Exploring Language-Agnosticity in Function Vectors: A Case Study in Machine Translation

Nurkhan Laiyk[♡*], **Gerard I. Gállego**[◇¶], **Javier Ferrando**[⋆], **Fajri Koto**[♡]
[♡]Mohamed bin Zayed University of Artificial Intelligence
[◇]Universitat Politècnica de Catalunya
[¶]Cantina Labs

Abstract

Function vectors (FVs) are vector representations of tasks extracted from model activations during in-context learning. While prior work has shown that multilingual model representations can be language-agnostic, it remains unclear whether the same holds for function vectors. We study whether FVs exhibit language-agnosticity, using word-level machine translation as a case study. Across three decoder-only multilingual LLMs, we find that translation FVs extracted from word-to-word prompts in a single English→X direction transfer to other target languages, consistently improving the rank of correct translation tokens across multiple unseen languages. We further find that the highest-gain tokens span multiple languages and that translation FVs across directions share most of their top-ranked heads, indicating that the FV encodes a largely language-agnostic translation signal.

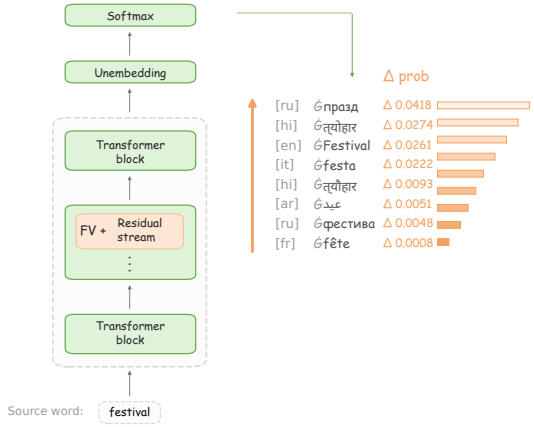


Figure 1: **Example of FV effects on the next-token distribution.** For a single source word, we show the top tokens whose probability (Δp) increases the most after FV, highlighting that the largest gains often correspond to plausible translations across multiple languages.

1 Introduction

Large language models (LLMs) have become central to multilingual NLP, showing strong performance across a wide range of cross-lingual and multilingual tasks (Brown et al., 2020). A growing body of mechanistic work has begun to ask how these models represent and process information across languages (Resck et al., 2025; Ferrando et al., 2024). One recurring finding is that many internal representations are *language-agnostic*: they encode task-relevant content, such as semantic meaning, independent of language identity, so that similar concepts in different languages are mapped to nearby states in activation space.

This property has been documented at several levels of abstraction. At the representation level, work on multilingual BERT (Devlin et al., 2019) showed that substantial cross-lingual alignment can emerge without explicit parallel supervision, enabling zero-shot transfer across languages and scripts (Pires et al., 2019; Wu and Dredze, 2019). More recent studies on autoregressive LLMs have confirmed that concept representations can be largely language-agnostic, with intermediate activations partially separating "what is being said" from "in which language" it is expressed (Wendler et al., 2024; Dumas et al., 2025). Beyond static representations, language-agnosticity has also been observed at the level of grammatical concepts (Brinkmann et al., 2025) and at the level of circuits, where prior work suggests that some internal mechanisms are shared across languages (Ferrando and Costa-jussà, 2024; Zhang et al., 2025b).

Building on this line of work, we ask whether *function vectors* exhibit the same language-agnostic property. Function vectors (FVs), introduced by Todd et al. (2024), are compact additive representations of tasks extracted from in-context

*Correspondence to: nurkhan.laiyk@mbzuai.ac.ae.
⋆This work is not related to the author's position at Amazon.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 23840–23851
October 24-29, 2026 ©2026 Association for Computational Linguistics

learning demonstrations. Recent work has shown them to be a central mechanism behind in-context learning (Yin and Steinhardt, 2025) and a useful analytical lens on model behavior more broadly (Jiang et al., 2025), making them a natural object through which to study how multilingual capabilities are internally represented. Using machine translation as a case study, we extract translation FVs from English→X word-level prompts in three multilingual decoder-only LLMs and apply them as residual-stream interventions. We find that FVs extracted from a single English→X direction promote correct translations across multiple target languages (Figure 1), and we support the cross-lingual nature of this effect through high-gain token analysis and substantial overlap in the selected FV heads across translation directions.[1]

2 Related work

Interpretability of machine translation in LLMs. Earlier work on understanding translation in neural models focused on observational analyses such as neuron activation patterns and representation probing (Mu et al., 2024; Tang et al., 2024; Wendler et al., 2024). More recent studies have moved toward causal, circuit-level analyses that identify the specific components responsible for translation behavior. Zhang et al. (2025a) study word-level translation in decoder-only LLMs and find that the behavior depends on a sparse subset of attention heads and MLPs with distinct functional roles, and that fine-tuning these components improves translation while preserving general capabilities. Lasnier et al. (2026) extend this to sentence-level few-shot MT, using activation patching to decompose translation into two separable subtasks, target language identification and sentence equivalence, each implemented by sparse attention-head sets, and use the identified heads to build steering vectors for instruction-free MT.

Language-agnosticity in multilingual LLMs. Dumas et al. (2025) use activation patching in word-to-word translation to show that concept and language information can be manipulated independently, pointing to a shared conceptual space. Brinkmann et al. (2025) extend this to grammatical structure, showing that morphosyntactic properties such as number, gender, and tense are often encoded in feature directions shared across typologically diverse languages. At the circuit level, Ferrando and Costa-jussà (2024) find that subject-verb agreement in English and Spanish relies on highly similar circuits, including a language-independent subject number direction that transfers across languages through activation steering. Fierro et al. (2025) show that some factual recall mechanisms generalize across languages while the final answer extraction step remains language-specific.

3 Methodology

Function vectors We briefly review the function vector framework of Todd et al. (2024), on which our analysis builds. For each task t, we assume a dataset of ICL prompts P_t and a corrupted version $\tilde{P}_t$ formed by shuffling demonstration labels. The procedure first selects a small set of attention heads using activation patching. For a given task t, it compares prompts from P_t (correct demonstrations) to corrupted prompts $\tilde{p} \in \tilde{P}_t$. For each head (ℓ, j), the head output on a corrupted prompt $\tilde{p}$ is patched with the head's mean output over $p \in P_t$, denoted $\bar{a}^t_{\ell j}$. The causal indirect effect (CIE) of head (ℓ, j) is defined as the resulting change in the probability of the correct answer token y:

$$\mathrm{CIE}_{\ell j}(\tilde{p}) = P\big(y \mid \tilde{p};\ a_{\ell j} \leftarrow \bar{a}^t_{\ell j}\big) - P(y \mid \tilde{p})\,.$$

Heads are ranked by their average indirect effect across corrupted prompts, and the top heads form the selected set A. The function vector for task t is then defined as the sum of the task-conditioned mean outputs of the selected heads,

$$v_t = \sum_{(\ell,j)\in A} \bar{a}^t_{\ell j},$$

and is applied by adding v_t to the residual stream at a chosen layer during inference. In our experiments, we apply a norm-matched variant, $h'_\ell = h_\ell + \alpha \|h_\ell\| \hat{v}_t$ with $\hat{v}_t = v_t / \|v_t\|$, to account for the growth of residual norms with depth.

Extracting FV for Machine Translation To clarify how the general FV framework applies to translation, we now detail the structure of the ICL prompts and the corruption scheme specific to word-level machine translation. Each ICL prompt p^t_i consists of N demonstration pairs followed by a single query word:

$$p^t_i = \big[(x_{i1}, y_{i1}), \ldots, (x_{iN}, y_{iN}), x_{iq}\big],$$

where each x_{ik} is a word in the source language (e.g., English) and y_{ik} is its translation in the target

[1] Code is available at https://github.com/nurkhanl5/fv_mt.

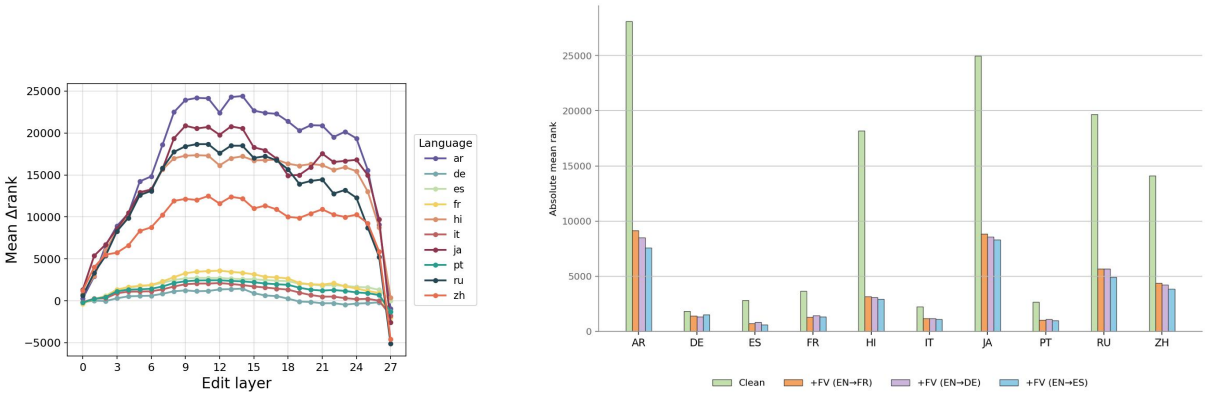


Figure 2: Left: average mean Δrank across edit layers for different target languages for Llama-3.2-3B. Curves correspond to target languages, and higher values mean that the FV moves the correct translation token closer to the top of the next-token distribution. Right: per-language absolute mean rank for LLaMA-3.2-3B under the clean setting and after FV intervention.

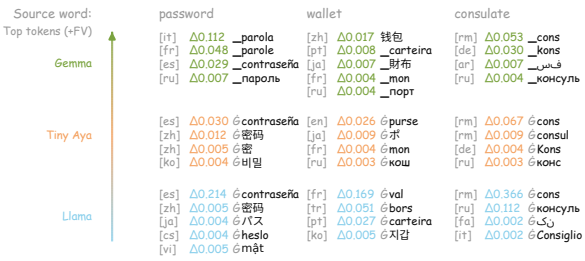


Figure 3: **Token-level effect of FV steering.** For source words, we compare the next-token distribution with and without FV injection and list tokens ranked by their probability increase under the intervention (+FV).

language (e.g., French). For instance, consider an English→French prompt of the following form:

```
Q: eggs\nA: oeufs\n\nQ: good\nA:
bon\n\nQ: sugar\nA: sucre\n\nQ:
loan\nA:
```

Given this prompt, the model must infer the translation mapping from the demonstrated pairs and produce `prêt`, the correct French translation of the query word `loan`. The correct answer token y used in the FV scoring procedure is the gold target token.[2] To form the corrupted prompts $\tilde{p} \in \tilde{P}_t$, the source-side words and the query remain fixed while the target words $\{y_{i1}, \ldots, y_{iN}\}$ are randomly permuted across the N demonstrations, breaking the source–target mapping. The FV extraction then proceeds as described in Paragraph 3.

Models and Data We focus on decoder-only autoregressive LLMs and evaluate three models: Gemma-2-2B (Rivière et al., 2024), Llama-3.2-3B (Grattafiori et al., 2024; Meta, 2024), and Tiny Aya (Cohere, 2026). We choose these models because they provide multilingual coverage at small parameter scales, which enables systematic FV extraction and intervention sweeps within our computational budget. We use the English–French, English–German, and English–Spanish word-pair datasets from Todd et al. (2024) to compute the translation FVs, with 4705, 5154, and 5200 pairs, respectively.

[2] If the gold translation tokenizes into multiple subwords, we use only its first token.

4 Experiments and Results

Cross-lingual rank improvement. We evaluate whether the translation FV exhibits language-agnostic transfer by promoting correct translations across target languages in the next-token distribution. Following the standard FV evaluation setup, we use zero-shot prompts at evaluation. For each English source word w_{src} and target language, we collect the gold translation token and measure its rank under the clean model and after FV intervention.[1] For evaluation, we build a separate held-out word-to-word set of 120 English source words in Spanish, French, German, Italian, Portuguese, Russian, Arabic, Chinese, Japanese, and Hindi.

We formalize this rank-based evaluation as follows. Let $M_{\text{clean}}(\cdot \mid w_{\text{src}})$ denote the model's next-token distribution at the answer position under the clean prompt, and let $M_{\text{FV}}(\cdot \mid w_{\text{src}})$ denote the corresponding distribution under the FV intervention. For each target token t_{tgt}, we measure the change in its rank:

$$\Delta\text{rank}(t_{\text{tgt}} \mid w_{\text{src}}) = \text{rank}\Big(M_{\text{clean}}(t_{\text{tgt}} \mid w_{\text{src}})\Big) - \text{rank}\Big(M_{\text{FV}}(t_{\text{tgt}} \mid w_{\text{src}})\Big). \quad (1)$$

A positive Δrank indicates that the FV moves the gold token closer to the top of the distribution. Figure 2 (left) shows the mean[3] Δrank per target language, averaged across the three experimental conditions (using the EN→ES, EN→DE, and EN→FR FVs separately). The intervention consistently improves the rank of the gold first translation token across target languages, even though each FV is extracted from only a single English→X direction. This suggests that the induced effect is not limited to the target language used to construct the FV, but extends to other target languages as well. A second consistent pattern is that these gains are concentrated in a similar band of middle layers, rather than being distributed uniformly across all layers. This is consistent with prior work that also finds translation-related effects to be strongest in intermediate layers (Zhang et al., 2025a). Although the magnitude varies across target languages, this middle-layer preference is stable across all three FVs.

One limitation of the Δrank metric on its own is that it conflates two qualitatively different situations: a small Δrank may reflect either a weak FV effect or, alternatively, a gold token that was already highly ranked in the clean distribution and therefore has little headroom for improvement. To distinguish these cases, we also examine the absolute mean rank of the gold translation token before and after FV intervention (Figure 2, right).

The results confirm that the cross-lingual effect holds across all evaluated languages, and clarify the role of starting rank. For Arabic, Hindi, Russian, Japanese, and Chinese, the gold token starts from a substantially lower clean rank, and these languages accordingly exhibit the largest absolute improvements under intervention, consistent with the higher Δrank values in Figure 2 (left). For languages where the gold token already ranks favorably in the clean distribution, Spanish, French, German, Italian, and Portuguese, the intervention still shifts it upward, though the room for improvement is inherently smaller. Notably, the FV produces a positive shift in both regimes, indicating that the smaller Δrank values for these latter languages reflect limited headroom rather than a weak intervention.

[3] per-direction results are reported in Appendix A.2

Promoted tokens span multiple languages. We then manually inspect which tokens gain the most probability under FV intervention. As shown in Figure 3, the highest-gain tokens are plausible translations of the source word and span multiple target languages rather than being limited to the single language used to extract the FV. Additional examples are provided in Appendix A.1. Together with the Δrank results, this confirms that the FV induces a broadly multilingual translation effect rather than a language-pair-specific mapping.

Top heads overlap across translation directions. A natural follow-up question is *why* a single-direction FV promotes translations across languages. One explanation is that the FV relies on attention heads that are themselves shared across translation directions. To test this, we compare the top-ranked heads selected by the FV scoring procedure across several English→X translation FVs. We quantify overlap among the top-10 FV-ranked heads across directions, reporting the number of heads shared across all directions. The overlap is substantial across all models: 8 of 10 heads are shared in Gemma-2-2B, 7 in Llama-3.2-3B, and 6 in Tiny Aya. This suggests that part of the FV-associated head set is shared across directions. This finding is consistent with concurrent work that has reported overlap of translation-relevant heads across directions using causal analyses outside the FV framework (Zhang et al., 2025a).

Ablation and specificity. To further confirm that the FV direction carries translation-relevant information rather than an artifact of the addition intervention, we ablate it by projecting it out of the residual stream. Removing the direction degrades the gold-translation rank consistently across all three FVs, and the degradation extends beyond the extraction direction to other target languages, mirroring the cross-lingual pattern observed under FV addition. To assess specificity, we evaluate the ablated model on a suite of standard NLP benchmarks (ARC-Easy (Clark et al., 2018), BoolQ (Clark et al., 2019), COPA (Roemmele et al., 2011), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SciQ (Welbl et al., 2017), WinoGrande (Sakaguchi et al., 2020), and MMLU in German, Spanish, and French); accuracy changes by at most 1–2 points

across all models and tasks, indicating that the ablation is largely selective to translation. Full results are in Appendix A.3.

5 Conclusions

We investigated whether function vectors exhibit language-agnosticity, using machine translation as a case study. Across three multilingual LLMs, we found that translation FVs extracted from a single English→X direction consistently promote correct translations into multiple target languages, with effects concentrated in a shared band of middle layers. Ablating the FV degrades translation across languages while leaving general task performance intact, a pattern compatible with the FV occupying a translation-specific subspace. The cross-lingual effect is further supported by substantial overlap in the top-ranked FV heads across translation directions. Together, these results suggest that function vectors encode a largely language-agnostic translation signal, extending the language-agnosticity property previously observed for representations, grammatical concepts, and circuits.

6 Limitations

Our study has several limitations that should be considered when interpreting the results.

Model scale. We evaluate on three models with 2–3B parameters. While this scale enables systematic intervention sweeps, computational constraints prevent us from testing whether the same patterns hold at larger scales.

Word-level focus. Our experiments operate at the word level, which provides a controlled setting for FV extraction and evaluation but does not capture the full complexity of translation. We do not evaluate on sentence- or document-level translation, nor on more challenging phenomena such as idioms, discourse-level coherence, or morphological agreement. Whether the language-agnostic transfer we observe extends to these settings remains an open question.

Evaluation metric. We rely on Δrank as our primary metric, which measures relative improvement in token ranking but does not directly correspond to translation accuracy or fluency. A token that moves from rank 10,000 to rank 5,000 shows a large Δrank but is still far from being the top prediction.

Interpretation of language-agnosticity. Our results establish behavioral patterns consistent with a language-agnostic translation signal: cross-lingual rank improvements, promoted tokens spanning multiple target languages, shared top-ranked heads, and symmetric degradation under ablation. As with most intervention-based analyses, alternative mechanisms remain possible. The FV may promote lexical items semantically related to the source word regardless of language, or shift probability mass toward translation-like outputs more generally. Testing FV transfer to non-translation semantic tasks would help disentangle these accounts, which we leave for future work.

References

Yonatan Bisk, Rowan Zellers, Roy Schwartz, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *AAAI*.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*.

Cohere. 2026. Cohere Labs Launches Tiny Aya, Making Multilingual AI Practical for Real-World

Languages. https://cohere.com/blog/cohere-labs-tiny-aya.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Javier Ferrando and Marta R. Costa-jussà. 2024. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125, Miami, Florida, USA. Association for Computational Linguistics.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *Preprint*, arXiv:2405.00208.

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.

Aaron Grattafiori, Abhimanyu Dubey, and 1 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Gangwei Jiang, Caigao JIANG, Zhaoyi Li, Siqiao Xue, JUN ZHOU, Linqi Song, Defu Lian, and Ying Wei. 2025. Unlocking the power of function vectors for characterizing and mitigating catastrophic forgetting in continual instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

Théo Lasnier, Armel Randy Zebaze, Djamé Seddah, Rachel Bawden, and Benoît Sagot. 2026. Translation heads: Disentangling meaning from language in LLM-based machine translation. In *Forty-third International Conference on Machine Learning*.

Meta. 2024. Llama 3.2 model cards and prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Yongyu Mu, Peinan Feng, Zhiquan Cao, Yuzhang Wu, Bei Li, Chenglong Wang, Tong Xiao, Kai Song, Tongran Liu, Chunliang Zhang, and JingBo Zhu. 2024. Revealing the parallel multilingual learning within large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6976–6997, Miami, Florida, USA. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and interpretability of multilingual large language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20454–20486, Suzhou, China. Association for Computational Linguistics.

Morgane Rivière, Shreya Pathak, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. Technical report.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial winograd schema challenge at scale. In *AAAI*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Kayo Yin and Jacob Steinhardt. 2025. Which attention heads matter for in-context learning? In *Forty-second International Conference on Machine Learning*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. 2025a. Exploring the translation mechanism of large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025b. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*.

A Appendix

A.1 Extended high-gain token examples

We provide additional high-gain token examples in Figures 4–5, complementing the examples shown in the main text.

Source word	High gained tokens
program	program | programma | प्रोग्राम | कार्यक्रम | programme | Programm | एक | programm | programming | πρόγραμμα | computer | programmi | برنامه | pro | програм | Program | programs | programmer | 프로그램 | कोड
restaurant	भोजन | र | रेस | restaurant | एक | rest | खादय | ख | घर | restaur | restoran | खाने | restaurante | 식당 | Restaurant | प | 레스토랑 | दुकान | क | होटल
shoes	ज | sapat | Schu | zapatos | Z | स | sand | च | प | Sch | πα | pares | Sand | giày | म | क | S | P | cal | sap
spinach	Spin | spinach | sp | Sp | سب | spin | kale | 녹색 | greens | am | Kale | पाल | Greens | p | Green | 잎 | be | ch | :
honey	hon | мед | honey | العسل | Honey | miel | med | h | mel | mad | ch | शहद | मध | de | Hon | miele | z | m | 꿀 |
deadline	termine | scadenza | la | deadline | data | plazo | de | término | limite | prazo | tempo | fecha | scad | f | termini | sc | dat | termin |
homework	tarea | le | homework | travail | devoirs | trabajo | зада | la | домаш | tareas | examen | zad | les | заня | cours | de | уче | doma | Homework | devoir
internet	Internet | internet | 인터넷 | Интернет | इंटरनेट | In | Intern | I | интернет | net | la | réseau | Net | internete | الإنترنت | Netz | اینترنت

Figure 4: **Example of FV effects on the next-token distribution for Tiny-Aya.** For a single source word, we show the top tokens whose probability increases the most after FV, highlighting that the largest gains often correspond to plausible translations across multiple languages.

Source word	High gained tokens
program	program | programa | plan | برنامه | to | 프로그램 | Program | programs | computer | программа | programming | software | programma | programmi | un | code
restaurant	restaurant | Restaurant | rest | ресторан | restaurante | | un | nhà | Rest | ristorante | eatery | 餐 | кафе | restoran | 餐厅 | 餐 | ресто | это | 식
notebook	laptop | | notebook | Laptop | portátil | ноутбу | 笔记本 | nb | Notebook | laptops | la | ノート | lap | net | Пор | лa | порта | 노 | máy | 태
spoon	spoon | Spoon | ложка | ложки | 匙 | лo | colher | ladle | cuchar | spoons | cuchara | - | cuillère | shovel | ку | spoon | teaspoon | ملعق | tablespoon | fork
swimmer	swimmer | 游 | swimmers | swimming | Sw | swim | Swim | sw | 游泳 | schw | n | 수 | người | nage | Swimming | nat | пла | سبا | 泳 | 游
pepper	pepper | paprika | to | pimienta | Pepper | poivre | peppers | chili | p | chile | pep | chi | (| pe | перец | poiv | piment | chilli | 胡椒 | pimenta
fridge	refrigerator | refrigerador | fridge | réfrigérateur | fri | холоди | frigor | kø | Kühlschrank | frigo | ф | Refrigerator | refriger | Frig | hũ | Fridge | kjø | 冰箱 | la | f
deadline	deadline | date | 截止 | 截 | 期 | dead | срок | due | 期限 | край | дата | deadlines | тер | fecha | сро | до | deadline | кон | Deadline | plazo

Figure 5: **Example of FV effects on the next-token distribution for Gemma-2-2B.** For a single source word, we show the top tokens whose probability increases the most after FV, highlighting that the largest gains often correspond to plausible translations across multiple languages.

A.2 Per-direction mean Δrank results

Figures 6–8 report the mean Δrank across edit layers for each target language on Llama-3.2-3B, Gemma-2-2B, and Tiny-Aya, respectively, providing a rank-based view of the function vector's effect on the next-token distribution. To test whether the observed cross-lingual transfer is specific to English as the source language, we repeat the experiments with Arabic and Russian as source languages (Figures 9 and 10). We observe similar patterns, suggesting that the effect is not tied to a particular source language.

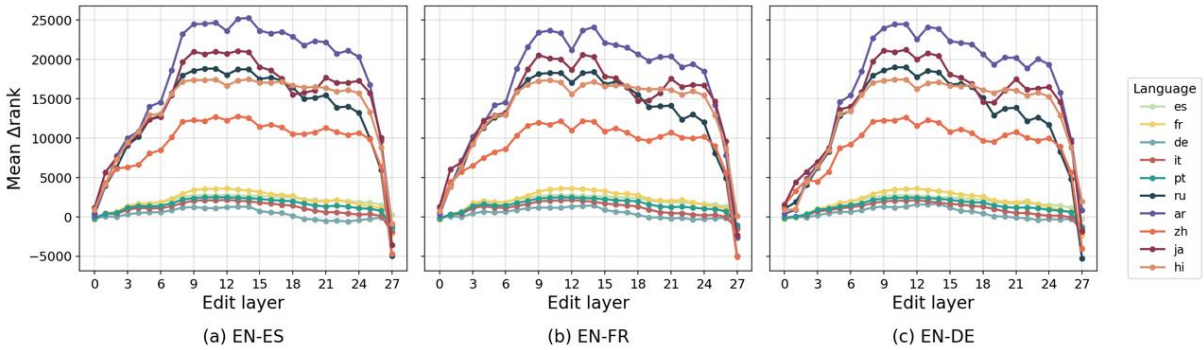


Figure 6: **Mean Δrank across edit layers for different target languages for Llama-3.2-3B.** Curves correspond to target languages, and higher values mean that the FV moves the correct translation token closer to the top of the next-token distribution.

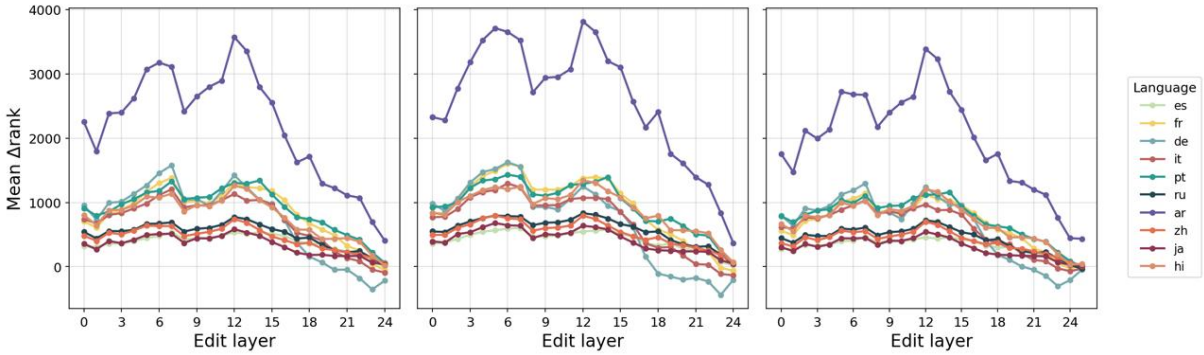


Figure 7: **Mean Δrank across edit layers for different target languages for Gemma-2-2B.** Curves correspond to target languages, and higher values mean that the FV moves the correct translation token closer to the top of the next-token distribution.

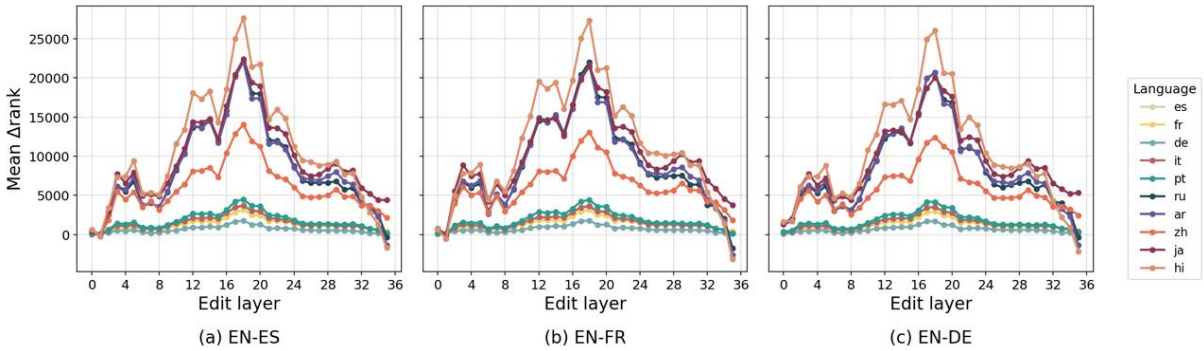


Figure 8: **Mean Δrank across edit layers for different target languages for Tiny-Aya.** Curves correspond to target languages, and higher values mean that the FV moves the correct translation token closer to the top of the next-token distribution.

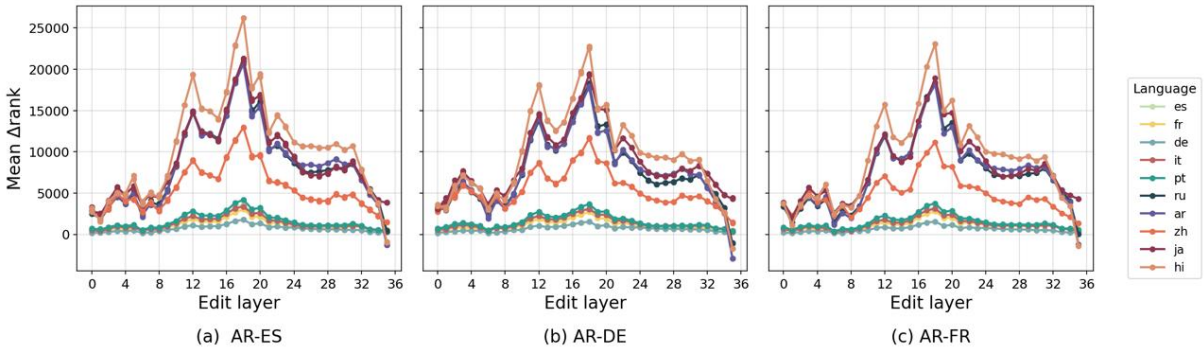


Figure 9: **Mean Δrank across edit layers for different target languages for Tiny-Aya, using FVs extracted from Arabic→{ES, FR, DE} directions.** Curves correspond to target languages, and higher values mean that the FV moves the correct translation token closer to the top of the next-token distribution.

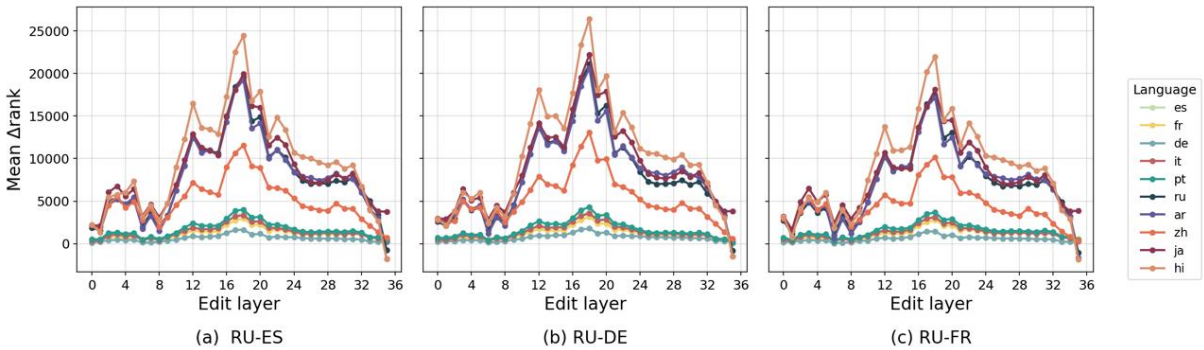


Figure 10: **Mean Δrank across edit layers for different target languages for Tiny-Aya, using FVs extracted from Russian→{ES, FR, DE} directions.** Curves correspond to target languages, and higher values mean that the FV moves the correct translation token closer to the top of the next-token distribution.

A.3 Ablation and specificity.

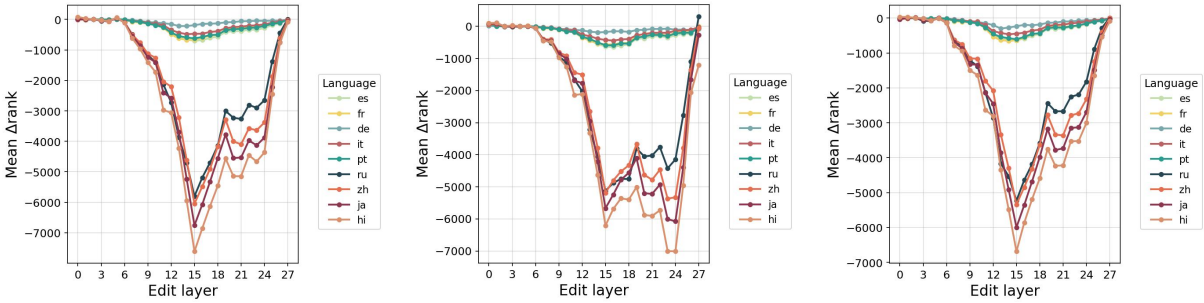
The preceding results show that adding the FV promotes correct translations across target languages. We now test the converse: whether removing the FV direction degrades translation, and whether any such degradation is specific to translation. At a chosen layer ℓ we project out the component of each residual-stream activation aligned with the FV direction:

$$\tilde{h}_{\ell,i} = h_{\ell,i} - \alpha\big(h_{\ell,i}u^{\top}\big)u, \qquad u = \frac{v}{\|v\|_2}, \quad (2)$$

where v is the EN→[X] translation FV and u its unit vector.

Removing the FV direction degrades the rank of the gold translation token consistently across all three FVs (Figures 11–13). Crucially, the degradation extends beyond the extraction direction to other target languages and is concentrated in the same middle-layer band where FV addition produces the largest gains. Addition and ablation thus produce symmetric, cross-lingual effects localized to the same intermediate representations. However, rank degradation is only informative if it reflects translation-specific computation rather than broad disruption of model capabilities. We therefore evaluate the ablated model on standard NLP benchmarks.

As shown in Table 1, accuracy changes by at most 1–2 points across all models and tasks, while translation rank degrades substantially. These results are consistent with the ablation being largely selective to translation. While preserved benchmark accuracy is a null result and does not rule out effects on untested capabilities, it is compatible with the interpretation that the FV direction encodes a translation-relevant subspace rather than a general-purpose feature.



(a) EN-ES (b) EN-DE (c) EN-FR

Figure 11: **Mean Δrank across edit layers for different target languages under FV ablation for Llama-3.2-3B.** Curves correspond to target languages. More negative values indicate that ablating the FV moves the correct translation token lower in the next-token ranking relative to the clean setting.

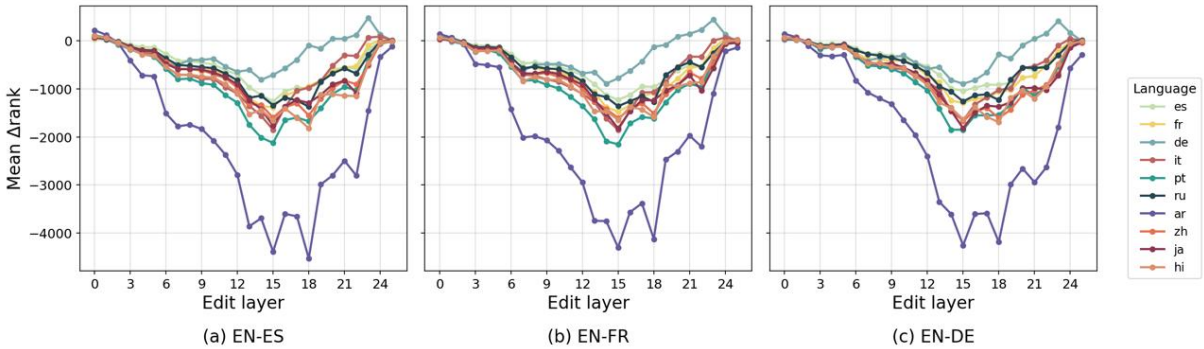


Figure 12: **Mean Δrank across edit layers for different target languages under FV ablation for Gemma-2-2B.** Curves correspond to target languages. More negative values indicate that ablating the FV moves the correct translation token lower in the next-token ranking relative to the clean setting.

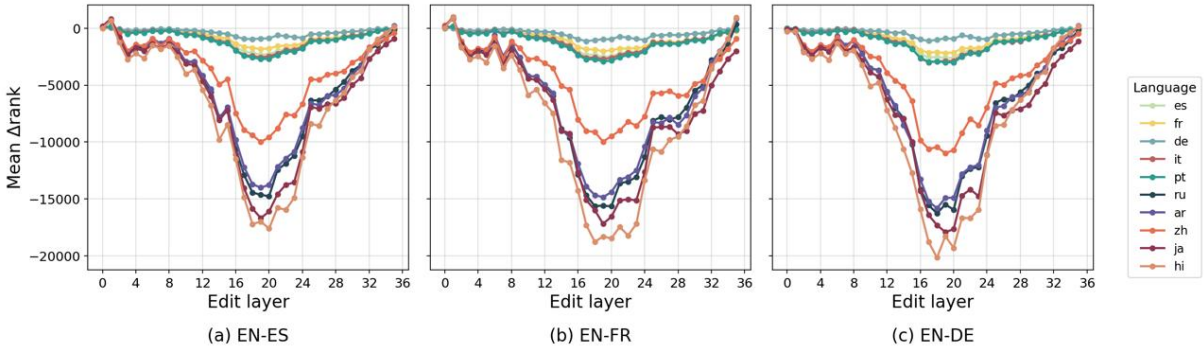


Figure 13: **Mean Δrank across edit layers for different target languages under FV ablation for Tiny-Aya.** Curves correspond to target languages. More negative values indicate that ablating the FV moves the correct translation token lower in the next-token ranking relative to the clean setting.

Data	ARC_EASY	BOOLQ	COPA	HELLA-SWAG	OPEN-BOOKQA	PIQA	SCIQ	WINO-GRANDE	MMLU_DE	MMLU_ES	MMLU_FR
Llama-3.2-3B											
Clean	0.74	0.74	0.82	0.56	0.31	0.76	0.95	0.70	0.52	0.55	0.55
Ablated	0.74 (+0.00)	0.74 (+0.00)	0.83 (+0.01)	0.56 (+0.00)	0.31 (+0.00)	0.76 (+0.00)	0.95 (+0.00)	0.69 (-0.01)	0.53 (+0.01)	0.55 (+0.00)	0.54 (-0.01)
Gemma-2-2B											
Clean	0.80	0.73	0.88	0.55	0.31	0.79	0.96	0.68	0.40	0.43	0.41
Ablated	0.79 (-0.01)	0.70 (-0.03)	0.88 (+0.00)	0.55 (+0.00)	0.31 (+0.00)	0.79 (+0.00)	0.96 (+0.00)	0.69 (+0.01)	0.40 (+0.00)	0.41 (-0.02)	0.41 (+0.00)
Tiny Aya											
Clean	0.76	0.81	0.78	0.50	0.28	0.76	0.94	0.64	0.57	0.52	0.54
Ablated	0.76 (+0.00)	0.81 (+0.00)	0.80 (+0.02)	0.50 (+0.00)	0.28 (+0.00)	0.76 (+0.00)	0.94 (+0.00)	0.65 (+0.01)	0.58 (+0.01)	0.53 (+0.01)	0.54 (+0.00)

Table 1: Clean vs. FV-ablation accuracy across evaluation tasks.