Language Steering for Multilingual In-Context Learning

Anonymous ACL submission

Abstract

If multilingual language model representations contain separable language-associated structure, then shifting between languages may require only simple activation-space interventions. To test this hypothesis, we study multilingual in-context learning as a controlled setting: few-shot demonstrations are provided in English while the test query is in a target language, creating a clear cross-lingual transfer challenge. We propose language vectors, which are computed as mean activation differences between parallel source and target language examples at a particular layer and applied as inference-time offsets to shift the model's internal representations toward the target language. Across three tasks spanning 19 languages and three models, our method consistently improves over baselines without any parameter updates. Beyond performance, the vectors encode interpretable linguistic structure: closely related languages form tight clusters, and vectors transfer effectively across tasks. Together, these findings confirm that language identity occupies structured directions in multilingual activation spaces that can be leveraged to improve cross-lingual transfer.

1 Introduction

Large language models have demonstrated remarkable multilingual capabilities, yet a persistent performance gap between English and low-resource languages limits equitable access across downstream tasks (Zhao et al., 2024; Qin et al., 2024). A key question is whether this gap reflects a fundamental limitation, or whether the necessary cross-lingual knowledge is already encoded internally, waiting to be activated. If so, redirecting a model's internal representations toward the target language should be enough to bridge it.

Recent work in mechanistic interpretability offers a reason to believe this redirection is possible. Transformer-based models have been shown to encode distinct patterns for different types of information within their internal representations (Rai et al., 2024; Li et al., 2024; Elhage et al., 2022), suggesting that concepts are not uniformly distributed but occupy structured, separable directions in activation space. This raises a natural question: is language identity encoded the same way? If so, a simple activation offset should suffice to steer the model toward target-language processing. Indeed, mean activation differences between contrastive inputs have been shown to encode interpretable directions in representation space (Marks and Tegmark, 2023; Rimsky et al., 2024), and such directions have been applied to address language confusion in multilingual models (Park et al., 2023; Yunfan et al., 2025; Sterz et al., 2025). However, these approaches target output language fidelity rather than downstream task performance, and whether language identity occupies directions that can be leveraged for cross-lingual task improvement remains an open question.

To investigate this, we need a setting with clear controllability over the language of inputs and outputs. In-context learning (ICL) (Brown et al., 2020) naturally provides this. When demonstrations are provided in a source language (English) where high-quality data exists but the test query is in a low-resource target language, models must internally recognise the language switch and transfer their task understanding across languages, a computationally and representationally demanding process that often fails. This cross-lingual ICL gap provides a natural testbed for our hypothesis.

To test this hypothesis, we apply *language vectors*, mean activation differences between hidden states of parallel source and target language examples, added as inference-time offsets, to the multilingual ICL setting. As shown in Figure 1, our method operates in a few-shot ICL setting, where we: (1) collect paired examples in the source and target language, (2) compute the activation differ-

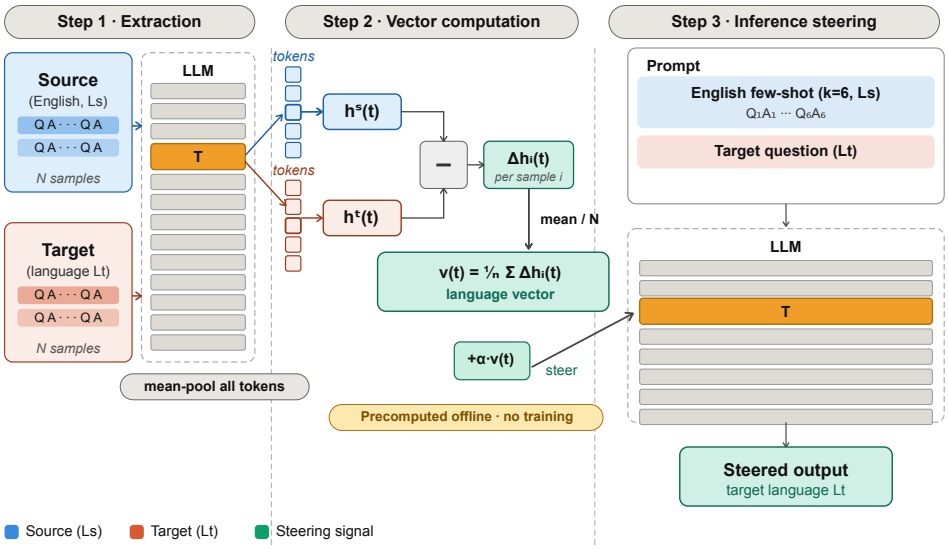

Figure 1: Overview of our language steering approach for multilingual in-context learning. **Step 1 (Extraction):** We extract hidden states from parallel source and target language question-answer pairs at layer t of the LLM. **Step 2 (Vector Computation):** The language steering vector $\mathbf{v}(t)$ is computed offline as mean of activation differences between target and source language representations across N samples. **Step 3 (Inference-Time Steering):** During inference with few-shot demonstrations from source language and a target language question, we steer the model by adding $\alpha \cdot \mathbf{v}(t)$ to the hidden states at layer t, guiding the model toward target language reasoning patterns.

ence to extract a language-specific steering vector, and (3) apply this vector during inference to shift the model's internal representations toward target-language processing behavior. We treat language vectors as an experimental probe of whether cross-lingual structure in activation space is exploitable.

We evaluate across three datasets spanning mathematical reasoning and natural language inference, 19 languages, and three models. Our results validate the hypothesis: consistent performance improvements over baselines suggest that language identity occupies structured directions in representation space, and the vectors encode interpretable linguistic structure, closely related languages cluster together, early-to-middle layers are most effective, steering the few-shot demonstrations is most effective and vectors transfer across tasks.

The contributions of our work are: (1) We use language vectors, a training-free adaptation of prior activation steering work (Marks and Tegmark, 2023; Park et al., 2023; Liu et al., 2023), and show that language identity occupies exploitable directions in multilingual LLM activation space, demonstrated through consistent downstream ICL performance gains. (2) We provide comprehensive empirical evaluation across three datasets, 19 languages, and three models. (3) We show that language vectors encode meaningful linguistic structure, with closely related languages forming tight clusters. (4) We demonstrate effective cross-task transfer, suggesting the vectors capture language-specific rather than task-specific representations.

2 Language Vectors: Method and Formulation

We consider multilingual in-context learning as our setup following Tu et al. (2025), where each instance consists of few-shot source (English) demonstrations and a test question in the target language, and aim to improve over this cross-lingual baseline toward the performance achievable with target-language demonstrations.

2.1 Problem Formulation and Notations

Let L_s denote the source language and L_t denote the target language. In the multilingual in-context learning setup, the model receives a prompt consisting of k few-shot demonstration examples $\mathcal{F} = \{(q_i^{L_s}, c_i^{L_s}, a_i^{L_s})\}_{i=1}^{k}$ followed by a test question $q_{\text{test}}^{L_t}$. Each demonstration includes a question $q_i^{L_s}$, chain-of-thought reasoning $c_i^{L_s}$ (if any), and answer $a_i^{L_s}$, all in the source language L_s. The model must leverage the task understanding from these source language demonstrations to generate both the reasoning $c_{\text{test}}^{L_t}$ and answer $a_{\text{test}}^{L_t}$ for the target language. Given a language model $\mathcal{M}$

with T layers, we denote the hidden state at layer $t \in \{1, \ldots, T\}$ and token position j for input text x as $\mathbf{h}_j(t) \in \mathbb{R}^d$, where d is the model's hidden state size.

2.2 Computing Language Directions

Our goal is to compute a language-specific steering vector $\mathbf{v}(t) \in \mathbb{R}^d$ that, when added to hidden states during inference, improves the model's performance on target language inputs q^{L_t} beyond the baseline. We fix the source language as English and compute steering vectors from activation patterns across parallel source-target examples.

Few-shot Format for Language Vector Computation. To capture richer contextual patterns, we construct $N = |\mathcal{D}_{\text{compute}}|$ samples, each consisting of k question-answer pairs concatenated together. Specifically, for each example $i \in \mathcal{D}_{\text{compute}}$, we construct a sample by placing example i in the first slot and filling the remaining $k-1$ slots by sampling uniformly with replacement from $\mathcal{D}_{\text{compute}}$. This ensures each example appears at least once across all constructed samples. For each sample i in our compute set $\mathcal{D}_{\text{compute}}$, we format parallel texts as:

$$x_i^{L_s} = \bigoplus_{m=1}^{k} [q_{i,m}^{L_s}, a_{i,m}^{L_s}], \quad x_i^{L_t} = \bigoplus_{m=1}^{k} [q_{i,m}^{L_t}, a_{i,m}^{L_t}] \tag{1}$$

where $q_{i,m}^{L_s}$ and $q_{i,m}^{L_t}$ are parallel questions in source and target languages respectively, $a_{i,m}^{L_s}$ and $a_{i,m}^{L_t}$ are the answers. For datasets with reasoning involved we also append $c_{i,m}^{L_s}$ and $c_{i,m}^{L_t}$ respectively in the text. $\oplus$ denotes concatenation, and k is the number of question-answer pairs per sample.

Activation Extraction and Vector Computation. We extract hidden states at layer t for both source and target formatted texts. Recall that each sample i is a concatenated sequence of k question-answer pairs. We pass each such sequence through model M and mean-pool the hidden states across all token positions to obtain a single vector per sample.

$$\mathbf{h}_i^{L_s}(t) = \frac{1}{|x_i^{L_s}|} \sum_{j=1}^{|x_i^{L_s}|} \mathbf{h}_{i,j}^{L_s}(t) \tag{2}$$

$$\mathbf{h}_i^{L_t}(t) = \frac{1}{|x_i^{L_t}|} \sum_{j=1}^{|x_i^{L_t}|} \mathbf{h}_{i,j}^{L_t}(t) \tag{3}$$

where $\mathbf{h}_{i,j}^{L_s}(t)$ denotes the hidden state at token position j of sample i in the source language at layer t, and $|x_i^{L_s}|$ is the total sequence length. We compare mean pooling against last-token extraction in Appendix C.3, finding mean pooling consistently superior. The steering vector $\mathbf{v}(t) \in \mathbb{R}^d$ at layer t is the mean of activation differences:

$$\mathbf{v}(t) = \frac{1}{N} \sum_{i=1}^{N} \left(h_i^{L_t}(t) - h_i^{L_s}(t) \right) \tag{4}$$

where $N = |\mathcal{D}_{\text{compute}}|$ is the number of samples in the compute set. This setup ensures that adding the steering vector shifts activations from the source language distribution (English) toward the target language distribution, effectively steering the model's internal representations to behave as if processing target language demonstrations.

2.3 Inference-Time Steering

During inference, we apply the pre-computed steering vector to hidden states at specific token positions within the input prompt. Given a test example from $\mathcal{D}_{\text{test}}$ with few-shot demonstrations $\mathcal{F}$ in the source language and a target language question $q_{\text{test}}^{L_t}$, the full input prompt is:

$$\text{prompt} = [\text{system_instruction}] \oplus \mathcal{F} \oplus q_{\text{test}}^{L_t} \tag{5}$$

where $\mathcal{F} = \{(q_1^{L_s}, c_1^{L_s}, a_1^{L_s}), \ldots, (q_k^{L_s}, c_k^{L_s}, a_k^{L_s})\}$ consists of k few-shot demonstrations with source language questions $q_i^{L_s}$, source language chain-of-thought reasoning $c_i^{L_s}$, and answers a_i.

Steering Positions. We define a set of token positions $\mathcal{P}$ where steering is applied. We experiment with four steering position configurations (Table 4): `on_fewshot` (all few-shot tokens), `after_fewshot` (boundary token only), `on_question` (test question tokens only), and `entire` (all prompt tokens).

Modified Forward Pass. We modify the hidden states for each position $p \in \mathcal{P}$ during the forward pass at layer t:

$$\mathbf{h}'_p(t) = \mathbf{h}_p(t) + \alpha \cdot \mathbf{v}(t) \tag{6}$$

where $\mathbf{h}_p(t) \in \mathbb{R}^d$ is the original hidden state at position p and layer t, $\alpha \in \mathbb{R}$ is a scaling hyperparameter controlling steering strength, and $\mathbf{h}'_p(t)$ is the steered hidden state. The hyperparameters $(t^*, \alpha^*, \mathcal{P}^*)$ are selected based on validation set performance ($\mathcal{D}_{\text{val}}$).

Language	MGSM				XNLI				MSVAMP			
	B	MFS	Ours	OR	B	MFS	Ours	OR	B	MFS	Ours	OR
Arabic	–	–	–	–	62.57	60.78	**62.87**	63.77	–	–	–	–
Basque	32.14	35.70	**36.90**	52.38	–	–	–	–	–	–	–	–
Bengali	57.14	55.95	**61.90**	58.33	–	–	–	–	57.49	**62.87**	59.58	61.38
Bulgarian	–	–	–	–	56.29	**61.98**	61.68	66.17	–	–	–	–
Catalan	64.29	64.28	**69.05**	76.19	–	–	–	–	–	–	–	–
Chinese	67.86	67.86	**71.43**	72.62	59.88	**61.38**	59.28	61.98	69.76	**73.05**	71.26	73.35
French	61.90	64.29	**70.24**	65.48	67.37	68.26	**72.75**	71.26	71.56	73.05	**74.55**	73.05
Galician	64.29	69.04	**73.81**	77.38	–	–	–	–	–	–	–	–
German	66.67	66.67	**75.00**	71.43	64.07	65.27	**66.47**	65.27	71.26	70.06	**71.26**	76.95
Greek	–	–	–	–	68.26	66.17	**70.06**	72.75	–	–	–	–
Hindi	–	–	–	–	61.38	57.78	**64.97**	61.98	–	–	–	–
Japanese	55.95	**61.90**	55.95	63.10	–	–	–	–	63.17	**68.26**	67.96	70.06
Russian	71.43	71.43	**72.62**	76.19	58.98	**64.37**	63.77	60.78	68.86	**72.46**	72.16	71.56
Spanish	**77.38**	70.24	76.19	78.57	66.77	66.17	**70.66**	67.37	74.55	73.95	**75.75**	74.55
Swahili	55.95	63.10	**65.48**	66.67	52.40	51.20	**55.99**	56.89	56.29	59.58	**60.78**	62.87
Thai	57.14	**67.86**	61.90	59.52	59.88	**64.37**	60.78	70.66	59.58	**64.67**	64.07	66.77
Turkish	–	–	–	–	62.28	57.19	**65.87**	65.57	–	–	–	–
Urdu	–	–	–	–	55.09	55.39	**56.29**	57.19	–	–	–	–
Vietnamese	–	–	–	–	64.07	63.17	**68.86**	68.86	–	–	–	–
Average	61.01	63.19	**65.87**	68.16	61.38	61.68	**64.31**	65.04	65.84	**68.66**	68.60	70.06

Table 1: Detailed per-language accuracy results on MGSM, XNLI, and MSVAMP for the Llama-3.1-8B-Instruct model. Gray dashes indicate that the dataset does not contain that language. B = Baseline, MFS = Multilingual Few-Shot Baseline (Tu et al., 2025), OR = Oracle. Oracle is the upper bound and is not for a direct comparison. **Bold** indicates the best score per language excluding the oracle.

3 Experiments

3.1 Setup

We test on three instruction-tuned models: Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct (Grattafiori et al., 2024; Team et al., 2024). We consider three datasets: MGSM (mathematical reasoning) (Shi et al., 2022), MSVAMP (arithmetic word problems) (Chen et al., 2024), and XNLI (natural language inference) (Conneau et al., 2018).

Data Splits. We partition the test set into three equal parts: compute ($\mathcal{D}_{\text{compute}}$) for steering vector calculation, validation ($\mathcal{D}_{\text{val}}$) for hyperparameter and configuration selection, and test ($\mathcal{D}_{\text{test}}$) for final evaluation. Since training data is minimal or unavailable for some benchmarks, we partition the test set uniformly across all datasets for consistency. We sample 6 examples from the available training data as few-shot demonstrations, kept fixed across all experiments. Task-level configurations (Appendix F) offer a practical alternative when per-language validation data is unavailable.

Implementation Details. While calculating the vector, we create $N = |\mathcal{D}_{\text{compute}}|$ samples, ensuring each example from the compute set appears at least once across all samples, with remaining slots filled through random sampling with replacement. We use $k = 6$ few-shot pairs per sample, mirroring the test-time prompt structure to ensure the steering vector is computed under conditions consistent with inference. We perform a grid search over steering layers $t \in \{5, 10, 15, 20, 25, 30\}$, scaling factors $\alpha \in \{0.5, 1.0, 2.0, 3.0\}$, and the four steering position configurations on $\mathcal{D}_{\text{val}}$. The reported results on $\mathcal{D}_{\text{test}}$ correspond to per-language validation-selected configurations.

Baselines. We compare our method against two baselines: (1) *Source baseline (B)*: few-shot prompts with source questions and source chain-of-thought and answer; (2) *Multilingual few-shot (MFS)*: few-shot examples drawn from multiple languages with their respective answers Tu et al. (2025), representing the approach of diversifying few-shot examples across languages. To understand what our upper bound would be when evaluating, we also test: *Oracle*: uses target-language few-shot demonstrations and serves as an upper bound. Following Tu et al. (2025), wherever there is a chain of thought reasoning involved, it's always in the source language (English).

Model	MGSM				XNLI				MSVAMP			
	B	MFS	Ours	OR	B	MFS	Ours	OR	B	MFS	Ours	OR
Llama-3.1-8B-Instruct	61.01	63.19	**65.87**	68.16	61.38	61.68	**64.31**	65.04	65.84	**68.66**	68.60	70.06
Qwen2.5-7B-Instruct	68.95	69.94	**71.83**	73.21	74.81	73.27	**75.51**	74.53	75.18	**76.94**	76.68	77.65
Qwen2.5-14B-Instruct	81.55	82.84	**84.23**	84.72	72.73	73.91	**74.66**	76.45	82.10	82.90	**83.90**	83.03

Table 2: Average performance comparison across three datasets and three models. B = Baseline, MFS = Multilingual Few-Shot Baseline (Tu et al., 2025), OR = Oracle. Oracle is the upper bound and is not for a direct comparison. Detailed per-language results are in Appendix A.

3.2 Results

Performance on Llama Model. Detailed accuracy results on $\mathcal{D}_{\text{test}}$ for the Llama-3.1-8B-Instruct model are shown in Table 1 across the three datasets, covering 19 languages. Our method improves over the source baseline (B) across most languages and datasets, demonstrating that activation steering enables effective cross-lingual transfer without any parameter updates. MGSM shows the most consistent and substantial improvements, indicating that structured mathematical reasoning particularly benefits from cross-lingual steering. On MSVAMP, our method performs comparably to MFS, with differences of less than 0.3 points on average. XNLI presents the most mixed results: gains vary considerably across languages, and MFS occasionally outperforms ours. Overall, our method shows the strongest gains on reasoning tasks, remains competitive with MFS on MSVAMP, and provides consistent improvements over the source baseline across settings.

Performance on Other Models. Table 2 summarizes average performance across all languages for each model and dataset combination. Our method consistently outperforms the source baseline on MGSM across all three models, with improvements ranging from 2.9% to 5.4%, confirming that activation steering is especially effective for structured mathematical reasoning. On XNLI and MSVAMP, improvements are generally smaller, with MFS remaining competitive especially on MSVAMP. Despite this, our method performs comparably to or better than the baselines in the majority of cases across models and datasets. Per-language results for Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct are in Appendix A.

Optimal Layer and α. Table 3 reports the optimal steering layer and scaling factor α per language and task for Llama-3.1-8B-Instruct. For XNLI, layer 10 dominates across 11 out of 14 languages, consistent with middle-layer representations encoding semantic and pragmatic features relevant to natural language inference. For MGSM and MSVAMP, layer 5 is most frequent, suggesting mathematical reasoning benefits from steering at earlier layers where surface-level language features are more prominent. The optimal α also varies: $\alpha = 3.0$ dominates for XNLI while MGSM and MSVAMP show lower values, suggesting natural language inference requires stronger steering than structured reasoning tasks. Task-level configurations (Appendix F) remain above the source baseline, confirming the method is effective without per-language validation data as well.

Language	MGSM	MSVAMP	XNLI
Arabic	—	—	10/3.0
Basque	5/1.0	—	—
Bengali	20/1.0	5/0.5	—
Bulgarian	—	—	10/3.0
Catalan	20/3.0	—	—
Chinese	30/0.5	5/0.5	5/3.0
French	5/1.0	15/0.5	10/3.0
Galician	5/1.0	—	—
German	5/2.0	30/1.0	5/3.0
Greek	—	—	10/3.0
Hindi	—	—	10/3.0
Japanese	5/1.0	10/0.5	—
Russian	30/0.5	25/0.5	10/3.0
Spanish	30/1.0	5/0.5	10/3.0
Swahili	5/1.0	5/0.5	5/3.0
Thai	5/0.5	5/2.0	10/3.0
Turkish	—	—	10/3.0
Urdu	—	—	10/3.0
Vietnamese	—	—	10/3.0

Table 3: Best layer/α per language and task for Llama-3.1-8B-Instruct. Each cell shows *layer*/α; — indicates the language was not evaluated on that task.

Steering Position. Table 4 shows accuracy across the four steering position configurations on MGSM. No single position dominates across all languages. On average, steering on the few-shot demonstrations (`on_fewshot`) achieves the highest accuracy (65.58%), roughly a 4-point im-

Language	B	OF	AF	OQ	ENT
Bengali	57.14	60.71	**61.90**	60.71	55.95
Catalan	64.29	66.67	**69.05**	67.86	**69.05**
German	66.67	**75.00**	–	69.05	64.29
Spanish	**77.38**	73.81	75.00	76.19	72.62
Basque	32.14	35.71	**36.90**	33.33	**36.90**
French	61.90	67.86	64.29	67.86	**70.24**
Galician	64.29	**73.81**	69.05	66.67	**73.81**
Japanese	**55.95**	–	**55.95**	–	–
Russian	71.43	69.05	**72.62**	**72.62**	69.05
Swahili	55.95	**65.48**	60.71	63.10	**65.48**
Thai	57.14	**61.90**	54.76	**61.90**	59.52
Chinese	67.86	**71.43**	**71.43**	**71.43**	**71.43**
Average	61.01	**65.58**	62.88	64.61	64.39

Table 4: Per-language accuracy across steering position configurations for Llama-3.1-8B-Instruct on MGSM. B = Baseline, OF = On Few-shot, AF = After Few-shot, OQ = On Question, ENT = Entire Prompt. Dashes indicate configurations where no validation improvement over baseline was found. **Bold** indicates the best score per language.

provement over baseline, likely because the model forms its task understanding while processing the demonstrations. In the absence of validation data, `on_fewshot` is the recommended default given its consistently strong average performance. Having established that language vectors improve cross-lingual ICL performance, we now ask what these vectors encode and whether the structure they capture is linguistically meaningful.

4 What Do Language Vectors Encode?

4.1 Cross-Task Transfer

We analyze whether language-specific steering vectors transfer across tasks by conducting systematic evaluation across three datasets: MGSM, MSVAMP, and XNLI, examining all six possible transfer directions. Hyperparameters are the same as those used to compute *Ours*. For each direction, we compute steering vectors from the source task and apply them during evaluation on the target task, using only languages common to both datasets.

Table 5 presents per-language cross-task transfer results for Qwen2.5-7B-Instruct, where we denote A→B as evaluation on B using vectors computed from A. Five out of six transfer directions demonstrate effective generalization, with high-resource languages such as Spanish and Russian showing consistent gains while Swahili exhibits higher variance. The transfer results reveal a consistent directional pattern: reasoning-to-reasoning transfers (MGSM↔MSVAMP) are largely symmetric, while transfers across the reasoning/inference boundary are more asymmetric. MSVAMP→XNLI represents a significant failure case, achieving only 47.98% compared to 75.19% baseline, a 27 percentage point drop. This failure is asymmetric: XNLI→MSVAMP works successfully, suggesting MSVAMP vectors encode task-specific mathematical patterns detrimental to natural language inference. These results indicate that language-specific steering vectors generally capture task-agnostic cross-lingual representations, but transfer effectiveness depends on task compatibility. Beyond task-specific data, we also evaluate vectors computed from general-purpose parallel sentences (Costa-Jussà et al., 2022) and single question-answer pairs; as shown in Table 12 in appendix, both improve over the source baseline for most configurations, though concatenated few-shot sequences consistently outperform both alternatives.

4.2 Do Vectors Reflect Linguistic Geometry?

To investigate whether language vectors encode linguistically meaningful structure beyond task performance, we apply hierarchical clustering to the steering vectors computed for each language as shown in Figure 2. If the vectors capture meaningful language-associated structure, we would expect typologically related languages to cluster together. The closest merges are Catalan–Galician (∼0.042) and Hindi–Urdu (∼0.043), consistent with their close linguistic relationships. Bulgarian and Greek form the next tight pair (∼0.065), with Russian joining at ∼0.140, likely reinforced by shared Cyrillic script. Spanish and French merge at ∼0.100, absorbing German at ∼0.218, while Bengali completes an Indo-Aryan cluster at ∼0.201 with Hindi and Urdu. Turkish, Basque, and Japanese group together (∼0.272–0.309) despite having no shared ancestry, and Chinese remains the most distinct language, merging last. Overall, these clusters capture a mixture of typological, script-level, and model-specific similarities rather than strictly genealogical language families, further supported by the L2 norms in Appendix E.2. Random steering vectors collapse into a near-flat hierarchy with all languages merging at a single high distance (1.967), confirming that the structure reflects genuine language-specific encoding rather than an artifact of hyperparameter optimization.

Language	MGSM (vector) → XNLI (eval)					XNLI (vector) → MGSM (eval)				
	B	MFS	Ours	CT	OR	B	MFS	Ours	CT	OR
Chinese	78.14	77.84	77.84	**78.14**	79.34	84.52	88.10	86.90	**90.48**	79.76
Thai	70.96	**73.05**	70.06	70.06	72.46	79.76	**82.14**	79.76	**82.14**	85.71
Swahili	53.29	47.31	**55.09**	54.79	54.79	19.05	16.67	**26.19**	25.00	33.33
Russian	78.14	79.04	**79.64**	**79.94**	78.74	85.71	85.71	**88.10**	**88.10**	85.71
French	81.44	82.34	82.34	**82.63**	82.63	78.57	80.95	**79.76**	**79.76**	84.52
Spanish	82.63	82.63	**83.53**	83.23	81.74	84.52	83.33	**89.29**	88.10	86.90
German	81.74	77.54	82.93	**82.98**	79.34	83.33	83.33	**85.71**	**85.71**	86.90
Average	75.19	74.22	75.92	**75.96**	75.58	73.64	74.32	**76.53**	76.36	77.55

Language	MGSM (vector) → MSVAMP (eval)					MSVAMP (vector) → MGSM (eval)				
	B	MFS	Ours	CT	OR	B	MFS	Ours	CT	OR
Bengali	65.57	70.96	67.07	**67.66**	76.35	58.33	**66.67**	58.33	58.33	65.48
German	79.04	79.94	**80.24**	79.64	81.14	83.33	83.33	**85.71**	77.38	86.90
Spanish	82.04	83.53	**83.83**	82.04	84.13	84.52	83.33	89.29	**90.48**	86.90
French	82.34	**85.03**	84.13	79.04	84.43	78.57	**80.95**	79.76	77.38	84.52
Japanese	78.74	**79.64**	**79.64**	79.34	79.94	71.43	75.00	75.00	**76.19**	73.81
Russian	79.64	**82.93**	81.74	80.54	81.44	85.71	85.71	**88.10**	83.33	85.71
Swahili	49.10	47.90	**50.60**	48.50	50.30	19.05	16.67	**26.19**	20.24	33.33
Thai	78.14	**79.34**	**79.34**	76.65	78.74	79.76	82.14	**86.90**	78.57	79.76
Chinese	82.04	83.23	**83.53**	83.23	82.34	84.52	**88.10**	86.90	**88.10**	79.76
Average	75.18	**76.94**	76.68	75.18	77.65	71.69	73.54	**74.34**	72.22	75.79

Language	MSVAMP (vector) → XNLI (eval)					XNLI (vector) → MSVAMP (eval)				
	B	MFS	Ours	CT	OR	B	MFS	Ours	CT	OR
Chinese	78.14	**77.84**	**77.84**	66.47	79.34	82.04	83.23	**83.53**	80.84	82.34
Thai	70.96	**73.05**	70.06	29.04	72.46	78.14	**79.34**	**79.34**	76.35	78.74
Swahili	53.29	47.31	**55.09**	33.83	54.79	49.10	47.90	**50.60**	48.80	50.30
Russian	78.14	79.04	**79.64**	43.70	78.74	79.64	**82.93**	81.74	79.64	81.44
French	81.44	**82.34**	**82.34**	50.60	82.63	82.34	**85.03**	84.13	81.44	84.43
Spanish	82.63	82.63	**83.53**	58.98	81.74	82.04	83.53	**83.83**	82.63	84.13
German	81.74	77.54	**82.93**	53.29	79.34	79.04	79.94	80.24	**81.14**	81.14
Average	75.19	74.22	**75.92**	47.98	75.58	75.18	**76.94**	76.68	75.83	77.65

Table 5: Cross-task transfer results across all task-vector combinations between MGSM, MSVAMP, and XNLI on Qwen2.5-7B-Instruct. For each task pair, we compute steering vectors from one task and evaluate on another task. B = Baseline, MFS = Multilingual Few-Shot Baseline (Tu et al., 2025), CT = Cross-Transfer, OR = Oracle. Oracle is the upper bound and is not for a direct comparison. **Bold** indicates the best score.

5 Related Work

Cross-Lingual Transfer and Alignment. Cross-lingual transfer learning investigates how multilingual models transfer knowledge across languages (Huang et al., 2021, 2022; Qin et al., 2025). To mitigate cross-lingual performance gaps, alignment strategies have been proposed ranging from prompt-level techniques such as multilingual prompting and code-switching, to representation-level methods and cross-lingual in-context pretraining (Li et al., 2024; Zhang et al., 2024; Wu et al., 2025; Tu et al., 2025; Ahuja et al., 2023). Translation-based approaches offer an alternative: Shi et al. (2023) demonstrate that multilingual chain-of-thought reasoning benefits from English-language reasoning chains, and Liu et al. (2025) and Zhu et al. (2024) examine when translating inputs into English yields downstream gains. Unlike these approaches, we avoid translation and additional training, steering representations during inference.

Multilingual Representations and Interpretability. Multilingual models encode language identity and linguistic structure as structured patterns within their activations (Zhao et al., 2024; Gurgurov et al., 2025; Tang et al., 2024; Pokharel et al., 2026). Wendler et al. (2024) and Schut et al. (2025) show that intermediate representations are biased toward English even for non-English inputs, while Dumas et al. (2025) demonstrate via activation patching that concept representations can be separated from language identity, suggesting that language occupies structured, manipulable di-

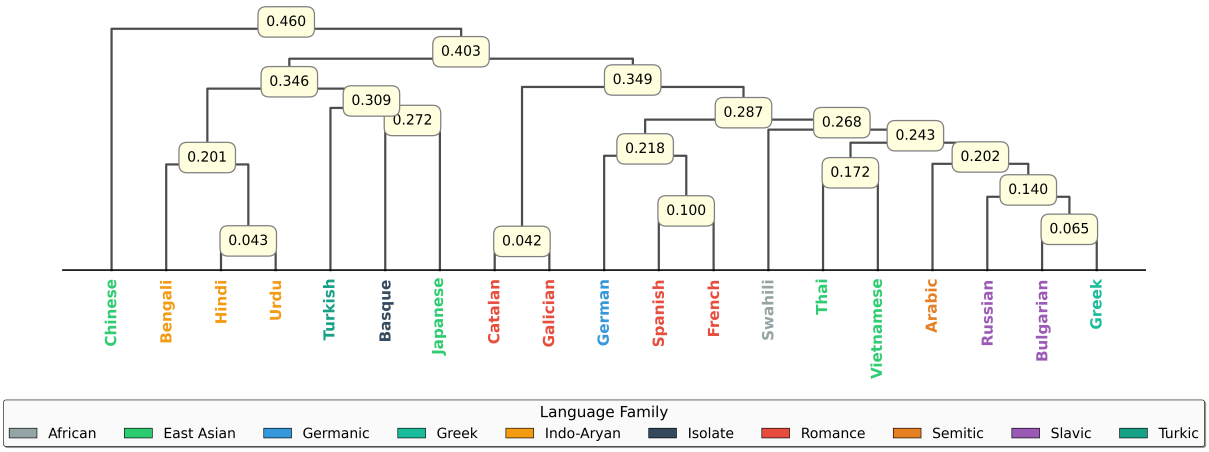

Figure 2: Hierarchical clustering of steering vectors at layer 10 of Llama-3.1-8B-Instruct. Edge labels show cosine distances (lower = more similar); colors indicate language families. Clusters align with typological similarity and script characteristics.

rections in activation space. These insights motivate activation-level interventions without retraining (Nie et al., 2025). Pokharel et al. (2026) propose CLAS, which modulates neuron activations to identify shared and language-specific neurons for zero-shot classification. Our approach instead targets ICL performance on reasoning and inference tasks via mean activation differences in a few-shot QA format.

Activation Steering and Language Vectors. Activation steering modifies internal activations at inference time to control model behavior (Turner et al., 2023). Zou et al. (2023) established a top-down framework for reading and writing interpretable directions in representation space; foundational work showed that mean activation differences between contrastive inputs encode such directions (Marks and Tegmark, 2023; Rimsky et al., 2024) and can shift model behavior as inference-time offsets (Park et al., 2023). Jorgensen et al. (2023) further showed that mean-centering improves steering vector quality, directly informing our computation procedure, and (Liu et al., 2023) showed that vectors from few-shot English demonstrations can steer downstream task behavior, forming the methodological precursor to our approach. More recent work proposes lightweight post-training steering methods, showing that simple residual stream interventions reliably shift model behavior (Cui and Chen, 2025; Stolfo et al., 2024). More recent work applies language-specific vectors to language confusion, where models output in an unintended language (Yunfan et al., 2025; Sterz et al., 2025; Lopo et al., 2025); notably, Sterz et al. (2025) observe that prior steering approaches tend to harm downstream task accuracy, while Lopo et al. (2025) similarly target output language fidelity rather than task accuracy. Unlike these approaches, our vectors shift language while preserving task structure, targeting downstream ICL accuracy rather than surface language control.

In-Context Learning. ICL enables task adaptation without gradient updates (Brown et al., 2020; Dong et al., 2024). Hendel et al. (2023) showed that ICL compresses demonstrations into implicit task vectors, and Todd et al. (2024) extended this to function vectors that generalize across contexts. Unlike these task-encoding approaches, our vectors capture language shift rather than task structure. Cross-lingual ICL studies show that performance degrades substantially when demonstrations are in a high-resource language but test queries are in a low-resource one (Tanwar et al., 2023; Winata et al., 2021), precisely the gap our method targets.

6 Conclusion

We introduce language vectors, a training-free approach that improves multilingual ICL by applying activation differences between source and target languages as inference-time offsets, extending prior activation steering work (Marks and Tegmark, 2023) to cross-lingual transfer. Our evaluation across three datasets, 19 languages, and three models demonstrates consistent gains, particularly on reasoning tasks. These findings suggest that language identity occupies structured directions in multilingual activation spaces that can be leveraged to improve cross-lingual transfer.

7 Limitations

Our evaluation is limited to instruction-tuned decoder-only models. Understanding why certain task transfers fail, extending the approach to languages without parallel data, and applying language vectors to generative tasks such as summarization remain promising directions for future work. Additionally, the relationship between steering vector norms and representational distance from English suggests a potential avenue for predicting steering effectiveness.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2475–2485.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Sasha Cui and Zhongren Chen. 2025. Painless activation steering: An automated, lightweight approach for post-training large language models. *arXiv preprint arXiv:2509.22739*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

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.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Bäumel, Josef van Genabith, and Simon Ostermann. 2025. Language arithmetics: Towards systematic language neuron identification and manipulation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2911–2937.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore. Association for Computational Linguistics.

Kuan-Hao Huang, Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1684–1697. Association for Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4633–4646. Association for Computational Linguistics.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. 2023. Improving activation steering in language models with mean-centring. *arXiv preprint arXiv:2312.03813*.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8058–8076.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? A study on solving multilingual tasks with large language models. 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 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2023. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.

Joanito Agili Lopo, Muhammad Ravi Shulthan Habibi, Tack Hwa Wong, Muhammad Ilham Ghozali, Fajri Koto, Genta Indra Winata, Peerat Limkonchotiwat, Alham Fikri Aji, and Samuel Cahyawijaya. 2025. Language surgery in multilingual large language models. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 438–467.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Ercong Nie, Helmut Schmid, and Hinrich Schütze. 2025. Mechanistic understanding and mitigation of language confusion in english-centric large language models. *arXiv preprint arXiv:2505.16538*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Rhitabrat Pokharel, Ameeta Agrawal, and Tanay Nagar. 2026. Cross-lingual activation steering for multilingual language models. *arXiv preprint arXiv:2601.16390*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in English? *arXiv preprint arXiv:2502.15603*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*. OpenReview.net.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Hannah Sterz, Fabian David Schmidt, Goran Glavaš, and Ivan Vulic. 2025. Recover the target language: Language steering without sacrificing task performance. *Preprint*.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2024. Improving instruction-following in language models through activation steering. *arXiv preprint arXiv:2410.12877*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne 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.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Eric Todd, Millicent L. 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*. OpenReview.net.

Yilei Tu, Andrew Xue, and Freda Shi. 2025. Blessing of multilinguality: A systematic analysis of multilingual in-context learning. *arXiv preprint arXiv:2502.11364*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

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.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. *arXiv preprint arXiv:2109.07684*.

Linjuan Wu, Hao-Ran Wei, Huan Lin, Tianhao Li, Baosong Yang, Fei Huang, and Weiming Lu. 2025. Enhancing LLM language adaption through cross-lingual in-context pre-training. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27152–27166, Suzhou, China. Association for Computational Linguistics.

Xie Yunfan, Lixin Zou, Dan Luo, Min Tang, Chenliang Li, Xiangyang Luo, and Liming Dong. 2025. Mitigating language confusion through inference-time intervention. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8418–8431.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024. PLUG: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046, Bangkok, Thailand. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *Preprint*, arXiv:2310.01405.

A Per-language results for Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct

Tables 6 and 7 show the results on the Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct models for all the datasets. Results demonstrate consistent improvements over the source baseline across most languages and datasets.

B Optimal Layer and Scaling Factor per Language

The optimal steering layer and scaling factor α are both task-dependent. For XNLI, layer 10 dominates (11 out of 14 languages), consistent with middle-layer representations encoding the semantic and pragmatic features relevant to natural language inference. For MGSM and MSVAMP, layer 5 is most frequent (7/12 and 5/9 languages respectively), suggesting that mathematical reasoning benefits from steering at earlier layers where surface-level language features are more prominent. The optimal α also varies: $\alpha = 3.0$ dominates for XNLI while MGSM and MSVAMP show more varied and generally lower optimal values, suggesting that natural language inference requires stronger steering than structured reasoning tasks. More details are in Table 3.

C Ablation Studies

C.1 Random Vector Baseline

To validate that our steering vectors capture meaningful language-specific information, we compare against random steering vectors generated from a standard normal distribution. As shown in Table 8, our method consistently outperforms random steering across all datasets and models. The relatively small absolute gap (0.94–1.29 percentage points) is expected given that hyperparameter optimization over layers, scaling factors, and positions can partially adapt even random vectors to the target language through validation performance. Crucially, however, the consistent advantage of our method across all settings confirms that explicitly computed language-specific vectors capture additional meaningful structure beyond what random perturbations can achieve through optimization alone, as further evidenced by the clustering analysis below.

To further validate that our vectors capture genuine language-specific structure, we apply hierarchical clustering to random steering vectors at their optimal layers. Unlike language vectors (Figure 2), random vectors collapse into a near-flat hierarchy with all languages merging at a single high distance ($\sim$1.967), showing no linguistically meaningful substructure. This confirms that the organized clustering of language identity observed in Figure 2 is not an artifact of hyperparameter optimization but reflects language-specific information encoded in our computed vectors.

Dataset	Model	Performance	
		R	Ours
XNLI	Llama-3.1-8B-Instruct	63.37	**64.31**
	Qwen2.5-7B-Instruct	75.32	**75.51**
	Qwen2.5-14B-Instruct	73.37	**74.66**

Table 8: Comparison of random steering vs. our method across models. R = Random Steering.

C.2 Steering Vector Ablations: Data Source and Format

Table 12 shows the accuracy when general non-task specific data or single question-answer pairs are used to compute the language vector. Single QnA uses the same approach as our method but computes vectors from individual question-answer pairs rather than concatenated few-shot sequences, serving as an ablation of our few-shot format for vector computation. Results show that vectors computed from concatenated few-shot QA sequences consistently outperform both FLORES-based vectors and single in-domain QA-pair vectors, suggesting that multi-example contextual composition provides a stronger steering signal than isolated examples or general-domain translation pairs.

C.3 Token Aggregation: Mean Pooling vs. Last Token

In Table 9 we compare mean-pooling against last-token extraction for the hidden state aggregation when computing steering vectors. Using mean tokens leads to a better performance as compared to just using the last token. Mean pooling captures language-level information distributed across the full sequence, whereas last-token extraction is sensitive to positional artifacts and punctuation tokens at sequence boundaries, which may not carry meaningful language-specific signal. This is consistent with our use of mean pooling for steering vector computation, where the goal is to capture aggregate language identity rather than position-specific features.

Language	MGSM				XNLI				MSVAMP			
	B	MFS	Ours	OR	B	MFS	Ours	OR	B	MFS	Ours	OR
Arabic	–	–	–	–	70.96	71.26	**71.56**	74.55	–	–	–	–
Basque	21.43	15.48	**26.19**	26.19	–	–	–	–	–	–	–	–
Bengali	58.33	**66.67**	58.33	65.48	–	–	–	–	65.57	**70.96**	67.07	76.35
Bulgarian	–	–	–	–	79.34	76.65	**80.24**	78.74	–	–	–	–
Catalan	80.95	78.57	**84.52**	85.71	–	–	–	–	–	–	–	–
Chinese	84.52	**88.10**	86.90	79.76	78.14	77.84	77.84	79.34	82.04	83.23	**83.53**	82.34
French	78.57	**80.95**	79.76	84.52	81.44	82.34	82.34	82.63	82.34	**85.03**	84.13	84.43
Galician	79.76	**83.33**	82.14	84.52	–	–	–	–	–	–	–	–
German	83.33	83.33	**85.71**	86.90	81.74	77.54	**82.93**	79.34	79.04	79.94	80.24	81.14
Greek	–	–	–	–	75.45	72.16	**76.95**	70.96	–	–	–	–
Hindi	–	–	–	–	73.95	72.75	**74.85**	73.35	–	–	–	–
Japanese	71.43	**75.00**	75.00	73.81	–	–	–	–	78.74	**79.64**	79.64	79.94
Russian	85.71	85.71	**88.10**	85.71	78.14	79.04	**79.64**	78.74	79.64	**82.93**	81.74	81.44
Spanish	84.52	83.33	**89.29**	86.90	82.63	82.63	**83.53**	81.74	82.04	83.53	**83.83**	84.13
Swahili	19.05	16.67	**26.19**	33.33	53.29	47.31	**55.09**	54.79	49.10	47.90	**50.60**	50.30
Thai	79.76	**82.14**	79.76	85.71	70.96	**73.05**	70.06	72.46	78.14	**79.34**	79.34	78.74
Turkish	–	–	–	–	76.35	74.55	**76.35**	72.75	–	–	–	–
Urdu	–	–	–	–	65.57	62.87	**66.17**	64.67	–	–	–	–
Vietnamese	–	–	–	–	79.34	75.75	**79.64**	79.34	–	–	–	–
Average	68.94	69.94	**71.82**	73.21	74.81	73.27	**75.51**	74.53	75.18	76.94	**76.68**	77.65

Table 6: Detailed per-language accuracy results on MGSM, XNLI, and MSVAMP on Qwen2.5-7B-Instruct. Gray dashes indicate that the dataset does not contain that language. B = Baseline, MFS = Multilingual Few-Shot Baseline, OR = Oracle. Oracle is an upper bound and not for direct comparison. **Bold** indicates the best score per language (excluding Oracle).

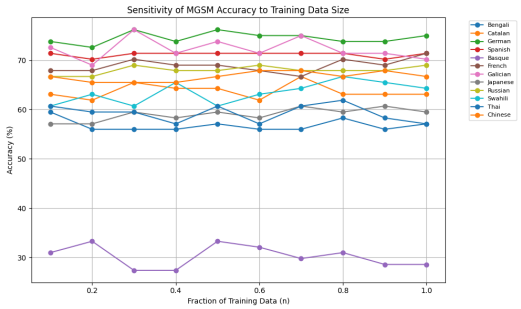

Figure 3: Sensitivity of test accuracy (%) to training data size for MGSM (Llama-3.1-8B-Instruct). *Base* is the unsteered baseline. n denotes the fraction of training examples used to compute steering vectors, with absolute counts in parentheses.

C.4 Sensitivity to Compute Set Size

We analyze the sensitivity of steering performance to the number of examples used to compute the steering vector across a range of data fractions. More details are in Figure 3. We observe substantial heterogeneity in sensitivity to training data size across languages. Low-resource or typologically distant languages (e.g., Basque) exhibit higher variance, whereas high-resource languages (e.g., Chinese, Spanish) remain relatively stable.

D Prompt Templates

We describe the exact prompt structure used for each dataset.

MGSM. The system message and few-shot demonstrations are passed as the USER turn; the model's generation forms the ASSISTANT turn. The system message is fixed across all languages and experiments:

```
You are a helpful assistant that solves
math word problems step by step. Show
your reasoning clearly and end with
'Final answer: ⟨number⟩'.
```

The USER turn contains k=6 demonstrations concatenated with double newlines, each formatted as:

```
Question: ⟨question in source/target
language⟩
Answer: ⟨chain-of-thought in English⟩
Final answer: ⟨numerical answer⟩
```

The test question is appended after the few-shot block as:

```
Question: ⟨question in target language⟩
Answer:
```

A key design choice: demonstration *questions* are in the source language (English) for the baseline,

Language	MGSM				XNLI				MSVAMP			
	B	MFS	Ours	OR	B	MFS	Ours	OR	B	MFS	Ours	OR
Arabic	–	–	–	–	73.35	**75.15**	74.85	78.44	–	–	–	–
Basque	47.62	**54.76**	47.62	59.52	–	–	–	–	–	–	–	–
Bengali	76.19	**80.95**	76.19	77.38	–	–	–	–	75.45	**76.95**	**76.95**	76.05
Bulgarian	–	–	–	–	76.35	**78.14**	77.84	79.64	–	–	–	–
Catalan	84.52	**89.29**	**89.29**	94.05	–	–	–	–	–	–	–	–
Chinese	90.48	91.67	**94.05**	89.29	77.25	76.05	**77.84**	78.74	86.23	84.73	**87.72**	85.63
French	90.48	**91.67**	**91.67**	90.48	78.44	80.24	**82.04**	82.93	85.33	86.83	**87.43**	85.93
Galician	84.52	**86.90**	**86.90**	92.86	–	–	–	–	–	–	–	–
German	92.86	90.48	**96.43**	91.67	78.74	78.74	**79.04**	81.14	85.03	**87.13**	86.83	88.62
Greek	–	–	–	–	72.16	**75.45**	**75.45**	77.54	–	–	–	–
Hindi	–	–	–	–	73.35	**75.75**	74.85	78.14	–	–	–	–
Japanese	88.10	86.90	**92.86**	88.10	–	–	–	–	84.73	82.93	**85.93**	82.93
Russian	91.67	**95.24**	94.05	95.24	75.15	**78.74**	76.65	79.94	83.83	85.33	**86.23**	86.23
Spanish	92.86	88.10	**95.24**	91.67	79.04	78.44	**81.44**	80.24	87.72	**89.22**	88.92	88.32
Swahili	50.00	51.19	**54.76**	57.14	53.29	51.80	**54.79**	58.08	69.46	71.26	**71.56**	73.05
Thai	89.29	86.90	**91.67**	89.29	72.16	72.46	**73.05**	74.25	81.14	81.74	**83.53**	80.54
Turkish	–	–	–	–	71.26	**74.85**	74.55	76.35	–	–	–	–
Urdu	–	–	–	–	63.77	64.07	**65.27**	67.07	–	–	–	–
Vietnamese	–	–	–	–	73.95	74.85	**77.54**	77.84	–	–	–	–
Average	81.55	82.84	**84.23**	84.72	72.73	73.91	**74.66**	76.45	82.10	82.90	**83.90**	83.03

Table 7: Detailed per-language accuracy results on MGSM, XNLI, and MSVAMP for the Qwen2.5-14B-Instructmodel. Gray dashes indicate that the dataset does not contain that language. B = Baseline, MFS = Multilingual Few-Shot Baseline, OR = Oracle. Oracle is the upper bound and is not for a direct comparison. **Bold** indicates the best score per language.

Language	Base	Mean	Last
Bengali	57.14	61.90	53.57
Catalan	64.29	69.05	61.90
German	66.67	75.00	65.48
Spanish	77.38	76.19	71.43
Basque	32.14	36.90	32.14
French	61.90	70.24	60.71
Galician	64.29	73.81	67.86
Japanese	55.95	55.95	55.95
Russian	71.43	72.62	66.67
Swahili	55.95	65.48	59.52
Thai	57.14	61.90	57.14
Chinese	67.86	71.43	66.67
Avg	61.01	65.87	59.92

Table 9: Accuracy using mean-pooling vs. last-token hidden states (Llama-3.1-8B-Instruct, MGSM). Mean = Mean, Last = Last Token.

Language	Resource level	L2 norm
Bengali (bn)	Low-resource	2.47
Basque (eu)	Low-resource	1.83
Thai (th)	Medium-resource	1.73
Swahili (sw)	Low-resource	1.72
Japanese (ja)	High-resource	1.52
Chinese (zh)	High-resource	1.29
Russian (ru)	High-resource	1.20
Catalan (ca)	Low-resource	1.12
Galician (gl)	Low-resource	1.10
German (de)	High-resource	1.09
French (fr)	High-resource	1.02
Spanish (es)	High-resource	0.96

Table 10: L2 norm of the steering vector at layer 10 of Llama-3.1-8B-Instructfor each target language on MGSM, sorted in descending order.

and in the target language for the oracle. Chain-of-thought reasoning and final answers are always in English, since the numerical answer is language-invariant. During steering, demonstration questions remain in English (source baseline setup) and only the hidden states are modified.

XNLI. The system message instructs the model to perform three-way natural language inference:

```
You are a helpful assistant that
performs natural language inference.
Given a premise and a hypothesis,
determine the relationship between them.
The relationship can be: 'entailment'
(hypothesis is true given the premise),
'neutral' (hypothesis might be true or
false), or 'contradiction' (hypothesis
is false given the premise). Answer
with only: entailment, neutral, or
contradiction.
```

Each of the k=6 demonstrations is formatted as:

```
Premise: ⟨premise⟩
Hypothesis: ⟨hypothesis⟩
```

```
Label: ⟨entailment | neutral | contradiction⟩
```

The test example is appended as:

```
Premise: ⟨premise in target language⟩
Hypothesis: ⟨hypothesis in target language⟩
Label:
```

Answer extraction takes only the *first word* of the model's response, checking for an exact match against the three label strings.

Ex. 1 — Gold: contradiction (French, XNLI)	
P	*Pour les aides à l'évacuation du 11 septembre, voir Civilian interview 14 (avril. . .)*
H	*Aucune aide n'est disponible concernant l'évacuation du 11 Septembre.*
Base.	entailment
Ours	**contradiction**
Ex. 2 — Gold: neutral (French, XNLI)	
P	*Au delà de la réputation de Las Vegas, nous n'avons pas vu de preuves expliquant pourquoi les employés y volèrent.*
H	*Les agents ont volé à Las Vegas plusieurs fois dans une courte période.*
Base.	entailment
Ours	**neutral**

Table 11: Qualitative examples from XNLI (French). P = Premise, H = Hypothesis.

MSVAMP. The prompt structure mirrors MGSM with k=6 demonstrations. The system message is identical to MGSM. Each demonstration follows the same `Question / Answer / Final answer` format, with chain-of-thought in English and the numerical answer language-invariant.

E Qualitative Analysis

E.1 Effect of Steering on Model Predictions

Table 11 shows the change in outputs when the output is steered versus when it is not.

E.2 Steering Vector Norms Across Languages

Table 10 reports the L2 norm of the steering vector at layer 10 of the Llama model for each target language on MGSM. Since each steering vector is computed as the mean difference between target and source (English) hidden states, its norm directly measures how far apart the two languages are in the model's representation space.

Languages that are typologically and scripturally similar to English French, Spanish, German, Russian cluster at lower norms, indicating their representations lie close to English in activation space. Languages with distinct scripts or typological structure Bengali, Basque, Thai show markedly larger norms, reflecting greater representational distance from English regardless of resource level. Japanese is a notable example: despite being high-resource, it exhibits the largest norm among that group (1.52), consistent with its non-Latin script and agglutinative morphology. This pattern complements the linguistic structure observed in the hierarchical clustering of steering vectors (Figure 2), where script and typological family rather than resource level alone drive the groupings.

F Hyperparameter Configuration Comparison

We perform a grid search over steering layers $t \in \{5, 10, 15, 20, 25, 30\}$, scaling factors $\alpha \in \{0.5, 1.0, 2.0, 3.0\}$, and four steering position configurations (`on_fewshot`, `after_fewshot`, `on_question`, `entire`) on the validation set $\mathcal{D}_{\text{val}}$. Only configurations achieving validation accuracy above the source baseline are carried forward to evaluation on $\mathcal{D}_{\text{test}}$.

To assess the impact of per-language hyperparameter selection, we compare two configuration levels: (1) **Per-language**: each language uses its own optimal layer, scaling factor, and steering position; (2) **Task-level**: a single configuration is shared across all languages within a task, selected as the one achieving the highest average validation accuracy. Results under both protocols are reported in Table 13. Task-level configurations remain consistently above the source baseline across all datasets, confirming that while per-language selection provides additional gains, the method does not rely on it.

The task-level optimal configurations are: for MGSM, layer 5 with $\alpha = 1.0$ and `on_fewshot`; for XNLI, layer 10 with $\alpha = 3.0$ and `on_fewshot`; for MSVAMP, layer 5 with $\alpha = 0.5$ and `on_fewshot`.

Language	MGSM					XNLI					MSVAMP				
	B	FLORES	Single	Ours	OR	B	FLORES	Single	Ours	OR	B	FLORES	Single	Ours	OR
Arabic	–	–	–	–	–	62.57	61.08	63.17	**62.87**	63.77	–	–	–	–	–
Basque	32.14	27.38	36.90	**36.90**	52.38	–	–	–	–	–	–	–	–	–	–
Bengali	57.14	57.14	60.71	**61.90**	58.33	–	–	–	–	–	57.49	57.49	59.28	**59.58**	61.38
Bulgarian	–	–	–	–	–	56.29	58.68	60.18	**61.68**	66.17	–	–	–	–	–
Catalan	64.29	65.48	66.67	**69.05**	76.19	–	–	–	–	–	–	–	–	–	–
Chinese	67.86	65.48	73.81	**71.43**	72.62	**59.88**	56.29	60.18	59.28	61.98	69.76	69.76	**71.26**	**71.26**	73.35
French	61.90	64.29	67.86	**70.24**	65.48	67.37	71.26	68.86	**72.75**	71.26	71.56	73.35	**73.35**	**74.55**	73.05
Galician	64.29	**75.00**	73.81	73.81	77.38	–	–	–	–	–	–	–	–	–	–
German	66.67	65.48	70.24	**75.00**	71.43	64.07	57.19	67.66	**66.47**	65.27	**71.26**	70.06	**71.26**	**71.26**	76.95
Greek	–	–	–	–	–	68.26	60.18	**70.06**	**70.06**	72.75	–	–	–	–	–
Hindi	–	–	–	–	–	61.38	60.78	64.67	**64.97**	61.98	–	–	–	–	–
Japanese	55.95	**60.71**	**63.10**	55.95	63.10	–	–	–	–	–	63.17	63.17	65.57	**67.96**	70.06
Russian	71.43	67.86	**72.62**	**72.62**	76.19	58.98	53.59	**64.97**	63.77	60.78	68.86	71.56	**72.16**	**72.16**	71.56
Spanish	**77.38**	71.43	76.19	76.19	78.57	**70.66**	**70.66**	69.16	**70.66**	67.37	74.55	73.95	75.45	**75.75**	74.55
Swahili	55.95	61.90	63.10	**65.48**	66.67	52.40	47.01	61.08	**55.99**	56.89	56.29	58.38	57.49	**60.78**	62.87
Thai	57.14	**61.90**	63.10	**61.90**	59.52	59.88	51.20	**62.87**	60.78	70.66	59.58	61.08	63.77	**64.07**	66.77
Turkish	–	–	–	–	–	62.28	64.67	66.17	**65.87**	65.57	–	–	–	–	–
Urdu	–	–	–	–	–	55.09	48.20	56.89	**56.29**	57.19	–	–	–	–	–
Vietnamese	–	–	–	–	–	64.07	63.47	67.96	**68.86**	68.86	–	–	–	–	–
Average	61.01	62.00	65.68	**65.87**	68.16	61.66	58.88	64.56	**64.31**	65.04	65.84	66.53	67.73	**68.60**	70.06

Table 12: Alternative vector configurations per language. B = Baseline, FLORES = vectors computed from general-purpose parallel sentences (Costa-Jussà et al., 2022), Single = our method with single question-answer pairs instead of concatenated few-shot sequences, Ours = our full few-shot method, OR = Oracle (upper bound, not for direct comparison). Bold indicates best among B, FLORES, Single, and Ours.

Language	MGSM					XNLI					MSVAMP				
	B	MFS	Per-Lang	Task-Lvl	OR	B	MFS	Per-Lang	Task-Lvl	OR	B	MFS	Per-Lang	Task-Lvl	OR
Arabic	–	–	–	–	–	62.57	60.78	**62.87**	61.38	63.77	–	–	–	–	–
Basque	32.14	35.70	**36.90**	32.14	52.38	–	–	–	–	–	–	–	–	–	–
Bengali	57.14	55.95	**61.90**	53.57	58.33	–	–	–	–	–	57.49	59.58	**62.87**	59.58	61.38
Bulgarian	–	–	–	–	–	56.29	**61.98**	61.68	60.18	66.17	–	–	–	–	–
Catalan	64.29	64.28	**69.05**	63.10	76.19	–	–	–	–	–	–	–	–	–	–
Chinese	67.86	67.86	**71.43**	70.24	72.62	59.88	**61.38**	59.28	60.78	61.98	69.76	71.26	**73.05**	71.26	73.35
French	61.90	64.29	**70.24**	66.67	65.48	67.37	68.26	**72.75**	68.86	71.26	71.56	73.05	**74.55**	73.65	73.05
Galician	64.29	69.04	**73.81**	73.81	77.38	–	–	–	–	–	–	–	–	–	–
German	66.67	66.67	**75.00**	70.24	71.43	64.07	65.27	**66.47**	65.27	65.27	71.26	70.06	**71.26**	69.16	76.95
Greek	–	–	–	–	–	68.26	66.17	**70.06**	**70.06**	72.75	–	–	–	–	–
Hindi	–	–	–	–	–	61.38	57.78	**64.97**	64.37	61.98	–	–	–	–	–
Japanese	55.95	**61.90**	55.95	**61.90**	63.10	–	–	–	–	–	63.17	67.96	**68.26**	67.66	70.06
Russian	71.43	71.43	**72.62**	67.86	76.19	58.98	**64.37**	63.77	60.48	60.78	68.86	72.16	**72.46**	70.66	71.56
Spanish	**77.38**	70.24	76.19	70.24	78.57	66.77	66.17	**70.66**	68.86	67.37	74.55	73.95	**75.75**	75.75	74.55
Swahili	55.95	63.10	**65.48**	65.48	66.67	52.40	51.20	**55.99**	55.09	56.89	56.29	59.58	**60.78**	57.49	62.87
Thai	57.14	**67.86**	61.90	60.71	59.52	59.88	**64.37**	60.78	60.78	70.66	59.58	64.07	**64.67**	61.98	66.77
Turkish	–	–	–	–	–	62.28	57.19	**65.87**	62.87	65.57	–	–	–	–	–
Urdu	–	–	–	–	–	55.09	55.39	**56.29**	54.49	57.19	–	–	–	–	–
Vietnamese	–	–	–	–	–	64.07	63.17	**68.86**	66.77	68.86	–	–	–	–	–
Average	61.01	63.19	**65.87**	63.88	68.16	61.38	61.68	**64.31**	62.87	65.04	65.84	68.27	**68.60**	67.44	70.06

Table 13: Detailed per-language accuracy results on MGSM, XNLI, and MSVAMP for the Llama-3.1-8B-Instruct model. Gray dashes indicate that the dataset does not contain that language. B = Baseline, MFS = Multilingual Few-Shot Baseline (Tu et al., 2025), OR = Oracle. Oracle is the upper bound and is not for a direct comparison. **Bold** indicates the best score per language excluding the oracle.