational Linguistics.

AI @ Meta, Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, and 543 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Duarte M. Alves, José Pombal, Nuno Miguel Guerreiro, Pedro Henrique Martins, João Alves, M. Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *CoRR*, abs/2402.17733.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *Preprint*, arXiv:1903.07091.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Seth Aycock and Rachel Bawden. 2024. Topic-guided example selection for domain adaptation in LLM-based machine translation. In *Proceedings of the*

18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 175–195, St. Julian's, Malta. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, Son Tran, Belinda Zeng, and Trishul Chilimbi. 2022. Why do we need large batchsizes in contrastive learning? A gradient-bias perspective. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Key ingredients for effective zero-shot cross-lingual knowledge transfer in generative tasks. 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 7222–7238, Mexico City, Mexico. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Alexandra Chronopoulou, Jonas Pfeiffer, Joshua Maynez, Xinyi Wang, Sebastian Ruder, and Priyanka Agrawal. 2024. Language and task arithmetic with parameter-efficient layers for zero-shot summarization. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 114–126, Miami, Florida, USA. 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 and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Paul-Ambroise Duquenne, Hongyu Gong, Benoît Sagot, and Holger Schwenk. 2022. T-modules: Translation modules for zero-shot cross-modal machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5794–5806, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *CoRR*, abs/2308.11466.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, 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.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In

Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15180–15190. IEEE.

Milan Gritta and Ignacio Iacobacci. 2021. XeroAlign: Zero-shot cross-lingual transformer alignment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 371–381, Online. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *Preprint*, arXiv:1611.04798.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742. IEEE Computer Society.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.

Mutian He and Philip N. Garner. 2023. Can chatgpt detect intent? evaluating large language models for spoken language understanding. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 1109–1113. ISCA.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jimin Hong, Gibbeum Lee, and Jaewoong Cho. 2024. Accelerating multilingual language model for excessively tokenized languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11095–11111, Bangkok, Thailand. Association for Computational Linguistics.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Vivek Iyer, Bhavitvya Malik, Wenhao Zhu, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Exploring very low-resource translation with LLMs: The University of Edinburgh‘s submission to AmericasNLP 2024 translation task. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 209–220, Mexico City, Mexico. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

João Maria Janeiro, Benjamin Piwowarski, Patrick Gallinari, and Loïc Barrault. 2024. Mexma: Token-level objectives improve sentence representations. *Preprint*, arXiv:2409.12737.

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. 2024. Mexa: Multilingual evaluation of english-centric llms via cross-lingual alignment. *Preprint*, arXiv:2410.05873.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Hyunji Lee, Danni Liu, Supriti Sinhamahapatra, and Jan Niehues. 2024. How do multimodal foundation models encode text and speech? an analysis of cross-lingual and cross-modal representations. *Preprint*, arXiv:2411.17666.

Daoyang Li, Haiyan Zhao, Qingcheng Zeng, and Mengnan Du. 2025. Exploring multilingual probing in large language models: A cross-language analysis. *Preprint*, arXiv:2409.14459.

Tianjian Li and Kenton Murray. 2023. Why does zero-shot cross-lingual generation fail? an explanation and a solution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12461–12476, Toronto, Canada. Association for Computational Linguistics.

Ziheng Li, Shaohan Huang, Zihan Zhang, Zhi-Hong Deng, Qiang Lou, Haizhen Huang, Jian Jiao, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. Dual-alignment pre-training for cross-lingual sentence embedding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3466–3478, Toronto, Canada. Association for Computational Linguistics.

Zhuoyuan Mao and Yen Yu. 2024. Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 1–25, Bangkok, Thailand. Association for Computational Linguistics.

Michael Matena and Colin Raffel. 2022. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nat.*, 630(8018):841–846.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6076–6085.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. 2024. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 868–874, St. Julian's, Malta. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Junhong Shen, Neil A. Tenenholtz, James Brian Hall, David Alvarez-Melis, and Nicolò Fusi. 2024. Tag-llm: Repurposing general-purpose llms for specialized domains. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dong Wang and Thomas Fang Zheng. 2015. Transfer learning for speech and language processing. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2015, Hong Kong, December 16-19, 2015*, pages 1225–1237. IEEE.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024a. 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.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2024b. Bridging the language gaps in large language models with inference-time cross-lingual intervention. *Preprint*, arXiv:2410.12462.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Di Wu, Yibin Lei, Andrew Yates, and Christof Monz. 2024a. Representational isomorphism and alignment of multilingual large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14074–14085, Miami, Florida, USA. Association for Computational Linguistics.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2024b. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. *Preprint*, arXiv:2411.04986.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Haoran Xu and Philipp Koehn. 2021. Cross-lingual bert contextual embedding space mapping with isotropic and isometric conditions. *Preprint*, arXiv:2107.09186.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113, Seattle, United States. Association for Computational Linguistics.

Katherine Yu, Haoran Li, and Barlas Oguz. 2018. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of the Third Workshop on Representation Learning for NLP*, pages 175–179, Melbourne, Australia. Association for Computational Linguistics.

Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. 2024. How vocabulary sharing facilitates multilingualism in LLaMA? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12111–12130, Bangkok, Thailand. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *Preprint*, arXiv:2408.10811.

A English-Only Fine-Tuning Results

Table 9 compares English-only and multilingual fine-tuning on MASSIVE. Multilingual fine-tuning substantially outperforms English-only in cross-lingual transfer performance.

B Dataset Details

All our task training data are retrieved from HuggingFace[14]. The translation test sets are hosted by WMT[15]. The alignment data are sourced from Tatoeba[16] with its default version of `v2021-07-22` at the time of writing. We filter out translations that are empty or include multiple sentences. The lowest-resource alignment languages have a few hundred parallel sentences: Javanese (264), Swahili (371), Welsh (823). The ablation de-en alignment data is from IWSLT 2017[17] (Cettolo et al., 2017).

[14]MASSIVE: `https://huggingface.co/datasets/AmazonScience/massive`
ALMA: `https://huggingface.co/datasets/haoranxu/ALMA-Human-Parallel`
UNER: `https://huggingface.co/datasets/CohereForAI/aya_collection/viewer/templated_uner_llm`

[15]`https://github.com/wmt-conference/wmt23-news-systems/tree/master/txt`

[16]`https://huggingface.co/datasets/Helsinki-NLP/tatoeba`

[17]`https://huggingface.co/datasets/IWSLT/iwslt2017`

	ar	en	es	ru	zh	cy	ja	jv	sw	tl	af	az	de	el	fr	hi	is	th	tr	ur
English-only	59.8	82.5	82.4	65.8	61.6	60.3	39.7	37.8	39.8	57.5	60.3	39.6	71.1	64.8	68.2	62.1	39.2	75.3	52.9	49.9
Multilingual	75.5	81.7	74.5	77.6	73.8	44.0	65.8	41.0	42.8	65.0	66.0	49.0	75.0	69.4	71.9	70.0	45.0	79.9	60.4	57.1

Table 9: Per-languages F_1 results on slot filling of English-only finetuning compared to multilingual fine-tuning on {ar, en, es, ru, zh}. Multilingual fine-tuning shows stronger transfer performance.

Code	**FLoRes Code**	**Full Name**	**Slot Filling**	**Machine Translation**	**JSON Generation**
af	afr_Latn	Afrikaans	✓		
az	azj_Latn	North Azerbaijani	✓		
ar	arb_Arab	Modern Standard Arabic	✓		
cs	ces_Latn	Czech		✓	
cy	cym_Latn	Welsh	✓		
da	dan_Latn	Danish			✓
de	deu_Latn	German	✓	✓	
el	ell_Grek	Greek	✓		
en	eng_Latn	English	✓	✓	✓
es	spa_Latn	Spanish	✓		
fr	fra_Latn	French	✓		
he	heb_Hebr	Hebrew		✓	
hi	hin_Deva	Hindi	✓		
hr	hrv_Latn	Croatian			✓
is	isl_Latn	Icelandic	✓	✓	
ja	jpn_Jpan	Japanese	✓	✓	
jv	jav_Latn	Javanese	✓		
pt	por_Latn	Portuguese			✓
ru	rus_Cyrl	Russian	✓	✓	
sk	slk_Latn	Slovak			✓
sr	srp_Cyrl	Serbian			✓
sv	swe_Latn	Swedish			✓
sw	swh_Latn	Swahili	✓		
th	tha_Thai	Thai	✓		
tl	tgl_Latn	Tagalog	✓		
tr	tur_Latn	Turkish	✓		
uk	ukr_Cyrl	Ukrainian		✓	
ur	urd_Arab	Urdu	✓		
zh	zho_Hans	Chinese (Simplified)	✓	✓	✓

Table 10: List of languages evaluated on different downstream tasks.

C List of Languages

The languages involved in our downstream tasks are listed in Table 10. The 35 languages in the initial analyses in §2 include all languages in slot fill and machine translation. They additionally include the following languages: am (Amharic), bn (Bengali), it (Italian), hu (Hungrian), hy (Armenian), id (Indonesian), kn (Kannada), ka (Georgian), mn (Mongolian), km (Khmer), ko (Korean), and lv (Latvian).

D Training and Inference Details

D.1 Training Hyperparameters

Fine-tuning is performed using LoRA (Hu et al., 2022) adapters with a rank of 8 for all attention components and linear projections (query, key, value, output, gate, up, down). We set LoRA's α parameter to 16 and dropout to 0.1. The number of trainable parameter is 20,971,520 on Llama 3, and 20,185,088 on Qwen 2.5. We train at most 5 epochs on the task data. Training on all our tasks converged before reaching the max number of epochs. The learning rate is set to 5e-4 with inverse square root schedule and warmup up ratio 0.03. We save checkpoints and evaluate every 200 optimization steps, and early stop if the development loss does not improve for 5 consecutive evaluations. For the temperature parameter τ in the contrastive loss, we searched among {0.1, 1.0, 1.5, 2.0} based on development loss on machine translation. For Llama we 0.1, for Qwen we use 1.5.

D.2 Prompt Format

Slot Filling The system prompt is shortened from He and Garner (2023).[18]

- **System**: Given a command from the user, a voice assistant will extract entities essential for carry out the command. Your task is to extract the entities as words from the command if they fall under a predefined list of entity types.
- **User**: wake me up at five am this week
- **Assistant**: time: five am; date: this week
- **User** (de): wecke mich in dieser woche um fünf uhr auf
- **Assistant** (de): date: dieser woche; time: fünf uhr

For **zero-shot slot filling** experiments, we need to specify more requirements in the system prompt with the template also following He and Garner (2023):

Given a command from the user, a voice assistant like Siri or Olly will extract entities from the command that are essential for carry out the the command. For example, for a command about playing a specific song, the name of the song mentioned by the user would be an entity, falling under the type of "song name".

Your task is to extract the entities as words from the command if they fall under any of the types given below according to the following description:

transport_descriptor house_place music_album sport_type playlist_name movie_name song_name place_name radio_name cooking_type weather_descriptor person email_folder business_type audiobook_author transport_type general_frequency meal_type game_name device_type transport_name time_zone joke_type drink_type email_address food_type date relation currency_name ingredient player_setting movie_type definition_word game_type list_name artist_name personal_info audiobook_name timeofday transport_agency media_type podcast_name coffee_type business_name news_topic app_name podcast_descriptor color_type music_genre event_name time change_amount alarm_type order_type music_descriptor

Please give answers like:

1. person: john; contact_field: phone number
2. transport_app: uber; time_of_day: tonight; time: ten pm
3. None
4. music_genre: jazz

etc., each taking a single line. The entity type must be one of the types given above, and the entity must be copied verbatim from the command. There could be zero, one, or multiple entities in a command.

[18] We stayed consistent to the original prompt text, preserving the typographical errors too.

Machine Translation

- **System**: Translate the following sentences from English to German.
- **User**: Police arrest 15 after violent protest outside UK refugee hotel.
- **Assistant**: Polizei verhaftet 15 Menschen nach gewalttätigen Protesten vor einer Flüchtlingsunterkunft in Großbritannien

JSON Generation

- **User**: Please identify all the named entities mentioned in the input sentence provided below. Use only the categories: PER - person, ORG - organization, and LOC - location. Remember, nationalities are neither locations nor organizations, and organizations can represent other groups of people. Pay attention to the provided example. You should only output the results in JSON format, following a similar structure to the example result provided. Example sentence and results: Where in the world is Iguazu? "Results": ["TypeName": "LOC", "Text": "Iguazu", "Start": 22, "End": 28] Considering the input sentence below, what is the output result? Widely considered to be one of the most spectacular waterfalls in the world, the Iguazu Falls on the border of Argentina and Brazil, are a certainly must see attraction in the area.
- **Assistant**: "Results": ["TypeName": "LOC", "Text": "Iguazu Falls", "Start": 81, "End": 93 , "TypeName": "LOC", "Text": "Argentina", "Start": 111, "End": 120 , "TypeName": "LOC", "Text": "Brazil", "Start": 125, "End": 131]

D.3 Inference Details

We use greedy decoding in all experiments for easily reproducible results. For the model merging experiments, we searched among weights {0.5, 0.7, 0.9} for the task-specific LoRA modules on the MASSIVE development set and chose 0.9 for our experiments.

	Supervised					Transfer (aligned)					Transfer (other)									
	ar	en	es	ru	zh	cy	ja	jv	sw	tl	af	az	de	el	fr	hi	is	th	tr	ur
Llama 3 SFT	75.5	81.7	74.5	77.6	73.8	44.0	65.8	41.0	42.8	65.0	66.0	49.0	75.0	69.4	71.9	70.0	45.0	79.9	60.4	57.1
+ align (middle)	75.1	82.0	74.9	78.0	74.9	49.4	66.5	48.2	47.7	65.5	66.2	47.9	74.7	72.4	72.1	69.6	48.0	79.1	62.2	56.1
Qwen 2.5 SFT	74.7	81.1	74.0	77.5	74.1	27.0	67.3	32.9	23.5	57.4	58.9	45.9	74.6	63.3	70.8	60.0	34.4	79.9	59.9	46.5
+ align (middle)	74.9	82.5	74.8	78.0	75.1	36.5	68.3	39.6	30.4	57.8	63.1	42.5	74.6	63.3	70.9	61.3	35.8	80.2	58.1	47.2

Table 11: Per-languages F_1 results on slot filling.

	Supervised X→En					Supervised En→X					Transfer X→En			Transfer En→X		
	cs	de	is	ru	zh	cs	de	is	ru	zh	he	ja	uk	he	ja	uk
BLEU																
Llama 3 SFT	37.8	43.0	28.3	32.0	22.5	25.9	35.5	10.6	25.2	38.9	39.3	17.5	38.7	14.5	14.2	17.7
+ align (middle)	38.4	43.1	29.1	32.4	23.0	24.7	34.7	10.9	24.4	38.1	39.8	18.8	38.4	16.0	15.6	19.5
Qwen 2.5 SFT	36.1	40.8	20.5	30.6	23.2	21.5	33.7	6.8	25.3	45.3	34.6	18.9	35.6	13.3	17.6	13.0
+ align (middle)	36.6	41.4	21.2	30.9	24.0	20.5	32.7	4.8	25.0	45.3	36.3	19.4	36.8	12.7	17.8	13.5
COMET																
Llama 3 SFT	85.2	84.9	81.0	82.4	79.7	84.3	81.8	68.7	83.3	84.2	83.6	79.8	85.1	75.7	83.5	79.7
+ align (middle)	85.5	84.9	81.1	82.4	79.8	83.8	81.6	69.0	83.3	84.0	83.6	80.1	85.2	77.1	84.2	80.8
Qwen 2.5 SFT	84.8	84.7	74.1	82.6	80.2	80.8	80.6	52.0	83.3	86.1	82.3	81.3	84.5	70.7	85.5	74.6
+ align (middle)	85.1	84.7	74.4	82.6	80.4	79.5	80.1	46.5	83.1	85.8	82.2	81.4	84.6	70.7	85.7	74.4

Table 12: Per-languages BLEU and COMET results on machine translation.

	Supervised↑	Transfer↑
Slot filling (MASSIVE): F_1		
SFT baseline	76.6	60.2
Middle (layer 16)	77.0	61.7
Top (layer 32)	76.6	62.0
Bottom (layer 8)	76.8	58.0
Middle + Bottom	77.6	62.5
Machine translation (WMT23): COMET		
SFT baseline	81.5	79.6
Middle (layer 16)	81.5	80.7
Top (layer 32)	82.0	80.2
Bottom (layer 8)	81.2	80.1
Middle + Bottom	81.5	80.6

Table 13: Impact of alignment loss placement on supervised and transfer performance on Llama 3.

D.4 Details for Retrieval

To evaluate cross-lingual retrieval performance, we adapt the implementation from LASER[19] (Schwenk et al., 2021) to process representations extracted offline.

E Results for Individual Languages

The detailed results for Table 2 are in Table 11 (slot filling) and Table 12 (machine translation).

F Results of Aligning at Several Layers

In Table 13, we show that adding alignment losses at both middle and top layers brings further improvements on slot filling, but does not on machine translation. This task-dependent behavior indicates that how to best align multiple layers still requires further investigation.

[19] https://github.com/facebookresearch/LASER/tree/main/tasks/xsim