Latent-Space Intervention for Cross-Lingual Factual Consistency: Consistency Improvements without Accuracy Drops

Anonymous ACL submission

Abstract

Large Language Models (LLMs) often answer the same factual question differently across languages. We study whether cross-lingual latent-space intervention can reduce this inconsistency. We train layer-specific autoencoders on parallel multilingual representations and apply inference-time corrections to factual QA prompts. We find that latent intervention substantially improves geometric alignment between languages, and that this improvement translates into consistent gains in cross-lingual consistency with English across both open-ended and multiple-choice QA formats, without degrading factual accuracy. In open-ended QA, Spearman's rank correlation between English and non-English languages improves substantially, with gains of up to 0.29 for English-Arabic and 0.16 for English-Russian pairs. In multiple-choice QA, answer agreement with English improves consistently across both KLAR and mParaRel. Ablations show that AE reconstruction drives these gains, while PCA-based and mean-shift subspace projection contribute marginally.

1 Introduction

Multilingual language models (MLMs) achieve strong performance on many cross-lingual tasks (Zhu et al., 2024), but factual knowledge retrieval remains uneven across languages (Jiang et al., 2020; Qi et al., 2023; Xing et al., 2024). Unlike many syntactic or semantic tasks, factual recall is closely tied to the distribution of facts in pretraining data (Liu et al., 2025): frequently observed facts are more likely to be recalled directly, whereas low-frequency facts often depend on cross-lingual transfer (Chang et al., 2024). Because pretraining corpora are highly imbalanced across languages, the same fact may be well represented in one language but rare or absent in another (Xu et al., 2023a). As a result, factual knowledge can be unevenly accessible across languages (Aggarwal et al., 2025; Goldman et al., 2025). This gives rise to cross-lingual factual inconsistency: a model may answer a factual query correctly in one language while failing or producing a different answer in another (Qi et al., 2023; Lu et al., 2025; Fierro et al., 2025).

A growing body of work suggests that part of this difficulty is representational (Ifergan et al., 2025; Lim et al., 2025; Wang et al., 2025a). MLMs encode factual queries in representations that may entangle factual content with language-specific surface form. When these factors are not cleanly separated, the same underlying fact can occupy different regions of the model's hidden space depending on the language of the query (Libovický et al., 2020; Zhao et al., 2025). Recent mechanistic analyses support this view: factual knowledge appears to be represented in relatively language-independent intermediate-layer spaces, while failures can arise during the transition to later language-specific representations (Wang et al., 2025a). Other work further suggests that semantically equivalent facts across languages can activate overlapping sets of language-agnostic factual neurons, although access to these shared mechanisms is not uniform across languages (Cao et al., 2026; Zhang et al., 2025).

This motivates a natural question: Can we construct a representation space in which factual content is more comparable across languages, and use it to test whether stronger alignment improves factual consistency? Such a space would serve two purposes. First, it would provide a diagnostic view of factual representations with reduced language-specific variation. Second, it would offer an intervention point for applying lightweight inference-time corrections without retraining the model or requiring language-specific factual supervision.

We propose that a cross-lingual autoencoder latent space, trained on parallel multilingual text, provides exactly this. By mapping LLM representations into a compact shared latent space where

the same content across languages is brought into alignment, we obtain a language-agnostic view of the model's internal representations (Peng et al., 2025). Within this space, we identify a factual inconsistency subspace representing directions of maximum cross-lingual divergence for the same facts, and apply lightweight inference-time corrections via forward hooks without modifying model weights. In other words, we project out the cross-lingual, latent inconsistency subspace via PCA. We evaluate on KLAR (Wang et al., 2025a) and mParaRel (Fierro and Søgaard, 2022) across typologically diverse languages in both multiple-choice and open-ended factual QA formats. Our findings show that latent-space intervention improves cross-lingual representation alignment and answer consistency without incurring a performance cost.

Contributions: (1) We show that autoencoder-based latent-space intervention consistently improves cross-lingual representation alignment, with larger gains for typologically distant languages and at middle-to-late layers. **(2)** We show that this alignment improvement translates into higher semantic consistency when projecting out the dimensions along which semantic inconsistencies occur. Question-answering accuracy is largely unaffected by this intervention.

2 Related Work

Several works correct cross-lingual gaps without retraining. Wang et al. (2025b) learn alignment matrices between language representations, and Kirtane and Huang (2026) apply steering vectors via forward hooks for multilingual in-context learning. A long line of work shows that shared latent spaces improve cross-lingual transfer (Conneau et al., 2020; Chi et al., 2021). More directly, Xu et al. (2023b) propose parameter-free projection modules that convert non-English representations into English-like equivalents to improve factual knowledge retrieval, but assume English as a fixed pivot. Our method requires no weight modification and no pivot language assumption.

Factual knowledge in MLMs is unevenly distributed across languages due to pretraining data imbalance (Jiang et al., 2020; Kassner et al., 2021), and models remain inconsistent across languages even for paraphrases of the same fact (Fierro and Søgaard, 2022). Wang et al. (2025a) provide a mechanistic account: MLMs encode facts in a language-independent concept space through most layers but fail during the final transition to language-specific representations, causing inconsistent predictions. Lim et al. (2025) and Tezuka and Inoue (2025) further confirm this through activations and latent process analysis. These works are diagnostic; we build on their benchmarks and metrics to evaluate a mitigation method.

3 Our Approach

Given a multilingual language model $\mathcal{M}$ and a set of factual relations $\mathcal{R}$, we consider a set of facts $\mathcal{F} = \{(s, r, o)\}$ where s is a subject, $r \in \mathcal{R}$ is a relation, and o is the correct object. Each fact is expressed as a prompt across a set of languages $\mathcal{L} = \{l_1, \ldots, l_n\}$, yielding language-specific prompts $\{x^l\}_{l \in \mathcal{L}}$ that ask the same factual question in different languages. We define $\mathcal{M}$ as *cross-lingually inconsistent* on a fact (s, r, o) if it predicts different answer choices for the same fact across languages, i.e., $\hat{y}^{l_i} \neq \hat{y}^{l_j}$ for some $l_i, l_j \in \mathcal{L}$, where $\hat{y}^l$ denotes the model's predicted choice for the prompt in language l. Our goal is to construct a cross-lingual latent space in which inconsistency can be measured and mitigated at inference time, without modifying model weights.

Cross-lingual Latent Space. A key challenge is that LLM representations conflate factual content with language-specific surface signals. Directly comparing or manipulating representations across languages in the LLM's native space is therefore noisy. Following Peng et al. (2025), we address this by mapping LLM representations into a compact cross-lingual latent space where the same content across languages is brought into closer alignment.

Specifically, we train an autoencoder (AE) in two phases. First, we optimize a shared encoder and decoder jointly. Given a parallel pair (h^{l_1}, h^{l_2}) of LLM hidden states for semantically equivalent sentences in languages l_1 and l_2, with latent representations $z^{l_i} = \text{enc}(h^{l_i})$ and reconstructions $\hat{h}^{l_i} = \text{dec}(z^{l_i})$, the training objective is:

$$\mathcal{L} = \zeta\, \mathcal{L}_{\text{self}} + \beta\, \mathcal{L}_{\text{align}} + \eta\, \mathcal{L}_{\text{var}} + \rho\, \mathcal{L}_{\text{cov}} \quad (1)$$

where the self-reconstruction loss

$$\mathcal{L}_{\text{self}} = \ell(\hat{h}^{l_1}, h^{l_1}) + \ell(\hat{h}^{l_2}, h^{l_2}) \quad (2)$$

encourages faithful reconstruction within each language. The alignment loss

$$\mathcal{L}_{\text{align}} = \text{MSE}(\tilde{z}^{l_1}, \tilde{z}^{l_2}), \quad \tilde{z}^{l_i} = \frac{z^{l_i}}{\|z^{l_i}\|} \quad (3)$$

directly minimizes the distance between normalized latent representations of parallel pairs. The variance loss

$$\mathcal{L}_{\text{var}} = \sum_j \text{ReLU}(\gamma - \text{Std}_j(Z)) \quad (4)$$

prevents dimensional collapse by penalizing dimensions whose standard deviation falls below a threshold γ. The covariance loss

$$\mathcal{L}_{\text{cov}} = \frac{1}{k} \sum_{i \neq j} [\text{Cov}(Z)]_{ij}^2 \quad (5)$$

reduces redundancy by penalizing off-diagonal entries of the latent covariance matrix. We use the Huber (Huber, 1992) loss for ℓ.

In the second phase, we freeze the shared encoder and train language-specific decoders $\{\text{dec}_l\}_{l \in \mathcal{L}}$ independently, one per language, minimizing $\ell(\text{dec}_l(z^l), h^l)$ for each language l. This allows each decoder to specialize in reconstructing its target language's representations from the shared language-agnostic latent space, while preserving the cross-lingual alignment learned in phase one.

Inference-time Correction. At inference time, we register a forward hook on a chosen transformer layer. When the model processes a prompt, the hook intercepts the hidden states $\mathbf{X} \in \mathbb{R}^{B \times T \times d}$, computes a mean-pooled sentence representation $\mathbf{h} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t \in \mathbb{R}^d$, and maps it to the latent space via the AE encoder: $\mathbf{z} = \text{enc}(\mathbf{h})$. We explore three correction strategies in the latent space.

The *identity* variant applies no correction; it simply encodes and decodes the representation without modification, serving as a baseline that isolates the effect of the AE bottleneck itself from any explicit correction: $\tilde{\mathbf{z}} = \mathbf{z}$. The *projection* variant identifies an inconsistency subspace via PCA on cross-lingual difference vectors and projects it out:

$$\tilde{\mathbf{z}}^l = \mathbf{z}^l - \sum_{j=1}^{m} (\mathbf{z}^l \cdot \mathbf{d}_j)\, \mathbf{d}_j \quad (6)$$

where $\mathbf{d}_1, \ldots, \mathbf{d}_m$ are the top-m principal components of the inconsistency subspace.

The *mean-shift* variant pulls each language's representation toward the cross-lingual mean without assuming any pivot language:

$$\tilde{\mathbf{z}}^l = \mathbf{z}^l + \lambda \left(\bar{\mathbf{z}} - \bar{\mathbf{z}}^l \right) \quad (7)$$

where $\bar{\mathbf{z}}^l = \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i^l$ is the mean representation of language l across all N facts in our evaluation sets and $\bar{\mathbf{z}} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \bar{\mathbf{z}}^l$ is the mean across all languages.

In all three cases, the corrected latent vector $\tilde{\mathbf{z}}$ is decoded back to the LLM representation space via the language-specific decoder $\text{dec}_l(\tilde{\mathbf{z}})$, and the difference between the decoded representation and the original mean-pooled hidden state is added as a residual to all token positions:

$$\mathbf{x}'_t = \mathbf{x}_t + \big(\text{dec}_l(\tilde{\mathbf{z}}) - \mathbf{h}\big), \quad \forall t \in \{1, \ldots, T\} \quad (8)$$

4 Experimental Setup

Datasets. For autoencoder training, we use parallel sentences from TED talk scripts (Salesky et al., 2021), and evaluate latent space alignment quality on *FLORES+* (Goyal et al., 2022). We use a subset of five typologically diverse languages: Arabic, English, Dutch, Russian, and Chinese. We evaluate on two multilingual factual knowledge benchmarks. *KLAR* (Wang et al., 2025a) is a cross-lingual factual consistency dataset, with prompts available in 17 languages. *mParaRel* (Fierro and Søgaard, 2022) is a multilingual extension of ParaRel (Elazar et al., 2021) containing paraphrase sets for relational facts across 45 languages. For both datasets, we convert facts into multiple-choice questions with four candidates, one correct answer, and three distractors drawn from the same relation, ensuring a consistent evaluation format across datasets and languages [1]. For more details on the datasets, see Appendix A.

Model. We use Aya-expanse-8B (Dang et al., 2024), a state-of-the-art open-source multilingual instruction-tuned language model covering 23 languages. Appendix B provides additional details about the models and hyperparameters.

Evaluation Metrics. We evaluate both the learned latent space and the downstream factual interventions. First, on FLORES+ parallel sentences, we assess the autoencoder before and after reconstruction using t-SNE, cosine similarity, and parallel-sentence retrieval accuracy. For factual QA, we report three metrics. *Cosine similarity to English* measures representation-level alignment between each non-English language and English. *Factual accuracy* measures the proportion of examples answered correctly. *Agreement* for multiple-choice QA measures how often two languages produce the same answer, regardless of whether that

[1] We do not evaluate open-ended QA on mParaRel, as its cloze-style prompts are designed for masked language models and do not place the answer at the end.

	Intervention	KLAR OE		KLAR MC		mParaRel MC	
		Acc	Rank Corr (en)	Acc	Agr (en)	Acc	Agr (en)
Raw	None	39.52	6.68	**85.59**	85.09	**85.91**	86.70
	+ Mean-shift	40.04	5.47	**85.59**	85.01	81.09	76.02
	+ PCA	39.51	6.53	85.56	84.97	85.86	86.29
Identity (AE)	None	**41.57**	**26.35**	85.32	85.43	**86.35**	**86.97**
	+ Mean-shift	41.18	8.49	**85.42**	**85.41**	85.91	86.68
	+ PCA	41.52	23.55	85.29	85.43	**86.37**	**86.97**

Table 1: Average accuracy (%) and consistency with English across strategies at the best layer per setting. Rank Corr (en) is the average Spearman rank correlation ($\times 100$) across English-paired language pairs. Agr (en) is the average answer agreement (%) with English. Bold indicates best result per column.

	KLAR OE Rank Corr				KLAR MC Agreement				mParaRel MC Agreement			
	ar-en	en-nl	en-ru	en-zh	ar-en	en-nl	en-ru	en-zh	ar-en	en-nl	en-ru	en-zh
Raw	2.59	16.28	4.13	3.70	66.17	94.65	90.87	88.66	84.31	90.06	87.33	85.09
Identity	**31.73**	**32.62**	**26.70**	**14.37**	66.36	**95.23**	**91.22**	**88.89**	**84.80**	**90.55**	**87.43**	85.19
+ PCA	29.36	31.73	25.61	7.49	66.44	95.19	91.22	88.85	**84.80**	90.45	**87.43**	85.19
+ Mean-shift	5.03	20.45	1.93	6.56	**66.82**	**95.23**	91.18	88.81	84.30	90.90	86.71	**85.80**

Table 2: Pairwise English consistency before and after intervention at the best layer per setting. Rank Corr values are Spearman rank correlation ($\times 100$). Agreement values are answer agreement (%).

answer is correct. For open-ended QA, we measure cross-lingual consistency as the Spearman correlation between the rank of the correct answer compared to the ranks of the distractors (Rank Corr).

5 Results

Across all layers and languages, AE-based intervention consistently improves cross-lingual representation alignment, with the largest gains for typologically distant languages and at later layers where raw representations become increasingly language-specific. A full evaluation of latent space alignment quality, is provided in Appendices C and D.

Tables 1 and 2 report our main results. We organize findings around three questions: whether the intervention improves consistency with English, whether factual accuracy is preserved, and which intervention strategy drives the gains. In open-ended QA, Spearman's rank correlation with English improves substantially at layer 24. AE-only intervention (identity) raises average rank correlation with English from 6.68 to 26.35, with the largest gains for English-Arabic (2.59 $\rightarrow$ 31.73) and English-Dutch (16.28 $\rightarrow$ 32.62). In multiple-choice QA, answer agreement with English improves consistently across both datasets. For KLAR at layer 28, average English-pair agreement rises from 85.09 to 85.43. For mParaRel at layer 20, it rises from 86.70 to 86.97. While gains in multiple-choice are modest, they are consistent across all English-paired language pairs.

Factual accuracy remains stable across all intervention strategies and datasets. In open-ended QA, identity improves average accuracy from 39.52 to 41.57 at layer 24. In multiple-choice QA, accuracy changes by less than 0.3 points in all conditions, confirming that consistency gains do not come at the cost of factual performance. Identity and Identity+PCA perform nearly identically across all settings, while Identity+Mean-shift, PCA, and Mean-shift contribute little or no improvement in consistency with English. In open-ended QA, Mean-shift yields an average rank correlation of only 5.47 with English, barely above baseline, despite achieving comparable accuracy. These results indicate that the cross-lingual alignment induced by the AE mapping is the operative mechanism, and that explicit inconsistency removal via subspace projection or mean-shift adds negligible value on top of it.

6 Conclusion

We investigated whether inference-time intervention in a cross-lingual latent space can reduce factual inconsistency in MLMs. Our results across KLAR (Wang et al., 2025a) and mParaRel (Fierro and Søgaard, 2022) show that AE-based intervention improves cross-lingual representation alignment, particularly for typologically distant languages and at middle-to-late layers, and translates into improved cross-lingual consistency.

Limitations

Our study has several limitations that also point to useful directions for future work. First, we evaluate on a focused set of languages and factual relations from KLAR, as well as on mParaRel. This controlled setting allows us to analyze cross-lingual factual inconsistency in detail, but broader evaluation on additional benchmarks, languages, and relation types is needed to test how general the observed patterns are.

Second, our experiments use a single multilingual LLM, Aya-expanse-8B. This makes the layer-wise analysis internally consistent, but the behavior of latent-space intervention may vary across model families, model scales, and pretraining mixtures. Extending the analysis to additional multilingual LLMs would clarify which findings are model-specific and which reflect more general properties of multilingual factual recall.

Third, our interventions are intentionally lightweight and inference-time only. PCA-based removal and mean-shift correction provide simple and interpretable ways to manipulate latent representations, but factual inconsistency may be distributed across relation-specific or nonlinear directions. More expressive correction methods may therefore yield stronger factual gains while preserving the interpretability of the current framework.

Finally, our results show that latent-space intervention improves geometric alignment more reliably than factual accuracy or agreement. Rather than treating this as a failure, we view it as an important finding: representation alignment is not sufficient by itself to guarantee factual consistency. Future work can build on this distinction by developing interventions that explicitly preserve answer-relevant information while reducing cross-lingual divergence.

References

Tushar Aggarwal, Kumar Tanmay, Ayush Agrawal, Kumar Ayush, Hamid Palangi, and Paul Pu Liang. 2025. Language models' factuality depends on the language of inquiry. *arXiv preprint arXiv:2502.17955*.

Pengfei Cao, Yuheng Chen, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2026. One mind, many tongues: A deep dive into language-agnostic knowledge neurons in large language models. *Artificial Intelligence*, page 104490.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. How do large language models acquire factual knowledge during pretraining? *Advances in neural information processing systems*, 37:60626–60668.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Ed Hovy, Hinrich Schutze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *ArXiv*, abs/2102.01017.

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. How do multilingual language models remember facts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16052–16106, Vienna, Austria. Association for Computational Linguistics.

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani,

et al. 2025. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer. *arXiv preprint arXiv:2502.21228*.

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.

Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer.

Maxim Ifergan, Leshem Choshen, Roee Aharoni, Idan Szpektor, and Omri Abend. 2025. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in llms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4630–4644.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Neeraja Kirtane and Kuan-Hao Huang. 2026. Language steering for multilingual in-context learning. *arXiv preprint arXiv:2602.02326*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. 2025. Language-specific latent process hinders cross-lingual performance. *arXiv preprint arXiv:2505.13141*.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schütze. 2025. Tracing multilingual factual knowledge acquisition in pretraining. *arXiv preprint arXiv:2505.14824*.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. 2025. Paths not taken: Understanding and mending the multilingual factual recall pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15077–15107.

Qiwei Peng, Guimin Hu, Yekun Chai, and Anders Søgaard. 2025. Debiasing multilingual LLMs in cross-lingual latent space. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22582–22593, Suzhou, China. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 10650–10666. Association for Computational Linguistics.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation.

Hinata Tezuka and Naoya Inoue. 2025. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31742–31792, Suzhou, China. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025a. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025b. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5418–5433, Vienna, Austria. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xiaolin Xing, Zhiwei He, Haoyu Xu, Xing Wang, Rui Wang, and Yu Hong. 2024. Evaluating knowledge-based cross-lingual inconsistency in large language models. *arXiv preprint arXiv:2407.01358*.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023a. Language representation projection: Can we transfer

factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023b. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702, Singapore. Association for Computational Linguistics.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. Multilingual knowledge editing with language-agnostic factual neurons. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5775–5788, Abu Dhabi, UAE. Association for Computational Linguistics.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, et al. 2025. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. *arXiv preprint arXiv:2505.15257*.

Shaolin Zhu, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, Deyi Xiong, et al. 2024. Multilingual large language models: A systematic survey. *arXiv preprint arXiv:2411.11072*.

A Data Details

We train the AE on parallel multilingual sentence pairs sourced from TED Talk transcripts (Salesky et al., 2021) across five languages: Arabic (ar), English (en), Dutch (nl), Russian (ru), and Chinese (zh). After deduplication, each language contains 212,698 unique sentences, yielding 2,126,980 parallel pairs across all ten language combinations. The data is split into train, development, and test sets. For evaluation of the cross-lingual latent space, we use 500 parallel sentences per language from the FLORES+ benchmark (Goyal et al., 2022).

For KLAR (Wang et al., 2025a), we use 2,619 facts spanning 20 factual relations (e.g. capital cities, places of birth, nationalities) across five languages: Arabic, English, Dutch, Russian, and Chinese. For each fact, one prompt template is selected and kept consistent across all languages to ensure comparable evaluation conditions.

For mParaRel (Fierro and Søgaard, 2022), we use 1,026 facts spanning 24 relations across the same five languages. As with KLAR, one template per fact is randomly selected and held fixed across languages. For both datasets, each fact is paired with four answer candidates, one correct object and three distractors drawn from the same relation, used for RankC evaluation.

B Experimental Details.

The AE encoder and decoders are three-layer MLPs with a latent dimension of $k = 256$. We train using the AdamW optimizer with a learning rate of 10^{-4} and weight decay of 10^{-8}, and a ReduceLROnPlateau scheduler with a reduction factor of 0.5, patience of 5, and minimum learning rate of 10^{-7}. Both training phases use a batch size of 1024, a maximum of 128 epochs, early stopping with patience of 16, and a maximum sequence length of 128 tokens. The loss weights are set to $\zeta = 1$, $\beta = 0.1$, $\eta = 1$, $\rho = 1$, and $\gamma = 1$. We train separate autoencoders for each transformer layer using mean-pooled hidden states extracted from the corresponding layer. At inference time, the forward hook is registered on each transformer layer output, using mean pooling over token positions. The correction is applied only during the prefill step and skipped for single-token autoregressive generation steps. We evaluate three correction strategies, *identity* (encode-decode only), *PCA projection*, and *mean-shift*, across layers $\{4, 8, 12, 16, 20, 24, 28, 31\}$, with $\alpha = 1.0$ and $n = 1$ principal component for PCA.

All experiments were conducted on a single NVIDIA A100 GPU (40/80 GB). We used the HuggingFace Transformers library (Wolf et al., 2020) with bf16 precision throughout. For each of the 8 evaluated layers, we extracted hidden state representations from Aya-expanse-8B (Dang et al., 2024), trained a separate autoencoder, and trained language-specific decoders independently, amounting to approximately 5 hours per layer. The autoencoder encoder and decoder are three-layer MLPs with hidden dimensions 1024, 512, and 256. In total, training across all layers required approximately 32 hours of GPU time.

C Latent Space Alignment on FLORES+

Figure 1 shows t-SNE visualizations of FLORES+ parallel sentence representations (Goyal et al., 2022) in the raw LLM space and the AE latent space at layers 20, 24, and 28. In the raw LLM space, language-specific clustering becomes increasingly pronounced at deeper layers, with languages forming nearly fully separated clusters at

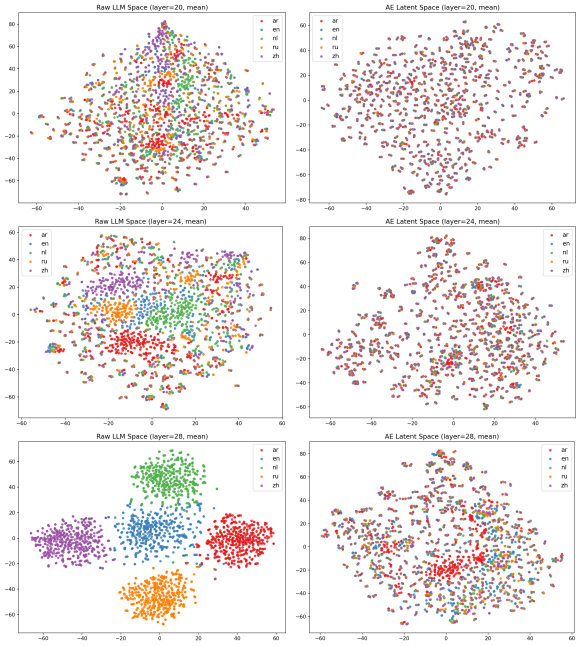

Figure 1: t-SNE visualizations of FLORES+ parallel sentence representations in the raw LLM space (left) and AE latent space (right) at layers 20, 24, and 28. Colors indicate languages.

L28. After mapping to the AE latent space, representations of parallel sentences across all five languages collapse into a single shared point cloud, confirming that the autoencoder successfully removes language-specific surface signals.

At L20, raw representations are already partially mixed across languages, and the AE provides modest gains in P1 (+0.038) with negligible cosine change, suggesting that at this layer the cross-lingual structure is partially present in the raw space and the AE refines rather than reconstructs it. At L24, language separation becomes more visible in the raw space, and the AE produces a clearer improvement in cosine similarity (+0.069), indicating that deeper layers require stronger correction to achieve alignment. At L28, where raw representations form fully separated language clusters, the AE achieves the largest gains in both cosine similarity (+0.150) and P1 (+0.042), demonstrating that the latent space can recover meaningful cross-lingual alignment even when the raw LLM space has become strongly language-specific. Together, these results confirm that the AE latent space provides consistent cross-lingual alignment across layers, with greater impact where language-specific signals are strongest.

D Additional Results.

Cross-lingual Latent Space Alignment The first row of Figure 2 shows cosine similarity to English before and after intervention across layers for open-ended and multiple-choice formats, respectively. Latent-space intervention consistently improves cross-lingual alignment across all languages and layers, with the gap between before and after being most pronounced at later layers, where unintervened representations diverge sharply from English. The improvement is largest for Chinese and Arabic, the most typologically distant languages in our set, while Dutch shows the smallest gain, consistent with its closer typological proximity to English. At middle layers (L8–L24), post-intervention similarities converge across languages, suggesting the AE latent space successfully reduces language-specific variation in this range. At the final layers (L28–L31), alignment degrades even after intervention, indicating that the AE cannot fully compensate for the strong language-specific signals that emerge in the model's final representations.

Factual Accuracy Bottom row of Figure 2 shows first-token accuracy before and after identity intervention across layers for open-ended and multiple-choice QA, respectively. For open-ended QA, intervention at early layers (L4–L8) substantially degrades accuracy across all languages, with the largest drops observed for English and Dutch. As the intervention layer increases, accuracy recovers and at middle-to-late layers (L20–L24) most languages modestly exceed their baseline. English and Dutch show the clearest gains, while Arabic remains at or below baseline throughout, and Russian shows negligible improvement. These results suggest that latent-space intervention can improve open-ended factual recall, but only when applied at the right layer. The next plot, which is for multiple-choice QA, shows a similar pattern of early-layer degradation, with accuracy recovering toward baseline at middle-to-late layers. However, unlike open-ended QA, no language consistently exceeds its baseline after intervention. Multiple-choice accuracy is already high, leaving little room for improvement, and the intervention neither meaningfully helps nor hurts at middle-to-late layers. The contrast with open-ended QA suggests that factual recall in the open-ended setting is more sensitive to representational intervention than structured answer selection.

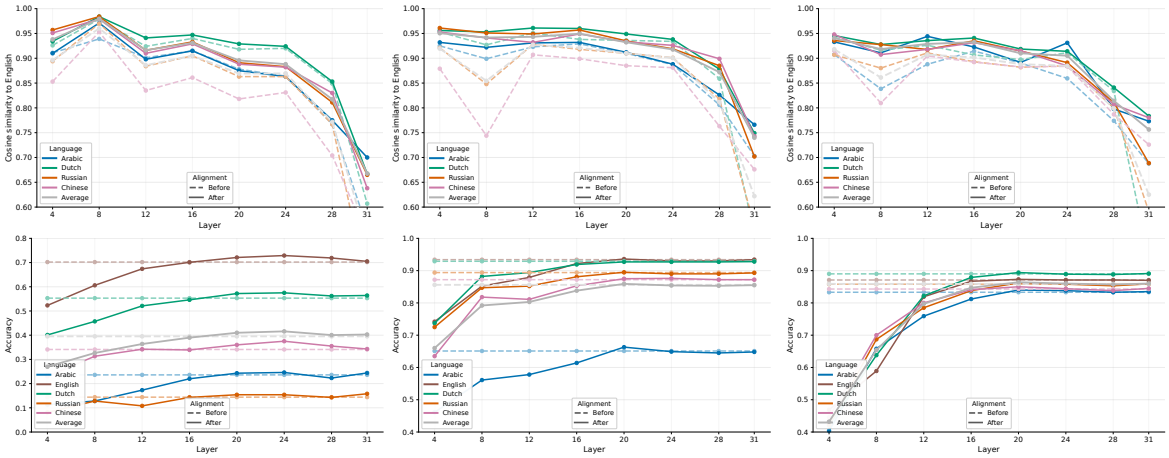

Figure 2: First row: Cosine similarity to English before and after latent-space intervention across layers for open-ended and multiple-choice QA (KLAR and mParaRel). Second row: First-token accuracy (%) before and after latent-space intervention across layers for open-ended and Multiple-choice QA (KLAR and mParaRel). Dashed lines show before intervention; solid lines show after intervention.

	ar	en	nl	ru	zh
ar	–	29.36	20.76	25.77	11.94
en	2.59	–	31.73	25.61	7.49
nl	31.96	16.28	–	21.41	14.07
ru	36.34	4.13	33.35	–	8.65
zh	36.72	3.70	25.78	28.33	–

Table 3: Pairwise rank correlation for KLAR open-ended QA at layer 24. Lower triangle shows rank correlation before intervention; upper triangle shows rank correlation after intervention.

	ar	en	nl	ru	zh
ar	–	66.4	66.5	67.4	65.4
en	66.2	–	95.2	91.2	88.8
nl	66.1	94.7	–	91.2	88.9
ru	66.9	90.9	90.8	–	88.2
zh	64.7	88.7	88.8	87.8	–

Table 4: Pairwise cross-lingual answer agreement for KLAR multiple-choice QA at layer 28. Lower triangle shows agreement before intervention; upper triangle shows agreement after intervention.

Cross-lingual Consistency Table 3 reports pairwise rank correlation for KLAR open-ended QA at layer 24. While the average pairwise rank correlation changes modestly from 21.92 to 19.68, the intervention substantially improves consistency for several English-centered pairs, including Arabic–English (+26.77), English–Russian (+21.48), and English–Dutch (+15.45). This is important because English often serves as the strongest factual anchor in multilingual LLMs. At the same time, gains are not uniform: several non-English pairs, especially those involving Chinese, decrease after intervention. Overall, these results suggest that latent-space intervention can strengthen consistency with the high-resource English factual space, but does not uniformly align all language pairs in open-ended generation. Table 5 reports pairwise answer agreement for mParaRel multiple-choice QA at layer 20. Agreement increases consistently after intervention, with average pairwise agreement rising from 86.31 to 86.68. The gains are small but broadly distributed across language pairs, with the largest increases for Arabic–Chinese (+0.88), Russian–Chinese (+0.68), Arabic–Dutch (+0.49), and Dutch–Chinese (+0.49). Table 4 reports detailed, pairwise cross-lingual consistency for multiple-choice QA at layer 28, for the KLAR dataset. Agreement increases slightly and consistently across all language pairs after intervention.

	ar	en	nl	ru	zh
ar	–	84.80	86.26	85.28	85.19
en	84.31	–	90.45	87.43	85.19
nl	85.77	90.06	–	88.69	86.94
ru	85.19	87.33	88.69	–	86.55
zh	84.31	85.09	86.45	85.87	–

Table 5: Pairwise cross-lingual answer agreement for mParaRel multiple-choice QA at layer 20. Lower triangle shows agreement before intervention; upper triangle shows agreement after intervention.

E Use of Artifacts

All artifacts used in this work are openly available and licensed for research use. We use artifacts in

accordance with their intended purpose and license terms. This work does not involve human subjects, crowd sourcing, or the collection of personal data.

F AI Usage Statement

We used AI assistance in several stages of this work. Claude (Anthropic) was used as a writing assistant to help draft, revise, and improve the clarity of the paper text. All AI-generated text was reviewed, edited, and verified by the authors. All experimental results, analyses, and scientific claims are the sole responsibility of the authors.