Representational Isomorphism and Alignment of Multilingual Large Language Models

Di Wu* **Yibin Lei*** **Andrew Yates** **Christof Monz**
University of Amsterdam
{d.wu, y.lei, a.c.yates, c.monz}@uva.nl

Abstract

In this extended abstract, 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 settings. Specifically, by applying a contrastive objective at the representation level with only a small number (e.g., 100) of translation pairs,[1] we significantly improve models' performance on Semantic Textual Similarity (STS) tasks across languages.[2]

1 Introduction

Representational isomorphism has been recognized as a key factor of few-shot capabilities (Lample et al., 2017; Søgaard et al., 2018). In this paper, we analyze multilingual sentence representations in LLMs through the lens of isomorphism. By examining the geometric properties of sentence pairs, we show that while embeddings from different languages are not well clustered in a common space, they exhibit high isomorphism. Projecting them via an orthogonal matrix effectively aligns representations across languages. It also explains the previous success of combining non-English inputs with English prompts (Etxaniz et al., 2023; Huang et al., 2023).

Building on this observation and previous studies highlighting representational isomorphism as a key factor in few-shot capabilities, we explore multilingual semantic alignment in LLMs. Using just 100 English-centric translation samples with contrastive loss across language pairs, we achieve effective representation space alignment. This significantly improves cross-lingual Semantic Textual Similarity (STS) task performance, proving more efficient than continued multilingual training. Notably, this also boosts STS performance within individual languages, even without a monolingual objective.

2 Representational Analysis

2.1 Representation Extraction

PromptEOL (Jiang et al., 2023) extracts sentence embeddings from causal language models like LLaMA (Touvron et al., 2023) using a simple prompting template:

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

The last hidden layer's vector for the final token is used as the sentence representation. This method has demonstrated strong performance on semantic representation tasks (Agirre et al., 2015, 2016).

We adopt PromptEOL for its simplicity and adaptability. For multilingual use, the English template is translated into corresponding languages, e.g., for German:

Dieser Satz: "[TEXT]" bedeutet in einem Wort:"

We use this method to derive multilingual LLM representations.

2.2 Cross-lingual Structural Analysis

We use Procrustes analysis (Schönemann, 1966) to assess the structural similarity of representations across languages. This method optimally rotates or reflects one set of points to align with another, preserving the shape. The accuracy of this alignment indicates the degree of isomorphism across spaces.

Formally, given two embedding sets, A and B, from LLMs using sentence pairs in different languages, Procrustes analysis learns an orthogonal projection W that maps A to a shared space with B by solving $min\,\|WA - B\|_F$ subject to

*These authors contributed equally to this work.

[1] Due to page limits, these results are not included in the extended abstract.

[2] 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. "From X" and "Into X" denote the average results for each column and row, respectively.

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.

$W^TW = I$. The solution $W = UV^T$ is derived from the singular value decomposition (SVD) of BA^T.

We conduct experiments on seven languages. We train W on translation pairs from NTREX (Federmann et al., 2022) and test on Flores (Goyal et al., 2022), merging 1,997 and 2,009 samples from the *dev* and *test* sets, respectively.

We then compute Precision@k by using embeddings in WA to retrieve those in B and checking if their counterparts are among the k-nearest neighbors based on cosine similarity, using this precision to quantify structural similarity in each translation direction.

2.3 Representation Discrepancy and Isomorphism

Table 1 shows the success rate of the resulting embeddings in cross-lingual retrieval before/after applying Procrustes projection (§2.2). It is clear that 1) the initial representation discrepancies are generally substantial across languages. 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.

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.

3 Conclusion

In this extended abstract, 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. Further experiments demonstrate that LLMs' semantic representations can be enhanced across languages through alignment using just 100 translation samples, offering a more efficient and effective approach than sample-level pretraining or instruction tuning.

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, 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).

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.

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.

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*.

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

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*.

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*.

A Appendix

A.1 Semantic Alignment across Languages on STS tasks

Table 3 shows the multilingual cross-lingual STS results in different settings after contrastive learning on both LLaMA2 and Tower models.

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.

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 using contrastive learning. $self$-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.

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.