

Non Monotonic Cross Lingual Representational Distances for Culturally Specific Japanese Concepts in Two Transformer Models

Khyati Chaurasia

khyatichaur8909@gmail.com

St. Norbert S. S. School

Research Article

Keywords: cross-lingual representations, culturally specific concepts, hidden states, cosine distance, transformer models

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11344284/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Cross-lingual representational similarity can vary across transformer depth. This exploratory study examines pooled hidden-state representations for 20 culturally specific Japanese concepts in TinyLlama-1.1B-Chat-v1.0 and Qwen1.5-1.8B-Chat. Each concept was represented by a Japanese prompt, a literal English rendering, a Spanish translation of an English baseline, and a token-scrambled Japanese input. Cosine distances were measured relative to the concept-specific English baseline at each hidden-state index. Across the full benchmark, Japanese inputs exhibited an early decrease in mean distance followed by later re-expansion. This non-monotonic pattern was also present in comparison conditions and was not unique to intact Japanese prompts. Within the originally selected analysis ranges, Japanese distances exceeded literal English and Spanish distances for all 20 concepts in each model. Mean paired differences ranged from 0.0834 to 0.2433; all four exact paired sign-permutation comparisons remained significant after Bonferroni adjustment (adjusted $p = 7.63 \times 10^{-6}$). These findings document layer-dependent differences in representational distance within this benchmark. Because language, prompt content, and cultural framing vary together across conditions, the experiment does not establish cultural asymmetry as their independent cause or demonstrate impaired translation or cultural understanding.

1. Introduction

Multilingual language models can transfer information across languages, as shown in research on multilingual BERT and XLM-R (Devlin et al., 2019; Pires et al., 2019; Conneau et al., 2020). Such transfer does not imply that all descriptions of related situations occupy identical positions in a model's internal representation space. Kudugunta et al. (2019) found that multilingual neural machine translation representations varied with linguistic similarity and translation direction.

Cultural key words can encode meanings associated with social relationships and values (Wierzbicka, 1997). Japanese concepts such as *giri*, *wa*, and *mono no aware* therefore provide an exploratory setting for examining differently framed descriptions of related situations. Their selection here motivates the benchmark; it is not an independently validated binary classification of translatability.

Internal representations can be investigated using several approaches. Alain and Bengio (2017) used linear classifier probes to examine intermediate features, while Belinkov (2022) reviewed methodological limitations of probing. The present study uses layerwise pooled-vector distances rather than trained probes. Geometric similarity alone does not establish equivalent meaning, causal mechanisms, or downstream competence. Multilingual behavioral evaluation, such as the text-analysis tasks studied by Rathje et al. (2024), addresses a different question from hidden-state geometry.

This study examines how distances from concept-specific English baselines change across hidden-state depth in two accessible language-model checkpoints. It asks whether trajectories are non-monotonic and whether Japanese prompts have greater distances than literal English and Spanish comparison prompts over selected ranges. Cultural specificity is not independently manipulated while language and

meaning remain constant; cultural explanations are therefore hypotheses rather than established mechanisms.

2. Methods

2.1 Models and benchmark

The model checkpoints were TinyLlama/TinyLlama-1.1B-Chat-v1.0 and Qwen/Qwen1.5-1.8B-Chat. The 20-item dataset contains six relational concepts, six social-space concepts, and eight existential concepts. The complete prompts are provided in the public repository dataset, `data/culturally_asymmetric_concepts.csv`.

Each concept includes an English baseline (E_base), a Japanese prompt containing the target concept (J_asym), a literal English rendering of the Japanese prompt (E_lit), and a Spanish rendering of the English baseline (S_sym). A further condition scrambles the Japanese input's tokenizer IDs (J_scram). These conditions are paired by concept but are not necessarily semantically equivalent. For example, the giri baseline describes attending a wedding to support a friend, whereas its Japanese and literal English prompts describe fulfilling a social obligation. Spanish prompts and Japanese prompts also differ in language. These differences limit causal interpretation.

2.2 Hidden-state extraction and scrambling

The extraction procedure requested hidden states from the causal language models under `torch.no_grad()`. It averages each hidden-state tensor over sequence positions and converts the pooled vector to a CPU float32 array. Index 0 denotes the embedding output; subsequent indices denote transformer-block outputs. The saved data contain 23 indices per concept for TinyLlama and 25 for Qwen. The notebook loads model weights using `torch.float16` and `device_map="auto"`, and converts pooled output vectors to float32 for distance calculation. Inputs are supplied directly rather than through an explicit chat template. Historical package versions and checkpoint revision hashes were not pinned in the original notebook.

For the scrambled condition, Japanese text is encoded without special tokens, and `random.shuffle` rearranges the resulting token IDs. A beginning-of-sequence token is prepended where available in the full-benchmark extraction function. This procedure shuffles subword tokens rather than linguistically segmented Japanese words and can disrupt both within-word structure and order. The original extraction notebook does not record a fixed random seed for scrambling. Mean pooling does not remove language, sequence-length, or tokenization effects.

2.3 Distance measurement

At each index l , the distance between a condition vector v_l and its English-baseline vector b_l is:

$$D_l = 1 - (v_l \cdot b_l) / (\|v_l\| \|b_l\|).$$

Smaller values indicate greater cosine similarity. These values describe pooled-vector geometry; they are not direct measures of translation accuracy, cultural competence, or a manifold's volume.

2.4 Statistical analysis

Statistical analysis used `sprint_master_controlled_manifold_data.csv`, containing 960 rows: 20 concepts $\times$ (23 TinyLlama indices + 25 Qwen indices). No duplicate model–concept–index records or non-finite distance values were found.

To retain the original analysis scope, distances were averaged within each concept over indices 11–18 for TinyLlama and 12–20 for Qwen. These ranges are exploratory and were retained from the original analysis; the comparisons are not presented as preregistered confirmatory tests. Japanese-minus-control differences were calculated separately for literal English and Spanish, producing 20 paired differences for each model and comparison.

Two-sided paired sign-permutation tests enumerate all 2^{20} sign assignments under a symmetric null distribution of concept-level differences. Bonferroni adjustment covers the four reported comparisons. Percentile 95% confidence intervals for mean differences use 50,000 concept-level bootstrap samples and seed 20261012. These intervals are marginal, not simultaneous family-wise intervals. Their uncertainty reflects variation across the selected concepts, not model-training variability or a random sample of all Japanese concepts.

For the trajectory figure, 10,000 bootstrap samples resample whole concepts while preserving their indices and conditions. Bands are pointwise percentile intervals. All inferential comparisons use the concept as the paired unit rather than treating repeated indices as independent observations.

3. Results

3.1 Full-benchmark trajectories

Japanese mean distance declined sharply from the embedding output, reached its minimum at index 3 for TinyLlama (0.0132) and index 7 for Qwen (0.0143), and increased toward the final index (0.8664 and 1.0646, respectively). These minima occur relatively early in the indexed sequences.

Scrambled Japanese inputs also showed a decrease followed by re-expansion, with minima at the same respective indices and final distances of 0.8856 for TinyLlama and 1.1601 for Qwen. Therefore, the saved measurements do not support the assertion that scrambling eliminates the decrease-and-re-expansion pattern.

Spanish controls likewise showed early minima and later increases. Their final mean distances were 0.6930 for TinyLlama and 0.7232 for Qwen. Literal English distances were smaller, with final means of 0.0628 and 0.0498. Accordingly, conditions differed in magnitude, but non-monotonicity was not exclusive to culturally specific Japanese inputs.

3.2 Paired comparisons within selected ranges

All 20 concept-level differences were positive for each of the four comparisons.

Table 1

Paired differences after averaging within each concept over the original selected index ranges. Each comparison has $n = 20$ paired concepts. The unadjusted exact p-value is $2/2^{20} = 1.907 \times 10^{-6}$ for each comparison. These tests assess selected-range mean differences, not the uniqueness or mechanism of the full trajectory.

Model	Comparison	Mean paired difference	Marginal 95% bootstrap CI	Bonferroni-adjusted p
TinyLlama	Japanese – literal English	0.2433	[0.2307, 0.2580]	7.63×10^{-6}
TinyLlama	Japanese – Spanish	0.1828	[0.1694, 0.1986]	7.63×10^{-6}
Qwen	Japanese – literal English	0.1512	[0.1325, 0.1725]	7.63×10^{-6}
Qwen	Japanese – Spanish	0.0834	[0.0663, 0.1023]	7.63×10^{-6}

4. Discussion

The saved measurements support larger Japanese-to-English-baseline distances than literal-English-to-baseline and Spanish-to-baseline distances within the selected ranges in both checkpoints. This finding survives a concept-level paired reanalysis. It should not be interpreted as evidence that cultural specificity independently causes the difference: Japanese and Spanish differ in language, while the target and baseline prompts can also differ in stated motivations and content.

The full-benchmark trajectories further show that decreases followed by re-expansion occur in several conditions, including token-scrambled Japanese. Thus, non-monotonicity cannot be described as a distinctive signature of intact culturally asymmetric semantics in this experiment. Early distance minima may reflect several properties of pooled hidden states; the present measurements do not identify their cause.

The study is limited to 20 hand-curated concepts and two checkpoints. There is one described framing per condition and concept, without an independently validated translation-equivalence protocol documented in the available materials. Language, tokenization, sequence length, and content differences are not fully controlled. The selected index ranges were exploratory. Scrambling was not repeated across multiple recorded seeds, and pooled cosine distance cannot establish translation quality or preservation of cultural meaning. The statistical analysis uses the saved distance outputs; model extraction was not rerun for this reanalysis.

Future studies should include culturally specific and ordinary concepts within the same languages, matched and independently reviewed prompt pairs, several paraphrases per concept, repeated scramble

draws, and behavioral evaluations. Such designs could test whether cultural specificity explains additional distance differences beyond language and prompt content.

5. Conclusion

Within this benchmark, Japanese prompts containing culturally specific concepts had greater cosine distance from their English baselines than literal English and Spanish comparison prompts over the selected index ranges. Multiple conditions exhibited non-monotonic trajectories. The results support layer-dependent representational-distance differences in the tested checkpoints, while leaving their cultural and linguistic causes unresolved.

Declarations

Data and code availability

The prompt dataset, original extraction notebooks, and saved distance measurements are publicly available at <https://github.com/khyatichaur8909-ctrl/cultural-semantic-geometry>. The source snapshot inspected for this analysis was committed. The full-benchmark results file is `results/sprint_master_controlled_manifold_data.csv`. The older `tinylama_control_results.csv` and `qwen_control_results.csv` files contain the three-concept pilot and should not be substituted for the full benchmark. The statistical procedures used for the revised results are specified in Section 2.4 and Appendix A; they differ from the original row-level tests in the repository notebooks.

References

1. Alain, G., & Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. International Conference on Learning Representations, Workshop Track. <https://openreview.net/forum?id=HJ4-rAVtl>
2. Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1), 207–219. https://doi.org/10.1162/coli_a_00422
3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451. <https://doi.org/10.18653/v1/2020.acl-main.747>
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
5. Kudugunta, S., Bapna, A., Caswell, I., & Firat, O. (2019). Investigating multilingual NMT representations at scale. Proceedings of the 2019 Conference on Empirical Methods in Natural

Language Processing and the 9th International Joint Conference on Natural Language Processing, 1565–1575. <https://doi.org/10.18653/v1/D19-1167>

6. Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4996–5001. <https://doi.org/10.18653/v1/P19-1493>
7. Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjeh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. Proceedings of the National Academy of Sciences, 121(34), e2308950121. <https://doi.org/10.1073/pnas.2308950121>
8. Wierzbicka, A. (1997). Understanding cultures through their key words: English, Russian, Polish, German, and Japanese. Oxford University Press. <https://doi.org/10.1093/oso/9780195088359.001.0001>

Figures

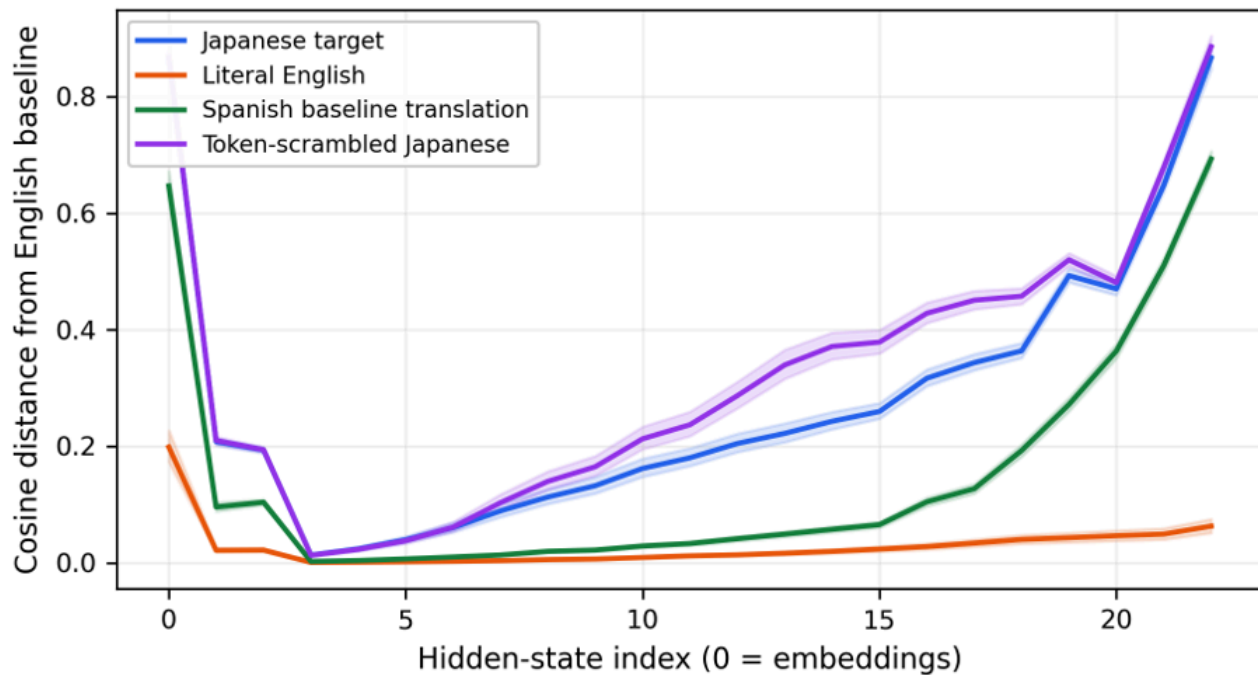


Figure 1

Full-benchmark trajectories in TinyLlama-1.1B ($n = 20$ concepts). Curves show mean cosine distance from the concept-specific English baseline; shaded bands are pointwise 95% concept-bootstrap intervals. Index 0 denotes the embedding output.

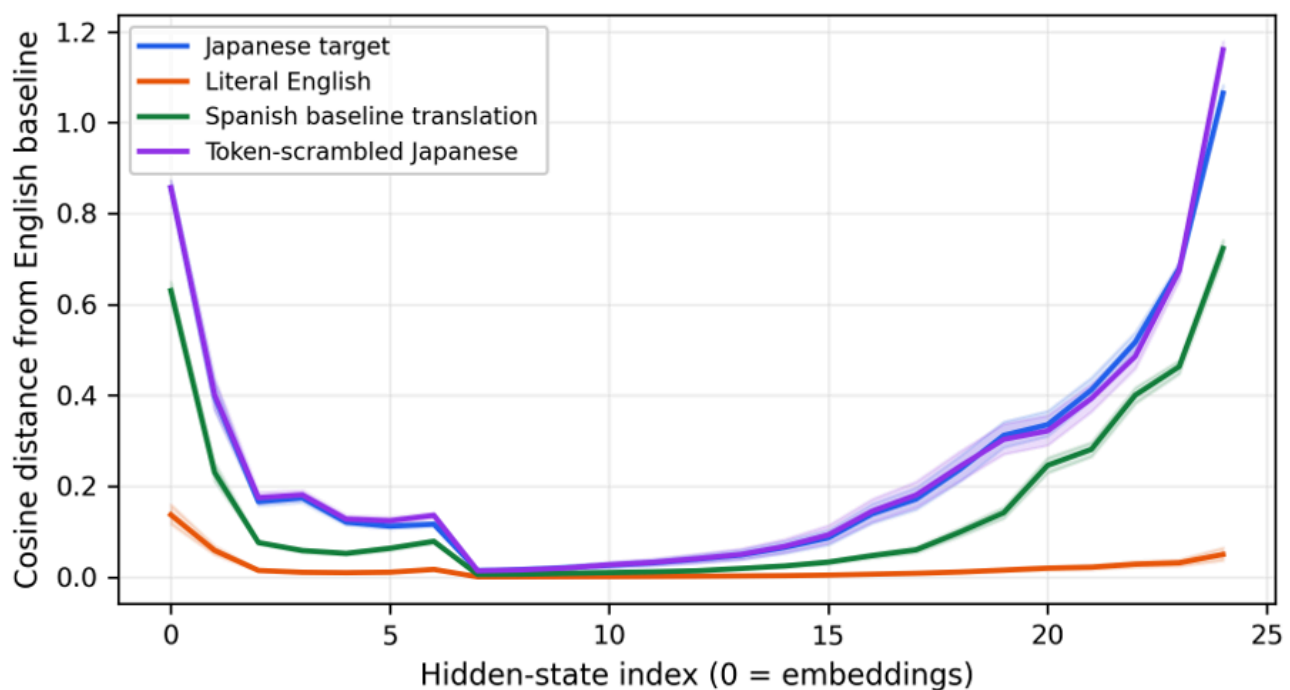


Figure 2

Full-benchmark trajectories in Qwen-1.8B ($n = 20$ concepts). Curves show mean cosine distance from the concept-specific English baseline; shaded bands are pointwise 95% concept-bootstrap intervals. Index 0 denotes the embedding output.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AppendixAComputationalspecification.docx](#)