Proceedings of CONF-MLA 2026 Symposium: Explainable Computing, Modeling & Data Science in Complex Systems
DOI: 10.54254/2755-2721/2026.GF35547

Layer-Wise English-Chinese Alignment and Target-Word Prediction in Transformer Language Models

Yuchen Zhang

Department of Computer Science, University of WashingtonSeattle, Seattle, USA
yuchz68@cs.washington.edu

Abstract. This paper examines whether layer-wise English-Chinese cross-lingual alignment in decoder-only language models is associated with controlled target-word prediction performance. We compare EleutherAI/Pythia-1.4B, Microsoft Phi-3.5-mini-instruct, and Qwen3-1.7B-Base. For each model, we extract target-word hidden representations across Transformer layers and evaluate four mapping conditions: no mapping, orthogonal Procrustes, ridge regression, and a residual multilayer perceptron. We also evaluate final-layer target-word prediction and an auxiliary multilingual language-identification task. Results show that English-Chinese alignment generally becomes stronger in deeper layers, and that ridge and MLP mappings usually outperform no mapping and Procrustes. However, alignment and prediction are related but not identical: some models show strong geometric alignment without the best target prediction. These findings suggest that representation alignment captures an important internal property of multilingual models, while final prediction also depends on tokenization, output distributions, instruction tuning, and training data.

Keywords: cross-lingual alignment, multilingual language models, Transformer representations, target-word prediction, English-Chinese NLP

1. Introduction

Modern language models show multilingual ability, but it remains unclear how much their internal representations are shared across languages. Earlier work on cross-lingual word embeddings found that representation spaces from different languages can often be aligned using simple mappings [1-3]. This paper asks whether a similar pattern appears inside decoder-only Transformer language models, and whether stronger English-Chinese alignment is associated with better target-word prediction.

We focus on three models: EleutherAI/Pythia-1.4B, Microsoft Phi-3.5-mini-instruct, and Qwen3-1.7B-Base. For paired English and Chinese target words, we extract layer-wise hidden states and measure how well Chinese representations can be mapped into the English representation space. We also evaluate whether the models can predict the target word from the preceding context. The study is guided by two questions: (1) do the three models show similar layer-wise English-Chinese alignment trends, and (2) is stronger alignment generally associated with better target-word prediction?

DOI: 10.54254/2755-2721/2026.GF35547

The main finding is that alignment usually improves in deeper layers, although the detailed pattern differs by model. Learned mappings, especially ridge regression and MLP, recover stronger cross-lingual structure than no mapping or Procrustes. Prediction results show a positive but imperfect connection between representation alignment and model behavior. This suggests that alignment is useful for interpreting multilingual representations, but final lexical prediction also depends on model-specific factors.

2. Related work

Multilinguality in LLMs. Large language models are still strongly shaped by English-dominant training data, and prior work has shown that performance in non-English languages often lags behind English performance [4-7]. This motivates internal analysis rather than relying only on final task accuracy.

Cross-lingual representation alignment. Prior research on bilingual lexicon induction and unsupervised machine translation shows that word embedding spaces can often be aligned with linear transformations [1-3, 8, 9]. More recent work suggests that contextualized Transformer representations may also retain alignable cross-lingual structure [10, 11].

Layer-wise multilingual representations. Interpretability studies suggest that multilingual models contain both shared and language-specific components. Some work finds greater cross-lingual similarity in middle or later layers, while language-specific neurons or components may appear in later layers [12-16]. These findings motivate the layer-wise design of this study.

Prediction-based analysis. Logit Lens-style analysis projects hidden states through the output head to inspect token predictions [17]. Claims that LLMs "think in English" have been debated because intermediate predictions can be affected by embedding geometry, frequency, and hubness [18-20]. We therefore use prediction as a final-layer behavioral measure, not as direct proof of internal language identity.

3. Method and experimental setup

3.1. Dataset and models

The dataset contains 1800 controlled English-Chinese samples generated with ChatGPT-5.5 using a prompt designed to cover nouns, verbs, and adjectives. Each sample contains an English target word, a Chinese target word, an English sentence, and a Chinese sentence expressing the same concept. Sentence contexts were designed to make the target word inferable while avoiding a fixed template.

For alignment, the target word is located inside each sentence and the model's hidden states over the target-token span are averaged. If a target word is split into multiple tokens, the span representation is the average of its token hidden states. For prediction, the model receives the sentence prefix before the target word and is evaluated on whether it predicts the target. The models are EleutherAI/Pythia-1.4B [21], Microsoft Phi-3.5-mini-instruct [22], and Qwen3-1.7B-Base [23].

3.2. Alignment methods

We compare four methods for mapping Chinese target-word representations into the English representation space. All methods are trained and evaluated separately at each Transformer layer.

DOI: 10.54254/2755-2721/2026.GF35547

No Mapping. The No Mapping baseline directly computes cosine similarity between the English target representation and the Chinese target representation without applying any transformation. This baseline measures how much English and Chinese representations are already aligned in the original hidden space.

Orthogonal Procrustes. The Procrustes method learns an orthogonal linear transformation from Chinese representations to English representations. Let $X \in \mathbb{R}^{N \times d}$ denote the Chinese target representations and $Y \in \mathbb{R}^{N \times d}$ denote the corresponding English target representations. The method solves

$$W^* = \arg \min_{W : W^T W = I} \| XW - Y \|_F^2. \tag{1}$$

The orthogonality constraint preserves the overall geometry of the representation space. Therefore, Procrustes tests whether English and Chinese hidden spaces are related mainly by a rotation or reflection.

Ridge Regression. Ridge regression learns a regularized linear mapping without the orthogonality constraint:

$$W^* = arg \|XW - Y\|_F^2 + \lambda \|W\|_F^2. \tag{2}$$

Compared with Procrustes, Ridge is more flexible because it can scale and reshape dimensions. The regularization term controls overfitting and is important because the number of training examples is relatively limited compared with the hidden-state dimension.

MLP Mapping. Finally, we use a nonlinear MLP mapper to test whether English-Chinese representation differences cannot be fully captured by a linear transformation. For a Chinese representation x , the MLP uses a residual form:

$$M_\theta(x) = x + f_\theta(x), \tag{3}$$

where f_θ is a two-layer feed-forward network:

$$f_\theta(x) = W_2 \, \text{Dropout} \left(\text{GELU}(W_1 x + b_1) \right) + b_2. \tag{4}$$

The residual connection allows the model to keep the original Chinese representation while learning a correction toward the English space. The mapper is trained with cosine-distance loss:

$$\mathscr{L} = 1 - \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \cos \left(M_\theta(x_i), y_i \right). \tag{5}$$

Before training, both English and Chinese representations are L_2 -normalized. The MLP uses a hidden size of 1024, dropout rate 0.1, AdamW optimization, and early stopping based on validation loss.

DOI: 10.54254/2755-2721/2026.GF35547

3.3. Prediction and auxiliary metrics

Prediction is evaluated at the final layer using five metrics: first-token Top-1 accuracy, first-token Top-5 accuracy, target probability, full-target exact match, and full-target geometric probability. Full-target metrics are important because Chinese targets may tokenize into multiple pieces. We also include a multilingual language-identification probe, where the model chooses between English and Chinese labels for a given sentence. This auxiliary task is interpreted cautiously because it can be sensitive to prompt wording and label tokens.

4. Results

4.1. Layer-wise alignment

Figures 1-3 show layer-wise alignment for the three models. Across models, English-Chinese alignment generally becomes stronger in deeper layers, but the shape of the trend is model-specific.

For Pythia-1.4B, alignment improves steadily from early to later layers. Ridge and MLP are stronger than No Mapping for most layers, suggesting that cross-lingual structure exists but is not fully aligned in the original hidden space. The final No Mapping score drops, which may indicate that the last hidden state becomes more specialized for output prediction.

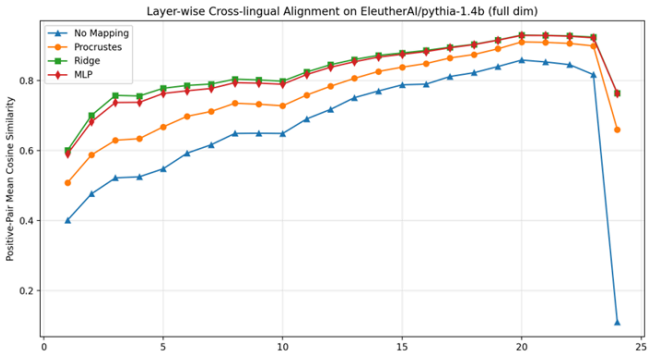

Figure 1. Layer-wise English-Chinese alignment on EleutherAI/Pythia-1.4B

For Phi-3.5-mini-instruct, alignment is strong across much of the model. MLP usually gives the highest scores, and No Mapping becomes strong in middle layers but declines later. This suggests that shared semantic structure is present, while later representations may become more language- or task-specific unless mapped.

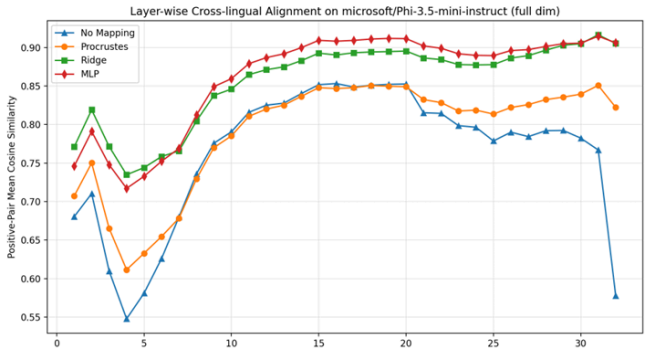

Figure 2. Layer-wise English-Chinese alignment on Microsoft Phi-3.5-mini-instruct

DOI: 10.54254/2755-2721/2026.GF35547

For Qwen3-1.7B-Base, No Mapping is already high in middle layers, indicating strong natural English-Chinese closeness. Procrustes sometimes underperforms No Mapping, which may happen when a global rotation slightly disrupts an already aligned local structure. Ridge and MLP still improve later-layer alignment, showing that flexible mappings recover additional shared structure.

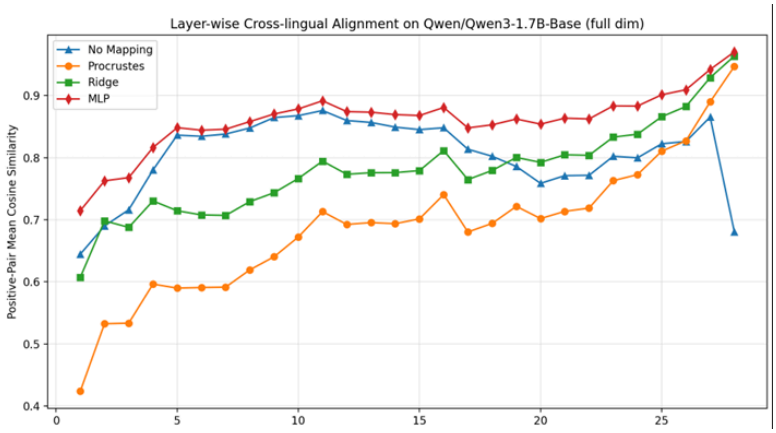

Figure 3. Layer-wise English-Chinese alignment on Qwen3-1.7B-Base

Overall, Ridge and MLP usually outperform No Mapping and Procrustes. This indicates that English and Chinese spaces are related but not perfectly overlapping. Ridge shows that much of the relation is linear, while MLP gains suggest that some cross-lingual differences may be nonlinear.

4.2. Target-word prediction

Table 1 summarizes final-layer target prediction. For English, Phi-3.5-mini-instruct achieves the strongest Top-1, Top-5, full exact match, and full geometric probability. Pythia-1.4B is lower but still reasonable, while Qwen3-1.7B-Base has lower English Top-1 and full exact match despite a strong Top-5 score. This suggests that Qwen often ranks the correct English target highly without making it the top prediction.

For Chinese, all models perform worse than in English. Phi-3.5-mini-instruct has the strongest first-token performance, but its full-target exact match remains low. Pythia-1.4B is weakest on Chinese, while Qwen3-1.7B-Base performs better than Pythia on Chinese exact-match metrics. These results suggest that Chinese target prediction remains difficult even in a controlled setup, partly because prefixes may allow multiple plausible continuations.

Table 1. Final-layer target-word prediction results

Model	Lang.	Top-1	Top-5	Prob.	Full Exact	Full Geom Prob.
Phi-3.5-mini-instruct	EN	0.565	0.918	0.333	0.565	0.425
Phi-3.5-mini-instruct	ZH	0.395	0.877	0.196	0.148	0.328
Pythia-1.4B	EN	0.425	0.822	0.169	0.425	0.173
Pythia-1.4B	ZH	0.178	0.578	0.073	0.044	0.195
Qwen3-1.7B-Base	EN	0.340	0.830	0.093	0.340	0.093
Qwen3-1.7B-Base	ZH	0.289	0.711	0.117	0.222	0.104

DOI: 10.54254/2755-2721/2026.GF35547

4.3. Language identification and alignment-performance correlation

Table 2 shows the auxiliary language-identification results. Phi-3.5-mini-instruct performs perfectly on both English and Chinese with high margins. Pythia and Qwen identify English almost perfectly but are weaker on Chinese. Because this task depends on prompt format and answer-label choice, it should be treated as a supplementary behavioral probe.

Table 2. Multilingual language-identification results. Margin is the mean log-probability difference between correct and incorrect language labels

Model	EN Acc.	ZH Acc.	EN Margin	ZH Margin
Phi-3.5-mini-instruct	1.00	1.00	3.05	4.66
Pythia-1.4B	1.00	0.80	4.08	0.49
Qwen3-1.7B-Base	0.99	0.77	3.05	1.14

To compare alignment and performance, we compute Pearson correlations across the three models between final-layer alignment and two overall prediction metrics: average English-Chinese Top-1 accuracy and average English-Chinese full exact match. As shown in Table 3, all correlations are positive, but the sample contains only three models, so the values are descriptive rather than statistically conclusive.

Table 3. Pearson correlations between final-layer alignment and target-word prediction across the three models

Alignment Method	Overall Top-1 r	Overall Full Exact r
No Mapping	0.407	0.675
Procrustes	0.140	0.446
Ridge	0.296	0.583
MLP	0.277	0.567

5. Discussion

The results support two main conclusions. First, deeper layers tend to show stronger cross-lingual structure, but this trend is not identical across models. Earlier layers may encode more local lexical features, middle layers may encode more transferable semantic information, and later layers may become specialized for prediction. The decline of No Mapping in some final layers suggests that representations can become more language-specific even when they remain alignable through learned transformations.

Second, alignment and prediction performance are related but distinct. Phi-3.5-mini-instruct performs strongly on both alignment and prediction, but Qwen3-1.7B-Base shows strong alignment without always achieving the best prediction scores. This means that geometric alignment is only one component of multilingual model behavior. Tokenization, output-head behavior, pretraining distribution, and instruction tuning can all affect whether aligned concepts are converted into the correct lexical output.

The comparison among No Mapping, Procrustes, Ridge, and MLP is also informative. Procrustes preserves geometry, so it is useful when the two spaces differ mostly by rotation. However, its constraint can be too strict when the model represents English and Chinese with different scaling or direction-specific distortions. Ridge improves alignment because it can adjust dimensions more flexibly while still remaining linear. MLP provides the strongest or near-strongest alignment because

DOI: 10.54254/2755-2721/2026.GF35547

it can model nonlinear corrections, but its advantage should be interpreted carefully because the dataset is small and nonlinear methods can overfit. The consistent strength of Ridge suggests that much of the English-Chinese relation remains approximately linear.

The prediction results also show that Chinese target prediction is harder than English target prediction for all three models. This may reflect weaker Chinese coverage in training data, tokenization differences, or ambiguity in the prefix-based prediction setup. A target word may be semantically supported by the context but not uniquely determined. Therefore, low exact-match accuracy should not be interpreted only as model failure. A useful follow-up would be a human baseline in which annotators predict the missing target from the same prefixes.

6. Limitations

This study has several limitations. First, the dataset is relatively small. Although the controlled target-word setup makes the experiment easier to interpret, the sample size limits the strength of statistical conclusions and makes the correlation analysis preliminary. A larger dataset would allow analysis by part of speech, word frequency, semantic category, and target-token length. Second, the study is limited to English-Chinese target-word prediction. English and Chinese differ strongly in script and tokenization, so the same conclusions may not hold for language pairs with closer scripts or different morphological properties. Future work should test both high-resource and low-resource language pairs. Third, the main alignment metric is positive-pair mean cosine similarity. This is useful for measuring whether corresponding English and Chinese targets are close, but it does not fully test whether the model distinguishes correct pairs from incorrect pairs. Hidden representations can also be anisotropic, which may inflate cosine similarity. Margin-based metrics comparing positive-pair and negative-pair similarity would provide a stronger test. Finally, the language-identification task is prompt-based. Because answer labels and prompt wording can affect language model outputs, this task should be viewed only as an auxiliary probe rather than a complete measure of multilingual ability.

7. Future work

Future work should expand the dataset and include more models from different families and training settings. This would make it possible to test whether the observed deeper-layer alignment trend is general or model-specific. Future experiments should also add negative-pair and retrieval-style alignment metrics, such as whether the mapped Chinese representation retrieves the correct English target among many candidates.

Another direction is to analyze language identity and semantic information separately. For example, dimensionality reduction or probing could identify which directions encode target meaning and which directions encode language identity. This would help explain why a model can show strong semantic alignment while still producing language-specific output behavior.

Finally, the relationship between alignment and performance should be evaluated on broader tasks. Target-word prediction is controlled and interpretable, but it does not fully represent multilingual ability. Bilingual sentence matching, translation retrieval, cross-lingual classification, and perplexity-based language modeling would provide a more complete evaluation.

DOI: 10.54254/2755-2721/2026.GF35547

8. Conclusion

This paper finds that English-Chinese cross-lingual alignment shows clear layer-wise trends across Pythia-1.4B, Phi-3.5-mini-instruct, and Qwen3-1.7B-Base. Learned mappings such as Ridge and MLP usually outperform No Mapping and Procrustes, suggesting that the English and Chinese spaces are related but not perfectly overlapping. Prediction results show that stronger alignment often corresponds to stronger performance, but the relationship is imperfect. Representation alignment therefore provides useful evidence about internal multilingual geometry, while final target-word prediction also depends on tokenizer behavior, output distributions, instruction tuning, and training data.

The study provides preliminary evidence that decoder-only language models can encode English and Chinese concepts in partially shared spaces, and that layer-wise alignment analysis can help explain differences in multilingual behavior across model families. At the same time, the mismatch between alignment and prediction shows that internal semantic similarity is not enough to guarantee correct lexical generation. A complete account of multilingual model behavior needs to consider both representation geometry and the output mechanisms that transform hidden states into tokens.

References

[1] Conneau, A., Lample, G., Ranzato, M. A., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. In *ICLR*.
[2] Artetxe, M., Labaka, G., & Agirre, E. (2019). Bilingual lexicon induction through unsupervised machine translation. In *ACL*.
[3] Marchisio, K., Koehn, P., & Xiong, C. (2021). An alignment-based approach to semi-supervised bilingual lexicon induction with small parallel corpora. In *Machine Translation Summit XVIII*.
[4] Shen, L., et al. (2024). The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of ACL*.
[5] Lu, T., & Koehn, P. (2025). Information access across language barriers in multilingual large language models.
[6] Pomerenke, D., Nothnagel, J., & Ostermann, S. (2025). The AI Language Proficiency Monitor: Tracking the progress of LLMs on multilingual benchmarks.
[7] Hu, Y., et al. (2025). Multilingual evaluation of large language models.
[8] Koehn, P., & Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.
[9] Lample, G., Ott, M., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Phrase-based and neural unsupervised machine translation. In *EMNLP*.
[10] Peng, Q., & Søgaard, A. (2024). Concept space alignment in multilingual LLMs. In *EMNLP*.
[11] Xu, H., & Koehn, P. (2021). Contextualized embeddings for word and word-sense alignment.
[12] Shani, N., & Basirat, A. (2025). Language dominance in multilingual large language models. In *BlackboxNLP*.
[13] Tezuka, T., & Inoue, K. (2025). Transfer neurons in multilingual large language models.
[14] Zhang, J., et al. (2026). Language-specific neurons and multilingual alignment in large language models.
[15] Zhao, H., et al. (2024). Language-specific components in multilingual language models.
[16] Xu, H., et al. (2025). Language-specific neurons in multilingual language models.
[17] Nostalgebraist. (2020). Interpreting GPT: The logit lens.
[18] Wendler, C., et al. (2024). Do large language models think in English?
[19] Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *ICDT*.
[20] Radovanović, M., Nanopoulos, A., & Ivanović, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *JMLR, 11*, 2487–2531.
[21] Biderman, S. et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of ICML*.
[22] Abdin, M. et al. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

DOI: 10.54254/2755-2721/2026.GF35547

[23] Yang, A. et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.