Representational Isomorphism and Alignment of Multilingual Large Language Models

Anonymous ACL submission

Abstract

In this paper, we investigate the capability of large language models (LLMs) to represent texts in multilingual contexts. Our findings reveal that sentence representations derived from LLMs exhibit a high degree of isomorphism across languages. This existing isomorphism facilitates representational alignments in few-shot or even zero-shot settings. Specifically, by applying a contrastive objective at the representation level with only a small number (e.g., 100) of translation pairs, we significantly improve models' performance on Semantic Textual Similarity (STS) tasks across languages. This representation-level approach proves to be more efficient and effective for semantic alignment than continued pretraining or instruction tuning. Interestingly, we also observe substantial STS improvements within individual languages, even without a monolingual objective specifically designed for this purpose.[1]

1 Introduction

Large Language Models (LLMs) demonstrate significant potential in solving multilingual tasks, e.g., machine translation (Kocmi et al., 2023) and multilingual QA (Agrawal et al., 2023). Notably, they exhibit strong few-show capacities (Xu et al., 2023; Lai et al., 2024), where a small number of samples can lead to substantial performance improvements.

Representational isomorphism has been identified as one key source of few shot capabilities in the context of word translation (Lample et al., 2017; Søgaard et al., 2018). In this paper, we analyze the multilingual sentence representation of LLMs from the perspective of isomorphism. We start by examining the geometric properties of representations derived from pairs of translation sentences. Using several widely used methods to extract embedding from LLMs, we show that although the resulting embeddings are not well clustered in a common space for different languages, they exhibit high isomorphism — projecting them through an orthogonal matrix allows the sentence representations to be effectively aligned across languages. Moreover, it also explains the previous success of combining a non-English input with an English prompt (Etxaniz et al., 2023; Huang et al., 2023).

Building on this observation, we further investigate the potential of multilingual semantic alignment upon LLMs. We show that by using a small number (e.g., 100) of English-centric translation samples equipped with contrastive losses (Gao et al., 2021) across language pairs, the representation spaces well converge and align effectively. This alignment largely enhances the performance on cross-lingual Semantic Textual Similarity (STS, Cer et al., 2017) tasks, proving to be more efficient and effective than continued language modeling training with multilingual samples. Interestingly, such progress also results in significant STS gains within each language, even in the absence of a monolingual objective specifically designed for this purpose. Given its high efficiency and effectiveness, we advocate for exploring representation-level alignment in the future.

2 Representational Analysis

2.1 Representation Extraction

Using prompts to extract sentence embeddings has been shown by Jiang et al. (2022) to yield strong performance on mask language models, e.g., BERT (Devlin et al., 2019). PromptEOL (Jiang et al., 2023) extends this method to causal language models, e.g., OPT (Zhang et al., 2023) or LLaMA (Touvron et al., 2023), by employing a prompting template as follows:

This sentence : "[TEXT]" means in one word:"

where the last layer's hidden vector for the last token """ is extracted as the sentence representation.

[1] Our anonymous code is available at `https://anonymous.4open.science/r/multilingual_reps`.

Precision@5	EN	AR	ZH	JP	RU	DE	ES	Into X
EN	- / -	0.33 / 0.67	0.61 / 0.97	0.03 / 0.82	0.36 / 0.96	0.82 / 0.96	0.76 / 0.99	0.49 / 0.90
AR	0.12 / 0.23	- / -	0.18 / 0.44	0.01 / 0.37	0.07 / 0.45	0.08 / 0.34	0.14 / 0.53	0.10 / 0.39
ZH	0.22 / 0.73	0.08 / 0.55	- / -	0.14 / 0.71	0.31 / 0.88	0.18 / 0.74	0.40 / 0.93	0.22 / 0.76
JP	0.04 / 0.33	0.02 / 0.34	0.21 / 0.59	- / -	0.17 / 0.56	0.03 / 0.56	0.06 / 0.62	0.09 / 0.50
RU	0.20 / 0.73	0.19 / 0.61	0.56 / 0.86	0.05 / 0.71	- / -	0.24 / 0.85	0.60 / 0.95	0.31 / 0.79
DE	0.67 / 0.88	0.09 / 0.62	0.37 / 0.89	0.01 / 0.80	0.36 / 0.92	- / -	0.83 / 0.96	0.39 / 0.85
ES	0.12 / 0.75	0.08 / 0.60	0.18 / 0.87	0.00 / 0.67	0.20 / 0.92	0.48 / 0.85	- / -	0.18 / 0.78
From X	0.23 / 0.61	0.13 / 0.57	0.35 / 0.77	0.04 / 0.68	0.24 / 0.78	0.3 / 0.72	0.47 / 0.83	0.25 / 0.71

Table 1: The success rate (Precision@5) for cross-lingual retrieval **before/after** applying Procrustes projection. The embeddings in each language are derived from the LLaMA2-7B model using the prompting method as described in §2.1. “From X” and “Into X” denote the average results for each column and row, respectively. The Procurstes projection W for each translation direction is trained on NTREX, while the Precision@5 is tested based on the translation sentences from Flores. We report results derived from LLaMA2-13B in Appendix A.4.

Precision@5	EN	AR	ZH	JP	RU	DE	ES	Into X
EN	- / -	0.78 / 0.73	0.93 / 0.94	0.95 / 0.93	0.76 / 0.94	0.96 / 0.96	0.97 / 0.97	0.89 / 0.91
AR	0.67 / 0.67	- / -	0.83 / 0.76	0.84 / 0.74	0.59 / 0.76	0.82 / 0.78	0.83 / 0.79	0.76 / 0.75
ZH	0.85 / 0.93	0.86 / 0.79	- / -	0.99 / 0.98	0.84 / 0.95	0.97 / 0.95	0.96 / 0.96	0.91 / 0.93
JP	0.88 / 0.92	0.86 / 0.78	1.0 / 0.97	- / -	0.83 / 0.95	0.96 / 0.95	0.95 / 0.95	0.91 / 0.92
RU	0.75 / 0.96	0.83 / 0.81	0.97 / 0.96	0.97 / 0.96	- / -	0.97 / 0.97	0.96 / 0.97	0.91 / 0.94
DE	0.9 / 0.96	0.68 / 0.79	0.91 / 0.94	0.89 / 0.94	0.75 / 0.96	- / -	0.99 / 0.97	0.85 / 0.93
ES	0.89 / 0.96	0.65 / 0.77	0.87 / 0.94	0.85 / 0.94	0.65 / 0.95	0.98 / 0.96	- / -	0.82 / 0.92
From X	0.82 / 0.9	0.78 / 0.78	0.92 / 0.92	0.91 / 0.92	0.74 / 0.92	0.94 / 0.93	0.94 / 0.93	0.86 / 0.90

Table 2: The success rate (Precision@5) for cross-lingual retrieval **before/after** applying Procrustes projection. Note that all embeddings are derived from the prompting template in English, instead of the same language with input sentences. We report results derived from LLaMA2-13B in Appendix A.4.

This method achieves competitive performance on semantic representation tasks (Agirre et al., 2015, 2016). Moreover, it provides an effective way to investigate the representations of LLMs.

Although some studies (Springer et al., 2024; Lei et al., 2024) have achieved more advanced performance using prompting, we adopt PromptEOL in this paper for its simplicity and generalizability. To adapt PromptEOL to a multilingual setting, we translate the English template mentioned above into other corresponding languages, e.g., a template,

Dieser Satz: “ [TEXT] ” bedeutet in einem Wort:“

is used for German. In the following sections, we derive representations of LLMs across languages by applying this method.

2.2 Cross-lingual Structural Analysis

We leverage *Procrustes* analysis (Schönemann, 1966) to measure the structural similarity of representations across languages. This method finds the optimal rotation and/or reflection (i.e., orthogonal linear transformation) to match points in a set of shapes, which ensures that the shape remains unchanged. Therefore, the precision in matching reflects the degree of isomorphism across spaces.

Formally, let’s assume there are two sets of embeddings, A and B, derived from LLMs using sentence pairs in two different languages. Procrustes analysis learns an orthogonal linear projection W to map A into a shared space with B, by solving $min \|WA - B\|_F$ subject to $W^T W = I$. A closed-form solution $W = UV^T$ can be advantageously obtained from the singular value decomposition (SVD) of BA^T.

In this paper, we conduct experiments on seven languages, namely EN, AR, ZH, JP, RU, DE, and ES, which encompass both similar and different language families and writing scripts. We investigate the structure similarity across all possible translation directions by training W on the corresponding translation samples built from NTREX (Federmann et al., 2022) and then testing on Flores (Goyal et al., 2022)[2]. Note that NTREX mainly focuses on the News domain while Flores is built from Wikipedia. Such out-of-domain testing helps to assess the robustness and generalization capabilities and provides a more realistic measure of how LLMs can handle diverse and unexpected inputs.

[2]NTREX and Flores are both multi-parallel. So it is easy to build translation data in each involved direction. Here, we merge all bitext in *dev* and *test* set for NTREX and Flores, resulting in 1,997 and 2,009 samples, respectively.

Specifically, we calculate Precision@k by using embeddings in WA to retrieve those in B and determine whether their counterparts are within the k-nearest neighbors based on cosine similarity. We use the precision after rotation to indicate the structural similarity within each translation direction.

2.3 Representation Discrepancy and Isomorphism

We begin our investigation by using sentence embeddings derived from prompting methods as mentioned in §2.1. Table 1 shows the success rate of the resulting embeddings in cross-lingual retrieval before/after applying Procrustes projection (§2.2). It is evident that 1) the initial representation discrepancies are generally substantial across languages, such as EN→JP (0.01), except for a few language pairs that are closer or use the same scripts, e.g., EN→DE (0.82). 2) However, after properly rotating (applying W), representations in most of the directions are well aligned, leading to clear gains from an average of 0.25 to 0.71. We also apply another representation extraction method that takes the last token's output embedding without prompting as representations (last token pooling). The results shown in Appendix A.3 demonstrate this method performs significantly worse than prompting-based methods. The results obtained from the LLaMA2-13B model are provided in Appendix A.4.

Overall, we argue that although representations from LLMs vary significantly across languages, they exhibit a high degree of isomorphism — properly rotating and/or reflecting the representation space can effectively align them.

2.4 Multilingual Representation via English Prompts

Previous studies show decent improvements can be achieved by simply adjusting/filling non-English instructions into English-centric prompting templates in the inference stage (Etxaniz et al., 2023; Huang et al., 2023). To explain the success, we investigate how the representations of LLMs change when using the prompting template in the predominant language, English, for different languages, rather than the same ones mentioned in §2.1.

Table 2 shows the success rate within the same data setting in §2.3. Notably, the initial representations' degree of alignment is much higher than that in Table 1 (0.86 v.s., 0.25), resulting in a similar alignment level with the latter after rotation. Also, the gain from applying Procrustes projection is marginal in this setting. We interpret the degeneration of the rotation gain as that English prompts, to some extent, have taken on the role of the corresponding spatial transformation, i.e., mapping representations into a shared English space.

In the following sections, we refer to using these English prompts (en-prompts) with non-English sentences as zero-shot representation alignment and conduct experiments based on this setting.

3 Semantic Analysis

3.1 Semantic Textual Similarity

In this section, we examine the multilingualism of LLM representations through the lens of Semantic Textual Similarity (STS) (Agirre et al., 2015, 2016). Each sentence pair in STS datasets is annotated from 0 to 5 indicating the pairwise semantic similarity. The Spearman correlation between the model-predicted and human-annotated similarity scores is used as the metric. The STS-17 shared task (Camacho-Collados et al., 2017) extends English-centric STS evaluation to multilingual settings. In this paper, we conduct experiments based on STS-17, which encompasses 3 monolingual STS (EN, AR, and ES) and 3 cross-lingual STS (AR-EN, ES-EN, and TR-EN) tasks.

Given the structure similarity of representations across languages, we test the few-shot capacity of aligning cross-lingual semantics within LLMs in the following sections.

3.2 Cross-lingual Contrastive Learning

Contrastive learning (Hadsell et al., 2006) learns effective representation by pulling semantically close neighbors together and pushing apart non-neighbors. Formally, given a set of paired examples $D = \{(x_i, x_i^+)\}_{i=1}^m$, where x_i and x_i^+ are semantically related, following Chen et al. (2020), a cross-entropy loss ℓ_i with in-batch negatives can be defined as follows:

$$\ell_i = -\log \frac{e^{sim(h_i, h_i^+)/\tau}}{\sum_{j=1}^{N} e^{sim(h_i, h_j^+)/\tau}}, \qquad (1)$$

where h_i is the representation of x_i, τ is a temperature hyperparameter, and $sim(h_i, h_j)$ is the cosine similarity. In this paper, we directly extend the objective (Eq. 1) into a cross-lingual setting, where x_i and x_i^+ refer to the i-th possible translation pair.

Training Setting. We select 1,000 multi-parallel samples from NTREX as the training set and

Model	Settings	EN	AR	ES	AR-EN	ES-EN	TR-EN	Avg
LLaMA2-7B	$self$-prompts	0.72	0.24	0.28	0.17	0.11	0.09	0.27
LLaMA2-7B	en-prompts	0.72	0.46	0.46	0.36	0.27	0.12	0.40
LLaMA2-7B	en-prompts (+100)	0.76	0.62	0.73	0.52	0.64	0.42	0.62
LLaMA2-7B	en-prompts (+1000)	0.82	0.62	0.80	0.54	0.75	0.55	0.68
Tower-7B	$self$-prompts	0.69	0.25	0.41	0.14	0.15	0.08	0.29
Tower-7B	en-prompts	0.69	0.45	0.70	0.26	0.35	0.11	0.43
Tower-7B	en-prompts (+100)	0.73	0.57	0.67	0.50	0.60	0.41	0.58
Tower-7B	en-prompts (+1000)	0.76	0.60	0.65	0.54	0.62	0.47	0.61

Table 3: The multilingual and cross-lingual STS results in different settings. $self$-prompts and en-prompts denote using prompting methods in §2.1 and §2.4, respectively. Tower continues to pre-train LLaMA2 with large amounts of multilingual data but fails to align semantics. However, aligning LLaMA2 at the representation level using a few translation samples from NTREX (e.g., 100), results in clear improvements from 0.40 to 0.68. We provide results derived from other sizes of LLMs in Appendix A.5.

construct pair-wise samples covering EN→AR, AR→EN, EN→ES, and ES→EN[3]. Meanwhile, we leave TR-involved data empty to investigate the potential impact on unseen languages. We apply the in-batch cross-entropy loss as the objective and fine-tune LLMs with LoRA (Hu et al., 2021). Detailed hyperparameters are in Appendix A.1.

We compare cross-lingual STS under varying settings, including 1) Zero-shot prompting using self-language for the template, see §2.1, 2) Zero-shot prompting using English templates, see §2.4, 3) Using Tower (Alves et al., 2024) as the backbone, a multilingual LLM extensively trained on multilingual data based on LLaMA2, and 4) Applying cross-lingual contrastive objective. The summarized results can be found in Table 3.

3.3 Results and Discussion

Semantic Alignment across Languages. In Table 3, we show that the initial semantic representation ($self$-prompts) is poor while applying en-prompts leads to relatively higher performance, which is in line with the representational analysis in §2.4. Applying contrastive objectives at the representation level, even with just 100 samples, results in strong overall STS improvements from 0.40 to 0.62. Further gain can be achieved by extending the training size from 100 to 1,000 samples.

Interestingly, although the training objectives (see §3.2) are designed from a cross-lingual perspective — aligning representations from other languages to English — the monolingual STS performances (EN, AR, ES) also show clear improvements. Notably, even the performance of the predominant language, English, improves significantly. We preliminarily interpret this phenomenon as representation alignment leads to better grounding across languages, however, we leave in-depth explorations for the future.

Sample- and Representation-Level Alignments. We observe that current studies (Xu et al., 2023; Alves et al., 2024; Lai et al., 2024; Gao et al., 2024) about the multilingualism of LLMs are mainly focusing on sample-level alignments, i.e., extending training or fine-tuning samples beyond English. E.g., Tower further pre-trained on a multilingual dataset encompassing 20 billion tokens based on LLaMA2. In Table 3, we clearly show that, despite extensive sample-level alignments, Tower's semantic representation still fails to effectively generalize across languages, resulting in marginal gains over the base model, LLaMA2. Also, Gao et al. (2024) demonstrate that neither multilingual pretraining nor instruction tuning can substantially improve cross-lingual knowledge conductivity. To this end, we advocate for exploring representation-level alignment in the future given its high efficiency and effectiveness in semantic alignments.

4 Conclusion

In this paper, we investigate the representation of LLMs from the perspectives of both geometric and semantic similarity. We show that LLMs' representations exhibit a high degree of isomorphism across languages, which explains their cross-lingual zero-shot or few-shot capabilities in a multilingual context. For example, we show that the semantics representation of LLMs can be easily enhanced across languages by alignment at the representation level using as few as 100 translation samples, which is much more efficient and effective than sample-level pretraining or instruction tuning.

[3]We cover both directions for each language pair to ensure all involved languages have a chance to be treated as negative samples in a batch.

Limitations

We conduct experiments exclusively on two families of LLMs, namely LLaMA2 and Tower. Therefore, the generalizability of our findings to other LLMs remains uncertain. Additionally, due to the limited language coverage in the STS17 task, our semantic analysis is restricted to a few languages.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2023. QAmeleon: Multilingual QA with only 5 examples. *Transactions of the Association for Computational Linguistics*, 11:1754–1771.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english? *arXiv preprint arXiv:2308.01223*.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. Ntrex-128–news test references for mt evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. *arXiv preprint arXiv:2404.04659*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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'06)*, volume 2, pages 1735–1742. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2023. Scaling sentence embeddings with large language models. *arXiv preprint arXiv:2307.16645*.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. PromptBERT: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837, Abu Dhabi, United Arab Emirates. 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, Maja Popović, and Mariya Shmatova. 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.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback. *arXiv preprint arXiv:2406.01771*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Yibin Lei, Di Wu, Tianyi Zhou, Tao Shen, Yu Cao, Chongyang Tao, and Andrew Yates. 2024. Meta-task prompting elicits embedding from large language models. *arXiv preprint arXiv:2402.18458*.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. *arXiv preprint arXiv:1805.03620*.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2024. Repetition improves language model embeddings. *arXiv preprint arXiv:2402.15449*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2023. Opt: Open pre-trained transformer language models, 2022. *URL https://arxiv. org/abs/2205.01068*, 3:19–0.

A Appendix

A.1 Fine-tuning Hyperparameters

We set the same hyperparameters for all experiments. The LoRA is applied to all linear layers. The LoRA rank is 64, alpha is 16, and dropout is 0.05. The batch size is set to 32 and the gradient is accumulated for 4 steps, resulting in an actual batch size of 128. The learning rate is set to 5e-4. For experiments of fine-tuning with 100 and 1,000 samples, we trained with 10 and 3 epochs.

A.2 Representation Isomorphism with Additional Metrics

We present the results of Precision@1 and Precision@10 on representation isomorphism with LLaMA-7B in Table 4, 5, 6, and 7.

A.3 Representation Isomorphism with Last Token Pooling-Derived Representations

Table 8 shows the results on representation isomorphism with last token pooling-derived representations of the LLaMA2-7B model.

A.4 Representation Isomorphism with LLaMA-13B

Table 9 and 10 show the results on representation isomorphism with the LLaMA2-13B model.

A.5 Semantic Alignment across Languages

Table 11 shows the multilingual cross-lingual STS results in different settings upon 13B LLMs.

Precision@1	EN	AR	ZH	JP	RU	DE	ES	Into X
EN	- / -	0.20 / 0.47	0.44 / 0.88	0.01 / 0.63	0.19 / 0.87	0.65 / 0.88	0.54 / 0.93	0.34 / 0.78
AR	0.06 / 0.09	- / -	0.10 / 0.26	0.00 / 0.2	0.03 / 0.26	0.02 / 0.21	0.06 / 0.33	0.05 / 0.23
ZH	0.07 / 0.52	0.02 / 0.36	- / -	0.07 / 0.50	0.12 / 0.71	0.07 / 0.57	0.11 / 0.79	0.08 / 0.57
JP	0.01 / 0.15	0.00 / 0.19	0.10 / 0.38	- / -	0.08 / 0.35	0.01 / 0.38	0.02 / 0.40	0.04 / 0.31
RU	0.01 / 0.52	0.01 / 0.43	0.38 / 0.72	0.02 / 0.54	- / -	0.09 / 0.73	0.36 / 0.86	0.14 / 0.63
DE	0.40 / 0.72	0.01 / 0.42	0.02 / 0.73	0.00 / 0.63	0.21 / 0.83	- / -	0.62 / 0.88	0.21 / 0.70
ES	0.02 / 0.55	0.04 / 0.41	0.09 / 0.72	0.00 / 0.49	0.11 / 0.80	0.26 / 0.73	- / -	0.09 / 0.62
From X	0.10 / 0.42	0.05 / 0.38	0.19 / 0.62	0.02 / 0.50	0.12 / 0.64	0.18 / 0.58	0.28 / 0.70	0.14 / 0.55

Table 4: The success rate (Precision@1) for cross-lingual retrieval **before/after** applying Procrustes projection with the **LLaMA2-7B** model. The embeddings in each language are derived from the LLaMA2-7B model using the prompting method as described in §2.1.

Precision@10	EN	AR	ZH	JP	RU	DE	ES	Into X
EN	- / -	0.40 / 0.73	0.67 / 0.98	0.05 / 0.88	0.44 / 0.98	0.86 / 0.97	0.82 / 0.99	0.54 / 0.92
AR	0.16 / 0.31	- / -	0.24 / 0.51	0.02 / 0.45	0.12 / 0.54	0.12 / 0.41	0.19 / 0.62	0.14 / 0.47
ZH	0.30 / 0.80	0.16 / 0.62	- / -	0.20 / 0.77	0.40 / 0.91	0.28 / 0.80	0.53 / 0.95	0.31 / 0.81
JP	0.06 / 0.41	0.06 / 0.42	0.28 / 0.69	- / -	0.23 / 0.64	0.06 / 0.65	0.13 / 0.70	0.14 / 0.58
RU	0.27 / 0.80	0.27 / 0.68	0.63 / 0.90	0.08 / 0.76	- / -	0.34 / 0.89	0.69 / 0.97	0.38 / 0.83
DE	0.78 / 0.92	0.16 / 0.69	0.46 / 0.92	0.04 / 0.84	0.43 / 0.95	- / -	0.88 / 0.97	0.46 / 0.88
ES	0.24 / 0.82	0.10 / 0.67	0.24 / 0.90	0.02 / 0.73	0.27 / 0.94	0.56 / 0.89	- / -	0.24 / 0.83
From X	0.30 / 0.68	0.19 / 0.64	0.42 / 0.82	0.07 / 0.74	0.32 / 0.83	0.37 / 0.77	0.54 / 0.87	0.32 / 0.76

Table 5: The success rate (Precision@10) for cross-lingual retrieval **before/after** applying Procrustes projection with the **LLaMA2-7B** model. The embeddings in each language are derived from the LLaMA2-7B model using the prompting method as described in §2.1.

Precision@1	EN	AR	ZH	JP	RU	DE	ES	Into X
EN	- / -	0.59 / 0.52	0.83 / 0.81	0.83 / 0.80	0.57 / 0.82	0.87 / 0.88	0.87 / 0.90	0.76 / 0.79
AR	0.50 / 0.44	- / -	0.68 / 0.56	0.69 / 0.56	0.41 / 0.58	0.63 / 0.61	0.65 / 0.63	0.59 / 0.56
ZH	0.70 / 0.79	0.67 / 0.60	- / -	0.96 / 0.92	0.68 / 0.86	0.89 / 0.87	0.80 / 0.88	0.78 / 0.82
JP	0.74 / 0.77	0.69 / 0.59	0.97 / 0.91	- / -	0.67 / 0.85	0.87 / 0.85	0.81 / 0.86	0.79 / 0.81
RU	0.51 / 0.84	0.63 / 0.64	0.91 / 0.88	0.88 / 0.87	- / -	0.88 / 0.93	0.86 / 0.91	0.78 / 0.85
DE	0.80 / 0.87	0.51 / 0.61	0.80 / 0.85	0.78 / 0.85	0.57 / 0.89	- / -	0.95 / 0.92	0.73 / 0.83
ES	0.76 / 0.87	0.45 / 0.58	0.73 / 0.83	0.69 / 0.82	0.46 / 0.87	0.94 / 0.91	- / -	0.67 / 0.81
From X	0.67 / 0.76	0.59 / 0.59	0.82 / 0.81	0.81 / 0.80	0.56 / 0.81	0.85 / 0.84	0.82 / 0.85	0.73 / 0.78

Table 6: The success rate (Precision@1) for cross-lingual retrieval **before/after** applying Procrustes projection with the **LLaMA2-7B** model. Note that all embeddings are derived from the prompting template in English as described in §2.4, instead of the same language with input sentences.

Precision@10	EN	AR	ZH	JP	RU	DE	ES	Into X
EN	- / -	0.83 / 0.80	0.95 / 0.96	0.97 / 0.95	0.80 / 0.96	0.98 / 0.97	0.98 / 0.98	0.92 / 0.94
AR	0.73 / 0.75	- / -	0.88 / 0.81	0.89 / 0.80	0.66 / 0.82	0.87 / 0.84	0.87 / 0.84	0.82 / 0.81
ZH	0.89 / 0.95	0.90 / 0.84	- / -	1.00 / 0.98	0.89 / 0.97	0.98 / 0.97	0.98 / 0.97	0.94 / 0.95
JP	0.91 / 0.94	0.90 / 0.83	1.00 / 0.98	- / -	0.88 / 0.97	0.98 / 0.97	0.98 / 0.97	0.94 / 0.94
RU	0.80 / 0.97	0.88 / 0.86	0.98 / 0.97	0.98 / 0.97	- / -	0.98 / 0.98	0.98 / 0.98	0.93 / 0.96
DE	0.93 / 0.97	0.74 / 0.84	0.94 / 0.96	0.92 / 0.96	0.79 / 0.97	- / -	0.99 / 0.98	0.89 / 0.95
ES	0.92 / 0.97	0.71 / 0.82	0.90 / 0.96	0.88 / 0.96	0.72 / 0.96	0.99 / 0.97	- / -	0.85 / 0.94
From X	0.86 / 0.92	0.83 / 0.83	0.94 / 0.94	0.94 / 0.94	0.79 / 0.94	0.96 / 0.95	0.96 / 0.95	0.90 / 0.93

Table 7: The success rate (Precision@10) for cross-lingual retrieval **before/after** applying Procrustes projection with the **LLaMA2-7B** model. Note that all embeddings are derived from the prompting template in English as described in §2.4, instead of the same language with input sentences.

Precision@5	EN	AR	ZH	JP	RU	DE	ES	Into X
EN	- / -	0.05 / 0.23	0.04 / 0.51	0.08 / 0.41	0.13 / 0.54	0.09 / 0.57	0.08 / 0.70	0.08 / 0.49
AR	0.03 / 0.07	- / -	0.02 / 0.13	0.02 / 0.08	0.03 / 0.13	0.01 / 0.12	0.02 / 0.16	0.02 / 0.12
ZH	0.19 / 0.24	0.08 / 0.18	- / -	0.46 / 0.34	0.15 / 0.37	0.19 / 0.40	0.11 / 0.44	0.20 / 0.33
JP	0.11 / 0.12	0.06 / 0.09	0.35 / 0.25	- / -	0.05 / 0.17	0.08 / 0.13	0.06 / 0.17	0.12 / 0.15
RU	0.15 / 0.23	0.05 / 0.12	0.08 / 0.30	0.06 / 0.15	- / -	0.19 / 0.36	0.18 / 0.45	0.12 / 0.27
DE	0.06 / 0.20	0.02 / 0.10	0.03 / 0.28	0.04 / 0.11	0.09 / 0.38	- / -	0.18 / 0.45	0.07 / 0.25
ES	0.07 / 0.28	0.02 / 0.14	0.02 / 0.33	0.02 / 0.15	0.08 / 0.45	0.13 / 0.43	- / -	0.06 / 0.30
From X	0.10 / 0.19	0.05 / 0.14	0.09 / 0.30	0.11 / 0.21	0.09 / 0.34	0.12 / 0.33	0.10 / 0.40	0.10 / 0.27

Table 8: The success rate (Precision@5) for cross-lingual retrieval **before/after** applying Procrustes projection with the **LLaMA2-7B** model. The embeddings are derived by taking the output hidden vector of the last token without prompting (**last token pooling**).

Precision@5	EN	AR	ZH	JP	RU	DE	ES	Into X
EN	- / -	0.26 / 0.72	0.66 / 0.90	0.66 / 0.88	0.22 / 0.96	0.56 / 0.85	0.30 / 0.83	0.44 / 0.86
AR	0.02 / 0.37	- / -	0.09 / 0.28	0.11 / 0.34	0.10 / 0.64	0.03 / 0.33	0.03 / 0.41	0.06 / 0.40
ZH	0.02 / 0.68	0.04 / 0.29	- / -	0.42 / 0.50	0.02 / 0.68	0.00 / 0.32	0.00 / 0.38	0.08 / 0.47
JP	0.02 / 0.62	0.05 / 0.40	0.74 / 0.54	- / -	0.05 / 0.86	0.01 / 0.57	0.01 / 0.53	0.15 / 0.59
RU	0.01 / 0.43	0.07 / 0.30	0.07 / 0.28	0.12 / 0.43	- / -	0.02 / 0.47	0.02 / 0.48	0.05 / 0.40
DE	0.47 / 0.84	0.24 / 0.61	0.19 / 0.57	0.52 / 0.79	0.20 / 0.95	- / -	0.41 / 0.80	0.34 / 0.76
ES	0.25 / 0.71	0.29 / 0.52	0.09 / 0.46	0.46 / 0.57	0.14 / 0.83	0.52 / 0.70	- / -	0.29 / 0.63
From X	0.13 / 0.61	0.16 / 0.47	0.31 / 0.51	0.38 / 0.58	0.12 / 0.82	0.19 / 0.54	0.13 / 0.57	0.20 / 0.59

Table 9: The success rate (Precision@5) for cross-lingual retrieval **before/after** applying Procrustes projection with the **LLaMA2-13B** model. The embeddings in each language are derived from the LLaMA2-13B model using the prompting method as described in §2.1.

Precision@5	EN	AR	ZH	JP	RU	DE	ES	Into X
EN	- / -	0.89 / 0.82	0.90 / 0.94	0.89 / 0.93	0.77 / 0.94	0.99 / 0.98	0.98 / 0.98	0.90 / 0.93
AR	0.81 / 0.80	- / -	0.82 / 0.86	0.86 / 0.85	0.78 / 0.85	0.94 / 0.88	0.94 / 0.88	0.86 / 0.85
ZH	0.59 / 0.95	0.89 / 0.88	- / -	1.00 / 0.98	0.88 / 0.97	0.97 / 0.97	0.99 / 0.98	0.89 / 0.96
JP	0.69 / 0.94	0.91 / 0.87	1.00 / 0.99	- / -	0.91 / 0.96	0.98 / 0.98	0.99 / 0.97	0.91 / 0.95
RU	0.44 / 0.95	0.94 / 0.89	0.94 / 0.98	0.95 / 0.97	- / -	0.98 / 0.99	0.98 / 0.98	0.87 / 0.96
DE	0.98 / 0.98	0.94 / 0.90	0.94 / 0.98	0.94 / 0.97	0.91 / 0.98	- / -	1.00 / 1.00	0.95 / 0.97
ES	0.95 / 0.97	0.93 / 0.88	0.90 / 0.97	0.91 / 0.96	0.86 / 0.97	0.99 / 0.98	- / -	0.92 / 0.96
From X	0.74 / 0.93	0.92 / 0.87	0.92 / 0.95	0.93 / 0.94	0.85 / 0.94	0.97 / 0.96	0.98 / 0.96	0.90 / 0.94

Table 10: The success rate (Precision@5) for cross-lingual retrieval **before/after** applying Procrustes projection with the **LLaMA2-13B** model. Note that all embeddings are derived from the prompting template in English as described in §2.4, instead of the same language with input sentences.

Model	Settings	EN	AR	ES	AR-EN	ES-EN	TR-EN	Avg
LLaMA2-13B	en-prompts	0.72	0.55	0.60	0.45	0.31	0.28	0.49
LLaMA2-13B	en-prompts (+100)	0.74	0.57	0.63	0.57	0.66	0.52	0.62
LLaMA2-13B	en-prompts (+1000)	0.77	0.62	0.71	0.61	0.63	0.55	0.65
Tower-13B	en-prompts	0.73	0.59	0.64	0.37	0.42	0.49	0.54
Tower-13B	en-prompts (+100)	0.66	0.60	0.67	0.51	0.53	0.45	0.57
Tower-13B	en-prompts (+1000)	0.69	0.63	0.68	0.57	0.61	0.51	0.62

Table 11: The multilingual and cross-lingual STS results derived from **LLaMA2-13B** and **Tower-13B** in different settings. $self$-prompts and en-prompts denote using prompting methods in §2.1 and §2.4, respectively.