Can Dialects Be Steered Like Languages? Sparse Neurons and Distributed Directions in Arabic LLMs

Kareem Elozeiri*[1], Mervat Abassy*[1], Omar Kallas[1], Fahim Dalvi[2], Preslav Nakov[1], Kentaro Inui[1,3,4], Nadir Durrani[2]

[1]Mohamed bin Zayed University of Artificial Intelligence
[2]Qatar Computing Research Institute, Hamad Bin Khalifa University
[3]Tohoku University [4]RIKEN
{kareem.ali, mervat.abassy}@mbzuai.ac.ae

Abstract

A key challenge in Arabic NLP is the scarcity of dialectal data relative to Modern Standard Arabic (MSA), causing LLMs to overproduce MSA and struggle with dialectally accurate generation. From an interpretability perspective, this raises a fundamental question: where and how are dialectal features encoded within model internals, and can these representations be leveraged to improve dialect generation without fine-tuning? This study investigates two complementary inference-time approaches that serve simultaneously as interpretability probes and control mechanisms. First, we conduct a neuron-level analysis, identifying sparse neuron populations that encode dialect-specific features and showing that amplifying or suppressing these neurons can steer model outputs toward target dialects. Second, motivated by the entanglement of dialectal features at the single-neuron level, we apply a vector-steering approach that extracts dialect-specific activation directions and injects them during inference. Together, these methods illuminate the geometry of dialectal knowledge in Arabic LLMs and offer a principled, interpretability-grounded framework for dialect control without requiring dialect-specific fine-tuning.[1]

1 Introduction

Arabic large language models (LLMs) have made strong progress on standard benchmarks, yet remain weak on Arabic dialects (Robinson et al., 2025). This partly reflects their training data: Modern Standard Arabic (MSA) dominates most Arabic LLM pretraining corpora, while dialectal text is comparatively scarce and inconsistently sourced (Bari et al., 2025). As a result, models often default to MSA or produce hybrid outputs that indiscriminately blend dialects. Nacar (2025) documents this pattern in ALLaM 34B, where Hijazi and Egyptian prompts consistently elicit MSA-style responses, limiting the model's usefulness for conversational agents, cultural content generation, and other applications requiring dialectal authenticity.

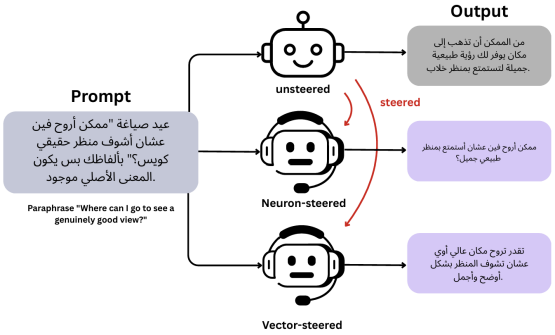


Figure 1: Example of unsteered, neuron-steered, and vector-steered outputs for an Egyptian Arabic prompt. Both steering methods shift the model away from MSA-like responses toward more authentic dialectal generation, with vector steering producing the most colloquial output.

Recent work on interpretability suggests that linguistic behavior in LLMs can be localized or linearly controlled: neurons can encode language-relevant features, and activation-space directions can steer generation toward target attributes (Tang et al., 2024; Rahmanisa et al., 2025; Turner et al., 2024; Rimsky et al., 2024; Chen et al., 2025). These findings raise a natural question for Arabic dialects: *Do dialectal features admit similar localized control, or does the lexical and syntactic continuity among Arabic varieties make them more distributed?* This distinction matters because dialects are not separate languages with clean boundaries; they share vocabulary, morphology, and orthography with MSA and with each other.

This task presents unique challenges: unlike independent languages, Arabic dialects have high lexical and syntactic overlap, and dialectal boundaries are fuzzy rather than discrete (Habash, 2010), making it harder to disentangle dialectal features at the neuron level. To address these open challenges,

* Equal contribution.
[1]The code and artifacts are available at https://github.com/mbzuai-nlp/arabic-dialect-steering

this work investigates two distinct yet complementary inference-time strategies for dialect control in Arabic LLMs. The first investigates whether Arabic LLMs encode dialect-specific neurons and whether manipulating these neurons can steer generation toward the desired dialect. The second takes a broader representational perspective, asking whether dialect-specific directions can be extracted from the model's activation space and injected during inference to achieve more robust dialectal fidelity without fine-tuning or modification.

Research Questions: We ask three questions:

1. Can Arabic dialects be steered in the same way as languages, despite their lexical and syntactic overlap with MSA and with each other?

2. Are dialectal features localized in neuron populations, distributed across residual space, or both?

3. Which inference-time intervention provides more reliable dialect control: neuron-level steering or activation-vector steering?

Our results reveal a nuanced picture: Arabic dialects are neither encoded as fully isolated neuron-level modules nor as entirely diffuse behaviors. Instead, dialectal information is both sparse and distributed. We find that dialect-associated neurons concentrate in late generation-facing layers, with MSA clearly separated from spoken dialects and with substantial sharing among regional dialects. However, these sparse neurons capture only a partial projection of the broader residual-space dialect direction. This explains why neuron steering can reinforce dialectal behavior, but distributed activation-vector steering provides more reliable control. Together, these findings suggest that Arabic dialect variation forms a structured and causally steerable dimension of LLM representation space. To the best of our knowledge, this is the first study to show that Arabic dialects are causally steerable within LLMs through both sparse neuron-level mechanisms and distributed activation-space directions.

2 Methodology

We study two inference-time interventions for Arabic dialect control. Neuron-based steering operates on sparse dialect-associated MLP neurons, while vector steering operates on dialect-specific directions extracted from contrastive dialect–MSA pairs. Both keep model weights fixed and serve as causal probes of dialectal representations.

2.1 Neuron-Based Dialect Steering

We test whether dialectal behavior can be controlled through sparse neuron-level interventions. Following the LAPE-based framework of Tang et al. (2024), we identify dialect-associated MLP neurons and rescale their activations during decoding to steer generation toward target dialects.

Neuron Identification: We treat dialect-associated neurons as those that are frequently active for one Arabic dialect but selective across dialects. Let L be the set of dialects used for extraction, S_k the tokenized corpus for dialect $k \in L$, and $T_k = \{(s,t) : s \in S_k,\ 1 \le t \le |s|\}$ the set of sentence–token-position pairs in that corpus. For neuron j in layer i, with scalar activation $\mathrm{act}(i,j;s,t)$, we estimate its activation probability for dialect k as $\hat{p}^k_{i,j} = \mathbb{E}_{(s,t)\in T_k}[\mathbf{1}\{\mathrm{act}(i,j;s,t) > 0\}]$. We collect these probabilities across dialects as $p_{i,j} = [\hat{p}^k_{i,j}]_{k\in L}$ and normalize the vector to unit ℓ_1 norm, yielding $p'_{i,j}$. The LAPE score is the entropy of this normalized distribution: $\mathrm{LAPE}_{i,j} = -\sum_{k\in L} p'^k_{i,j} \log p'^k_{i,j}$. Low LAPE indicates that a neuron is selective for a small number of dialects. We select a neuron as dialect-associated if it falls in the top $n\%$ of activation probability for at least one dialect and in the bottom $m\%$ of LAPE scores across neurons, where n and m are selection-percentile hyperparameters. These neurons are then used as targets for inference-time steering.

Neuron Steering: We steer the model at inference time by rescaling the activations of selected dialect-associated MLP neurons during decoding, leaving model parameters unchanged. For a target dialect k, let $\mathcal{D}_k(i)$ denote the set of selected target-dialect neurons in layer i. We define $\mathcal{D}_{\mathrm{MSA}}(i)$ as the selected MSA neuron set and, optionally, $\mathcal{D}_{\mathcal{C}}(i)$ as the union of selected neurons for non-target competitor dialects. Let $a_{i,t,j}$ be the activation of neuron j in layer i at decoding step t. We apply the following multiplicative update:

$$\tilde{a}_{i,t,j} = \lambda_{i,j} a_{i,t,j},$$

where

$$\lambda_{i,j} = \begin{cases} \alpha, & j \in \mathcal{D}_k(i), \\ \gamma, & j \in \mathcal{D}_{\text{MSA}}(i), \\ \gamma_{\text{comp}}, & j \in \mathcal{D}_{\mathcal{C}}(i), \\ 1, & \text{otherwise.} \end{cases}$$

Here, α is the target-dialect amplification factor, γ is the MSA suppression factor, and γ_{comp} is the optional competitor-dialect suppression factor. We use $\alpha > 1$ to strengthen target-dialect neurons and $\gamma, \gamma_{\text{comp}} < 1$ to suppress MSA and non-target dialect neurons. Thus, decoding is biased toward the target dialect while weights remain unchanged.

2.2 Vector Steering

While neuron-level steering offers fine-grained interpretability, dialect features in Arabic LLMs may be distributed across many neurons simultaneously, making isolated neuron interventions insufficient. To address this, we explore a complementary representation-level approach: *vector steering* (Turner et al., 2024; Rimsky et al., 2024), which operates on the full activation space of a layer rather than on individual neurons.

Steering Vector Extraction. For a target dialect k and a contrastive variety (MSA), we collect a parallel corpus of sentence pairs $\{(s_i^k, s_i^{\text{MSA}})\}_{i=1}^N$. Each sentence is passed through the model independently, and we extract the mean hidden-state activation across all response tokens at layer ℓ, denoted $h^\ell(s)$. The dialect steering vector is then computed as the mean difference between dialect and MSA activations:

$$v_\ell^k = \frac{1}{N} \sum_{i=1}^{N} \left(h^\ell(s_i^k) - h^\ell(s_i^{\text{MSA}}) \right)$$

This difference vector encodes the directional shift in activation space corresponding to moving from MSA toward dialect k, capturing distributed dialect features that may not be localized to any individual neuron. Crucially, because activations are extracted from the model's response tokens rather than the input token, the vector reflects the model's *generated* dialectal behavior rather than its encoding of the input, making it a more direct target for generation-time intervention.

Inference-Time Injection: During generation, we register a forward hook at layer ℓ that intercepts the activation tensor at every token generation step and adds the scaled steering vector in-place, $h_{\text{new}}^\ell = h^\ell + \alpha \cdot v_\ell^k$, where α is a scalar coefficient controlling the strength of the intervention. Downstream layers then receive the modified activations and generate dialect-influenced tokens accordingly. No retraining or parameter updates are required.

3 Experiments

3.1 Experimental Setup

Models. We experiment with two Arabic-centric large language models: ALLaM-7B-Instruct-preview (Bari et al., 2025) and Fanar-1-9B-Instruct (Fanar Team et al., 2025). Both are instruction-tuned specifically for Arabic. ALLaM comprises 32 transformer layers, while Fanar has 42. Experimenting across both models allows us to assess whether findings generalize across architectures trained on different Arabic corpora and with different instruction-tuning procedures.

Dialects and data. Ablation experiments are conducted on Egyptian Arabic (Cairo) and Moroccan Arabic (Rabat), as these dialects offer both larger amounts of parallel data for steering vector extraction and direct evaluation availability in the benchmark used in this work. They were also chosen because they represent linguistically diverse dialects with substantial variation between them, allowing for a more informative analysis of the proposed methods.

Final model performance is then evaluated across four broader dialect groups: Egyptian, Moroccan, Levantine, and Gulf Arabic. Steering vectors and dialect-specific neurons are extracted using parallel dialect data from the MADAR corpus. Specifically, Cairo, Rabat, Beirut, and Doha each provide 12,000 sentence pairs, which are used for steering vector extraction and neuron identification. In contrast, Riyadh and Aleppo provide only 2,000 sentence pairs each.

For evaluation, Cairo and Rabat are mapped directly to the Egyptian and Moroccan benchmark subsets. Beirut and Aleppo are evaluated using the Syrian subset, as they are the closest available dialects in the benchmark, while Doha and Riyadh are evaluated on the Saudi subset.

Evaluation. Primary evaluation uses the LLM-as-a-judge protocol described in Section 3.2. We additionally evaluate using the AL-QASIDA benchmark as a complementary automatic metric. For the best-performing model configurations for each

dialect, human annotations are collected to measure inter-rater agreement with the LLM judge.

3.2 LLM-as-a-Judge

We assess generation quality, using Gemini 2.5 Flash (Comanici et al., 2025) as the judge. Each response is evaluated independently against the target dialect across four integer-scored (1–5) dimensions: *dialect authenticity*, *coherence*, *Arabic fluency*, and *MSA formality*. The first three serve as quality measures while MSA formality is used diagnostically to capture drift toward formal Modern Standard Arabic. We report mean scores per dimension over all judged samples. The full prompt, rubric, and implementation details are in Appendix D.1.

3.3 Human Evaluation

For each human-evaluation task, we recruited eight Arabic annotators in pairs per dialect. They used the same 1–5 dimensions as the LLM-as-a-judge setup. The model identities are hidden from the annotators to reduce bias. For each output and metric, we first average the two annotator scores and then average these item-level scores across all samples. We report the human average score as the mean of fluency, coherence, and dialect authenticity, and report MSA formality separately.

3.4 AL-QASIDA Evaluation

We evaluate using the AL-QASIDA framework (Robinson et al., 2025), scoring responses with the ADI2 metric:

$$\text{ADI2}(y) = P(y \text{ is dialectal}) \times P(y \text{ is dialect } C)$$

which jointly penalizes MSA responses and responses in the wrong dialect. We also report macro-ADI2, which aggregates dialect probabilities at the regional level to account for overlap between geographically proximate varieties. We use the monolingual generation task of the benchmark throughout, as it best reflects real-world usage and aligns with our inference-time intervention. Full details on the framework, prompt construction, and corpora are in Appendix E.

3.5 Neuron-based Steering

Neuron identification. Following the selection procedure defined in Section 2.1, we select neurons that fall in the top 5% of activation probability for at least one dialect and in the bottom 1% of LAPE-score distribution across all neurons. These thresholds keep neurons that are both frequently active for a dialect and highly selective across dialects.

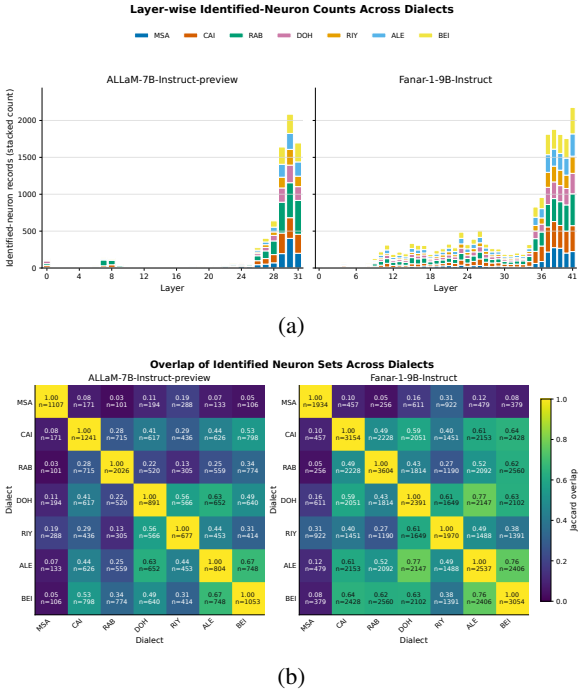


Figure 2: Dialect-specific neuron analysis for ALLaM and Fanar across MSA, Cairo, Rabat, Doha, Riyadh, Aleppo, and Beirut. Top: identified neurons concentrate in later layers. Bottom: Jaccard overlap shows stronger sharing among spoken dialects than with MSA.

Layer-wise distribution. Figure 2a shows that dialect-specific neurons are depth-localized rather than uniformly distributed. In ALLaM, they form a narrow late-layer bottleneck, with layers 29–31 accounting for 69.5% of all dialect-neuron records and layer 30 as the main peak. Fanar shows a similar late-layer pattern, but spread over a wider terminal band, with layers 37–41 accounting for 50.6% of records, plus a visible mid-layer tail. This suggests that dialectal behavior is not primarily encoded as early surface cues such as orthography or local lexical features. Instead, it emerges closer to the generation-facing layers where internal Arabic representations are mapped into next-token choices, making these neurons plausible control points for dialectal style, lexical choice, and morphosyntactic realization.

Cross-dialect overlap. Figure 2b shows that the identified neuron sets are neither isolated dialect modules nor a single generic Arabic circuit. MSA is consistently separated from the dialects, with average MSA–dialect overlap of only 0.09 in ALLaM and 0.14 in Fanar, compared with dialect–dialect overlap of 0.40 and 0.55. Since all varieties

share substantial Arabic lexical and orthographic structure, this separation suggests that the selected neurons are not merely generic Arabic detectors. Instead, dialects share part of their internal representation while retaining distinct neuron sets. High-overlap pairs such as Aleppine–Beiruti and Doha–Riyadh suggest reusable regional subcircuits that may support transfer, while low overlap with MSA marks sharper boundaries where MSA-based interventions are less likely to transfer.

Ablations. We ablate the target-dialect amplification factor α, the MSA suppression factor γ, and the competitor-dialect suppression factor γ_{comp}, varying one parameter while keeping the other two fixed. Ablations are run on Egyptian and Moroccan Arabic across both models. Our results show that α is the main parameter affecting dialectal generation performance, with the best setting being $\alpha = 2.0$ for ALLaM and $\alpha = 4.0$ for Fanar. Varying γ and γ_{comp} does not yield consistent improvements and in some cases degrades performance. Full ablation settings and results are reported in Appendix A.

3.6 Vector Steering

To construct the parallel sentence pairs $\{(s_i^k, s_i^{\text{MSA}})\}$ required by Equation 1, we use the MADAR corpus. Each dialect sentence is placed in the assistant turn of the chat template, formatted as if produced by the model, so that hidden states are extracted from response token positions, which is consistent with how vectors are applied at generation time.

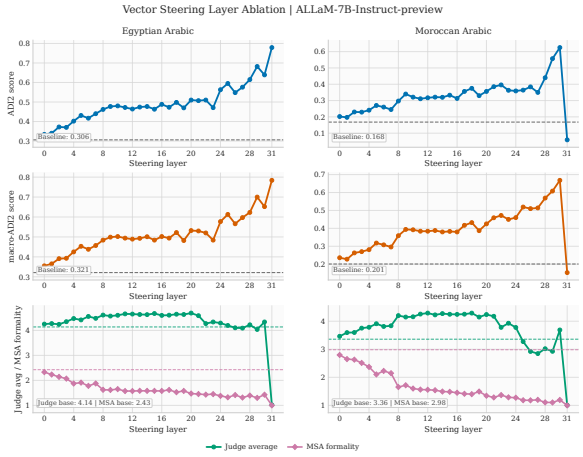
Before applying steering vectors at inference time, we first ask a more fundamental interpretability question: do the hidden states of Arabic LLMs encode dialect identity in a structured and geometrically separable way? If dialect-specific directions exist in activation space, they should be recoverable from the mean hidden-state representations used to construct our steering vectors. We investigate this through PCA of the extracted representations at a fixed intermediate layer across three Arabic LLMs (full analysis in Appendix B.1). We find that dialects form geographically coherent clusters in representation space, with dialect varieties occupying distinct regions, and that representational distance roughly tracks real-world linguistic proximity. This structured geometry confirms that dialect identity is encoded in a recoverable and separable way, motivating the use of mean activation differences as steering directions.

Ablations. We conduct ablations across layers, coefficients and number of steered tokens.

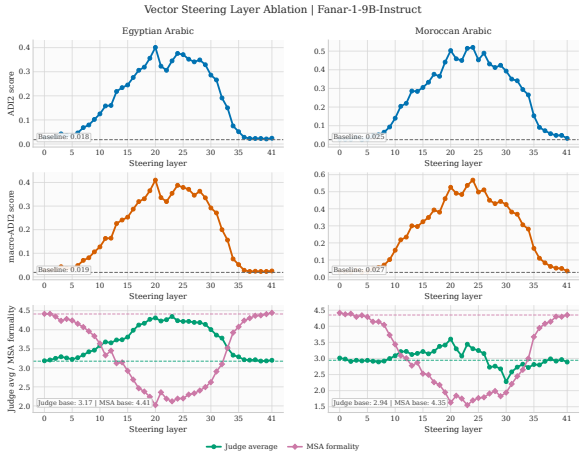
Layer selection. All layer ablations are conducted with a fixed steering coefficient of $\alpha = 2$. As shown in Figure 3, the two models reveal strikingly different dialectal encoding profiles. Fanar concentrates dialect information in its middle layers, with both ADI2 and macro-ADI2 peaking around layers 19–24; the LLM judge corroborates this, showing a corresponding rise in dialect quality scores that peaks in the same region before declining. ALLaM tells a different story: judge scores remain broadly stable across most layers, suggesting that generation quality is largely insensitive to intervention depth, while ADI2 rises monotonically, most sharply at the final layers, indicating that benchmark gains at depth come without a corresponding perceptual improvement. In both models, MSA formality moves inversely with dialect quality scores, confirming that the steering vector shifts register rather than introducing unrelated artifacts. Based on the highest LLM-as-a-judge score, we select candidate layers from the high-quality plateau of each model (layers 19–24 for Fanar, 11–20 for ALLaM) for the subsequent coefficient ablation.

Coefficient selection. α is swept over $\{1, 2, 3, 4, 5\}$ across the candidate layers (Figure 9, Appendix B.2). ALLaM and Fanar differ notably in robustness: Fanar's Egyptian Arabic shows no divergence between ADI2 and judge scores even at $\alpha = 5$, whereas ALLaM degrades in coherence at higher coefficients. Moroccan Arabic is more sensitive than Egyptian Arabic for both models, with judge quality declining while ADI2 continues to rise. The best layer–coefficient pair per model per dialect is carried forward to all subsequent experiments.

Token budget. Using the optimal layer and coefficient per model–dialect pair, steering is restricted to the first N decoded tokens, with N increased in steps of 10 (Appendix B.3). The two models behave differently: ALLaM commits to the target dialect after minimal intervention, with judge scores and ADI2 plateauing from as few as 10–20 tokens onward, suggesting the model sustains the dialect autonomously once primed. Fanar requires sustained injection, with ADI2 and macro scores rising steadily with N, consistent with its near-zero unsteered baseline. Full-sequence steering on Moroccan Arabic for Fanar reveals a tension between dialect authenticity and output quality that ADI2



(a) ALLaM-7B-Instruct-preview



(b) Fanar-1-9B-Instruct

Figure 3: Layer ablation on Egyptian and Moroccan Arabic with a fixed coefficient $\alpha = 2$. ADI2, macro-ADI2, LLM judge average, and MSA formality are shown across all layers for each model. Dashed lines indicate unsteered baselines.

alone would not expose.

Extraction-size stability. We test whether steering directions are stable across extraction-corpus sizes by re-estimating vectors from 1k–12k parallel pairs and comparing layer-wise cosine similarity. The directions stabilize by 4k examples at the selected steering layers, suggesting that vector orientation is robust to sampling variation; full results are in Appendix B.4.

4 Results

We evaluate two generation settings. In the *mono-dialect* setting, the prompt is written in the target dialect, reflecting the natural use case where a user writes in a dialect and expects a response in kind. In the *MSA-prompt* setting, the prompt is written in MSA while the response is steered toward a target dialect, testing whether internal steering can override the prompt variety rather than simply preserve it.

Final evaluation setup. All methods use deterministic decoding with a maximum of 128 new tokens. The unsteered baseline receives the original prompt without intervention. Neuron steering uses the LAPE-selected neurons with $\alpha = 2.0$ for ALLaM and $\alpha = 4.0$ for Fanar, without MSA or competitor suppression. Vector steering injects response-side dialect–MSA directions for the first 30 generated tokens, following the token ablation results. This token budget is stable for ALLaM and avoids the quality degradation observed for Fanar under longer Moroccan steering. For Egyptian and Moroccan, we report the best layer and coefficient settings from the ablations. For non-ablated dialects, we use layer 20 as the default based on its strong performance across the ablated settings.

4.1 Mono-Dialect Prompts

Table 1 shows that vector steering demonstrates clear effectiveness, matching or surpassing explicit prompting on judge averages for Fanar across all dialects and for ALLaM on Egyptian and Moroccan Arabic, despite explicit prompting (Details in Appendix F)leading on raw ADI2 and macro-ADI2. This dissociation reinforces that ADI2 is not always a representative metric, as it does not account for coherence or overall output quality. For ALLaM on Saudi and Syrian, explicit prompting retains a judge advantage, which is partly explained by the fact that vector steering parameters were ablated and optimized on Egyptian and Moroccan only and applied to these dialects without further tuning. Neuron steering improves over the unsteered baseline, particularly for Egyptian and Moroccan for ADI2 scores but shows no improvement in LLM as a judge scores. Human evaluation (Table 2) broadly validates these findings, with a minor fluency tradeoff on Moroccan Arabic for Fanar. Together, these results support a key interpretability finding: dialectal behavior is partly localized in identifiable neurons, but robust control requires the broader distributed direction captured by vector steering. More fine-grained results are provided in Appendices D.2 and G.

4.2 MSA Prompts

We evaluate on 300 MSA prompt synthetic samples, steering each toward the target dialects. Neuron steering failed entirely, with dialect authenticity

Model	Method	ADI2				macro-ADI2				Judge avg.			
		EGY	MOR	SAU	SYR	EGY	MOR	SAU	SYR	EGY	MOR	SAU	SYR
ALLaM-7B-Instruct	Unsteered	0.306	0.168	0.036	0.094	0.321	0.201	0.084	0.252	4.137	3.359	3.580	3.873
	Explicit	**0.631**	**0.518**	**0.194**	**0.260**	**0.640**	**0.570**	**0.387**	**0.613**	4.677	4.469	**4.298**	**4.527**
	Neuron	0.468	0.308	0.027	0.062	0.471	0.318	0.072	0.234	3.978	3.766	3.282	3.494
	Vector	0.512	0.310	0.124	0.156	0.532	0.380	0.252	0.407	**4.712**	**4.712**	4.151	4.470
Fanar-1-9B-Instruct	Unsteered	0.018	0.025	0.004	0.006	0.019	0.027	0.022	0.017	3.175	2.944	3.462	3.172
	Explicit	0.383	0.345	**0.028**	**0.158**	0.388	0.347	**0.076**	**0.279**	3.986	3.599	3.396	3.509
	Neuron	0.119	0.070	0.002	0.007	0.119	0.078	0.020	0.017	3.029	2.580	3.244	2.974
	Vector	**0.384**	**0.410**	0.010	0.044	**0.398**	**0.428**	0.035	0.200	**4.446**	**4.060**	**3.536**	**3.648**

Table 1: Mono-dialect results across automatic and LLM-as-a-judge metrics. Higher is better for ADI2, macro-ADI2, and judge average. EGY, MOR, SAU, and SYR denote Egyptian, Moroccan, Saudi, and Syrian Arabic, respectively. Neuron and Vector denote the best neuron-steering and vector-steering configurations. Within each model, metric, and dialect, the best score is shown in bold and the second-best score is underlined.

collapsing to the minimum judge score and ADI2 reaching zero, confirming that sparse interventions can reinforce but not induce dialectal generation. Only vector steering results are therefore reported in Table 3. Vector steering produces dialectal signal across all targets, though performance is lower than in the mono-dialect setting with ALLaM being more responsive than Fanar. Crucially, the fact that models produce dialectal output despite receiving MSA prompts shows that vector steering exerts influence at the activation level, overriding the prompt's linguistic register. These results reinforce the mono-dialect finding that distributed residual-space directions are more effective than sparse neuron interventions.

5 Analysis and Discussion

To explain the gap between neuron and vector steering, we compare the residual dialect directions used by vector steering with the subspace spanned by the down-projection directions of the LAPE-selected MLP neurons. This yields a residual-subspace coverage score measuring how much of each dialect vector is captured by the sparse neuron set. Formal details and random-baseline tests are provided in Appendices C.1 and C.2.

Figure 4 shows that LAPE-selected neurons capture a meaningful but incomplete projection of the residual dialect directions. Although the selected neurons occupy less than 1% of MLP intermediate dimensions, they explain a non-trivial portion of the dialect direction, especially for Cairo and Rabat. However, most of the residual direction remains outside this sparse subspace. This explains the main empirical gap: neuron steering provides localized interpretability, while vector steering better captures the distributed activation shift needed for robust dialect control.

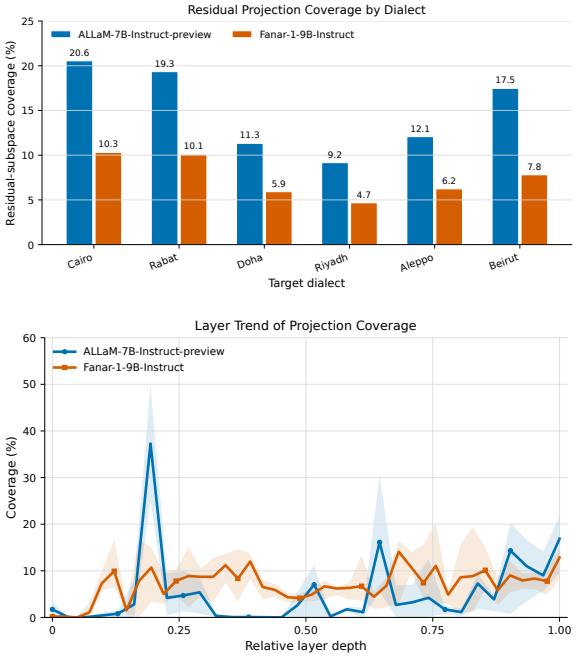


Figure 4: Residual-subspace coverage of vector-steering directions by LAPE-selected neuron output subspaces. Top: aggregate coverage by dialect. Bottom: mean coverage across normalized layer depth, with shaded variation across dialects. Higher values indicate stronger overlap between sparse neuron subspaces and distributed dialect directions.

6 Related Work

Our work on dialectal Arabic control in LLMs sits at the intersection of Arabic NLP and mechanistic interpretability.

Evaluation of Arabic linguistic variation: LLMs remain weak on dialectal Arabic, often defaulting to MSA rather than generating dialectal content (Robinson et al., 2025). Prior work analyzes Arabic model internals across varieties (Abdelali et al., 2022) and shows that MSA-to-dialect transfer is uneven, geographically shaped, and prone to negative interference (Khalak et al., 2026). Relatedly, Elshabrawy et al. (2025) find that excessive entanglement with MSA can hurt

Model	Dialect	Layer	α	Human avg. ↑			MSA formality ↓		
				Unst.	Neu.	Vec.	Unst.	Neu.	Vec.
ALLaM-7B-Instruct	EGY	20	2	3.698	4.111	**4.231**	3.195	2.470	**1.795**
	MOR	19	1	3.300	3.621	**3.628**	4.165	3.905	**3.722**
	SAU	20	2	3.538	3.438	**4.066**	4.410	4.473	**3.005**
	SYR	20	2	3.693	3.724	**4.138**	3.177	3.075	**1.945**
Fanar-1-9B-Instruct	EGY	23	3	3.506	3.339	**3.858**	4.953	4.430	**2.393**
	MOR	20	2	**3.005**	2.864	2.657	4.340	3.998	**2.707**
	SAU	20	2	**3.333**	3.259	3.285	4.723	4.842	**4.620**
	SYR	20	2	3.526	3.221	**3.617**	4.292	4.210	**2.825**

Table 2: Human evaluation summary for mono-dialect outputs. Each score is averaged over two blinded annotators and then over samples. Human avg. averages fluency, coherence, and dialect authenticity, while lower MSA formality is better. Layer and α denote the vector-steering layer and coefficient. Bold and underline mark the best and second-best values within each model and dialect. Unst., Neu., and Vec. denote unsteered, neuron steering, and vector steering.

Dialect	ALLaM-7B-Instruct			Fanar-1-9B-Instruct		
	ADI2	macro	Judge	ADI2	macro	Judge
EGY	0.215	0.221	4.523	0.215	0.221	3.884
LEB	0.108	0.291	4.014	0.108	0.291	3.511
SYR	0.080	0.219	4.143	0.080	0.219	3.620
SAU	0.046	0.102	3.918	0.046	0.102	3.611
QAT	0.069	0.158	3.676	0.069	0.158	3.522
MOR	0.293	0.350	4.321	0.102	0.139	3.640

Table 3: MSA-to-dialect vector-steering results using layer 20 and 30 steered tokens. Higher is better for all metrics. Judge denotes the average LLM-as-a-judge score. EGY, LEB, SYR, SAU, QAT, and MOR denote Egyptian, Lebanese, Syrian, Saudi, Qatari, and Moroccan Arabic.

dialect generation and use subspace interventions to decouple dialect modeling from the dominant standard variety. Benchmarks such as ARADiCE further enable fine-grained evaluation of dialectal and cultural capabilities (Mousi et al., 2025).

Neuron-level language control and activation steering: Tang et al. (2024) identify language-specific neurons with LAPE and show that activating them can steer output language. Subsequent work shows that amplifying such neurons improves steering (Rahmanisa et al., 2025), while language arithmetic extends LAPE-based analysis by systematically activating or suppressing language neurons and showing overlap among related languages (Gurgurov et al., 2025).

A foundational line of work has established that high-level concepts can be extracted as linear directions in activation space and used to steer model behavior at inference time. Turner et al. (2024) introduced activation addition, showing that adding a contrastive steering vector to residual stream activations reliably shifts model behavior toward a target concept. Rimsky et al. (2024) found that mean differences between contrastive activations at the last token position consistently outperform other extraction strategies. Zou et al. (2025) demonstrated that traits such as honesty and toxicity can be identified and manipulated via linear directions. Chen et al. (2025) extended this paradigm to personality traits, showing that persona vectors extracted via contrastive prompting can monitor and control sycophancy, hallucination, and malicious behavior, and that activations from response tokens yield more effective steering directions than prompt tokens. Aneja et al. (2026) show that personality-related semantic directions can transfer across model variants and regulate emergent misalignment.

Unlike prior work on cross-lingual transfer, we operate within a single language family where dialectal boundaries are fuzzy and MSA overlap is high. We show that Arabic LLMs nonetheless encode dialect identity as structured geometric properties of their hidden states, and that contrastive activation directions can steer generation toward a target dialect at inference time, revealing intra-linguistic variation as an underexplored dimension of LLM representation space.

7 Conclusion and Future Work

We explored the mechanistic interpretability of Arabic dialect encoding in LLMs, finding that dialectal information is neither fully localized nor entirely diffuse, with dialect-associated neurons concentrating in late layers yet capturing only a partial projection of the broader residual-space dialect direction. We applied both neuron-based and distributed vector steering, with vector steering providing more reliable dialect control across both models, even when overriding MSA prompts.

Future work should extend this analysis to additional dialects and mixed-dialect settings, explore adaptive and compositional steering methods, and investigate how dialect representations interact with other stylistic and sociolinguistic attributes in multilingual and multimodal models.

Limitations

Our experiments focus on a limited set of Arabic dialects and two Arabic-centric LLMs, and results may not fully generalize to other dialects, architectures, or multilingual models. In addition, evaluation of dialectal generation remains inherently challenging: automatic metrics and LLM-as-a-judge scores capture complementary aspects of performance, but neither provides a complete measure of dialect authenticity or sociolinguistic naturalness. Human evaluations are inherently subjective, and inter-annotator agreement, while reasonable, reflects the ambiguity inherent in dialect perception. Finally, while our analysis identifies interpretable dialect-related structures, the relationship between localized neuron representations and distributed residual-space features remains only partially understood, and further work is needed to characterize these mechanisms more precisely.

Ethics and Broader Impact

This work aims to improve dialectal inclusivity in Arabic LLMs by reducing the strong bias toward Modern Standard Arabic and enabling more natural generation for underrepresented dialect communities. We focus on inference-time steering methods that do not require retraining or modification of model parameters, which may lower the computational cost of dialect adaptation and improve accessibility for low-resource varieties.

At the same time, dialect steering technologies may be misused to imitate regional linguistic styles in deceptive or manipulative contexts. Our methods are intended for research on controllable and interpretable language generation, not for impersonation or misinformation. In addition, Arabic dialect boundaries are inherently fluid and socially complex, and automatic evaluations may not fully capture sociolinguistic authenticity or community perceptions of dialect use. We therefore encourage future work incorporating broader human evaluation across speaker communities and dialect backgrounds.

Annotations were annotated by trained Arabic annotators recruited through a thirdparty company and compensated at the standard hourly rate for their location. Annotators signed non-disclosure agreements prior to participation. We provide clear task instructions, avoid collecting annotator personal data beyond what is required for payment and administration by the vendor, and restrict dataset release to permitted uses consistent with the underlying licenses.

References

Ahmed Abdelali, Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. 2022. Post-hoc analysis of Arabic transformer models. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 91–103, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Krishak Aneja, Manas Mittal, Anmol Goel, Ponnurangam Kumaraguru, and Vamshi Krishna Bonagiri. 2026. Intrinsic guardrails: How semantic geometry of personality interacts with emergent misalignment in llms. *Preprint*, arXiv:2605.10633.

M Saiful Bari, Yazeed Alnumay, Norah Alzahrani, Nouf Alotaibi, Hisham Alyahya, AlRashed AlRashed, Faisal Mirza, Shaykhah Alsubaie, Hassan Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman I Alsubaihi, Maryam Al Mansour, Saad Hassan, Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. Allam: Large language models for arabic and english. In *International Conference on Learning Representations*, volume 2025, pages 34179–34214.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Mahmoud El-Haj. 2020. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France. European Language Resources Association.

Ahmed Elshabrawy, Hour Kaing, Haiyue Song, Alham Fikri Aji, Hideki Tanaka, Masao Utiyama, and Raj Dabre. 2025. When alignment hurts: Decoupling representational spaces in multilingual models. *Preprint*, arXiv:2508.12803.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform.

Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Baeumel, 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, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*, volume 3 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Abdulmuizz Khalak, Abderrahmane Issam, and Gerasimos Spanakis. 2026. From FusHa to folk: Exploring cross-lingual transfer in Arabic language models. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 196–209, Rabat, Morocco. Association for Computational Linguistics.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

Omer Nacar. 2025. Ui-level evaluation of allam 34b: Measuring an arabic-centric llm via humain chat. *Preprint*, arXiv:2508.17378.

Inaya Rahmanisa, Lyzander Marciano Andrylie, Mahardika Krisna Ihsani, Alfan Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. 2025. Unveiling the influence of amplifying language-specific neurons. 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 919–968, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

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, Bangkok, Thailand. Association for Computational Linguistics.

Nathaniel Romney Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. ALQASIDA: Analyzing LLM quality and accuracy systematically in dialectal Arabic. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22048–22065, Vienna, Austria. Association for Computational Linguistics.

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.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2025. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

Appendix

A Neurons-Based Steering Ablations

We conducted ablation experiments to assess the contribution of each neuron-steering parameter and to select an effective steering configuration. Specifically, we varied the target-dialect amplification factor α, the MSA suppression factor γ, and the competitor-dialect suppression factor γ_{comp}. In each ablation, one parameter was varied while the remaining two were held fixed, allowing us to isolate the effect of each component.

Experiments were conducted using ALLaM-7B-Instruct-preview and Fanar-1-9B-Instruct, with Egyptian Arabic and Moroccan Arabic as target dialects.

A.1 Target Dialect Coefficient Ablations

For this ablation, we vary the target-dialect amplification coefficient α while fixing the MSA suppression coefficient and the non-target competitor suppression coefficient to $\gamma = \gamma_{\text{comp}} = 0.9$. This isolates the effect of strengthening target-dialect neurons during decoding. As shown in Figure 5, ALLaM exhibits a non-monotonic response to α. Moderate amplification improves dialectal performance, with $\alpha = 2.0$ providing the best overall trade-off across Egyptian Arabic and Moroccan Arabic. Larger values lead to a sharp degradation in LLM-as-a-judge scores, particularly fluency and coherence, suggesting that excessive amplification can distort generation quality even when dialectal signals are strengthened.

For Fanar, increasing α produces stronger gains in the automatic dialectal metrics, especially ADI2 scores, which improve consistently as the coefficient increases. However, the LLM-as-a-judge results show that very large amplification reduces fluency and coherence, while dialect authenticity improves only modestly. We therefore select $\alpha = 4.0$ for Fanar as a balanced setting: it substantially improves dialectal metrics while avoiding the larger quality degradation observed at $\alpha = 5.0$. Overall, these results indicate that the target-dialect amplification coefficient is a main driver of neuron-steering performance, but its optimal value is model-dependent.

A.2 MSA Neurons Suppression Ablations

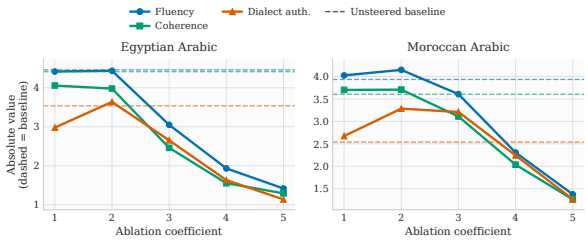
In this ablation, we vary the MSA suppression coefficient γ while fixing the target-dialect amplification coefficient to the best value selected in the previous ablation, namely $\alpha = 2.0$ for ALLaM and $\alpha = 4.0$ for Fanar. We also fix the competitor-dialect suppression coefficient to $\gamma_{\text{comp}} = 0.9$. This setup isolates the effect of suppressing MSA-specific neurons. As shown in Figure 6, varying γ has a weaker and less consistent effect than varying the target-dialect amplification coefficient. For ALLaM, weaker MSA suppression, represented by larger values of γ, generally preserves or improves performance, while strong suppression does not provide clear gains and can reduce fluency and coherence, especially for Egyptian Arabic.

For Fanar, the effect of γ is also limited and non-monotonic. Although automatic dialect scores remain above the unsteered baseline across most settings, no single MSA suppression value yields consistent improvements across both dialects and evaluation metrics. The LLM-as-a-judge scores further indicate that suppressing MSA neurons does not reliably improve dialect authenticity and may coincide with lower fluency and coherence. Overall, the results suggest that MSA-neuron suppression is not a driver of neuron-steering performance. The gains are primarily attributable to target-dialect neuron amplification, while aggressive MSA suppression can be unnecessary or harmful.

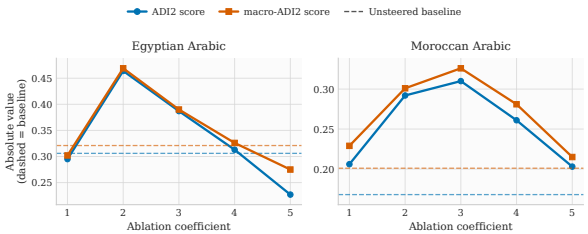
A.3 Non-Target Dialects Neurons Ablations

In this ablation, we vary the non-target dialect suppression coefficient γ_{comp} while fixing the target-dialect amplification coefficient to the best value selected earlier, namely $\alpha = 2.0$ for ALLaM and $\alpha = 4.0$ for Fanar. We also fix the MSA suppression coefficient to $\gamma = 1.0$. Since γ_{comp} is a multiplicative suppression factor, smaller values correspond to stronger suppression, while $\gamma_{\text{comp}} = 1$ corresponds to no suppression of non-target dialect neurons.

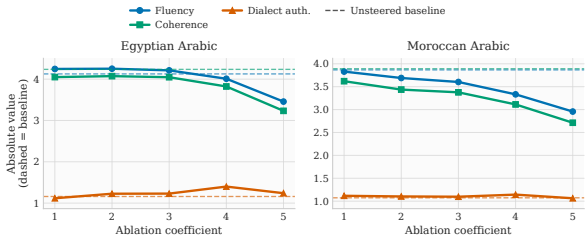
As shown in Figure 7, varying γ_{comp} has a limited and inconsistent effect compared with target-dialect amplification. For ALLaM, automatic dialect scores remain above the unsteered baseline across settings, but stronger non-target suppression does not consistently improve LLM-as-a-judge scores. For Fanar, automatic scores also remain above baseline, yet the LLM-as-a-judge results show no reliable gains in fluency, coherence, or dialect authenticity. These results suggest that suppressing non-target dialect neurons is not a primary driver of steering performance, and that aggressive non-target suppression is generally unnecessary.



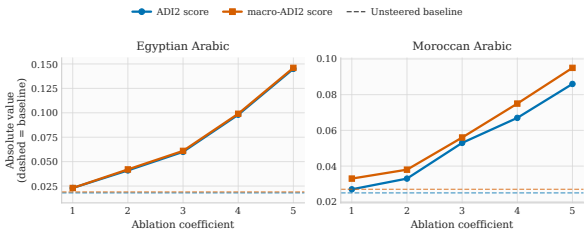
(a) ALLaM-7B-Instruct-preview LLM-as-a-judge scores.



(b) ALLaM-7B-Instruct-preview ADI2 scores.

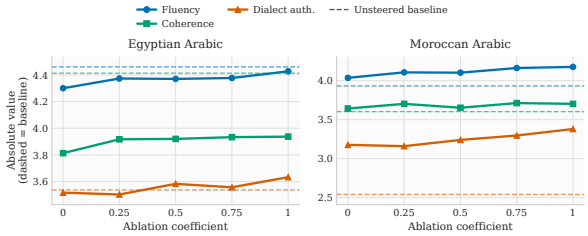


(c) Fanar-1-9B-Instruct LLM-as-a-judge scores.

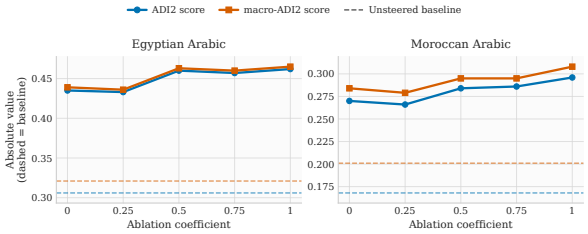


(d) Fanar-1-9B-Instruct ADI2 scores.

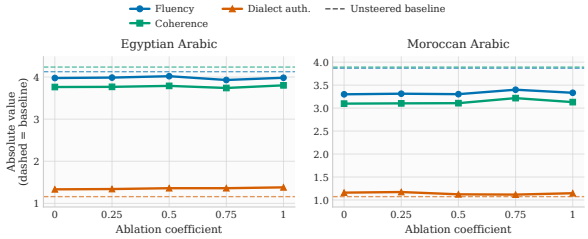
Figure 5: Neuron-steering target-dialect amplification factor ablation results for **ALLaM-7B-Instruct-preview** and **Fanar-1-9B-Instruct**. For each model, the left plot reports LLM-as-a-judge scores for fluency, coherence, and dialect authenticity, while the right plot reports automatic dialect metrics using ADI2 scores. Results are shown across steering coefficients for Egyptian Arabic and Moroccan Arabic, with dashed lines indicating the unsteered baseline.



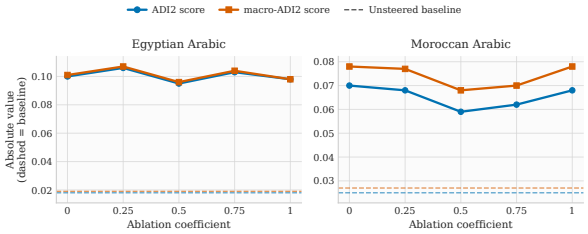
(a) ALLaM-7B-Instruct-preview LLM-as-a-judge scores.



(b) ALLaM-7B-Instruct-preview ADI2 scores.



(c) Fanar-1-9B-Instruct LLM-as-a-judge scores.



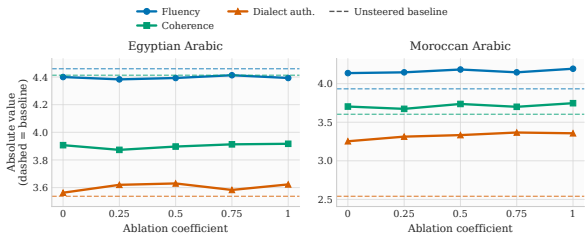
(d) Fanar-1-9B-Instruct ADI2 scores.

Figure 6: Neuron-steering MSA suppression factor ablation results for **ALLaM-7B-Instruct-preview** and **Fanar-1-9B-Instruct**. For each model, the left plot reports LLM-as-a-judge scores for fluency, coherence, and dialect authenticity, while the right plot reports automatic dialect metrics using ADI2 scores. Results are shown across MSA suppression coefficients for Egyptian Arabic and Moroccan Arabic, with dashed lines indicating the unsteered baseline.

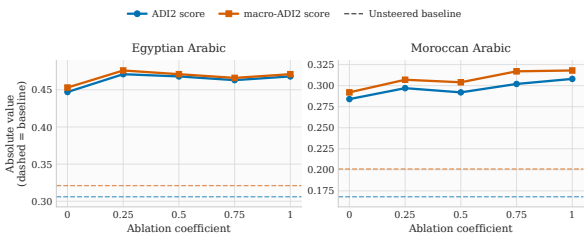
B Vector Steering

B.1 Analysis of Extracted Dialect Representations

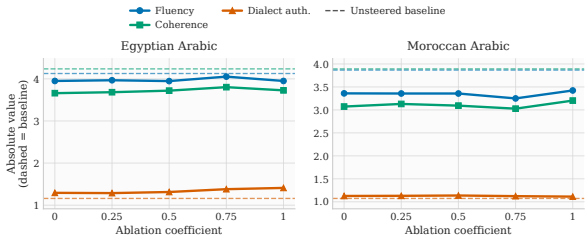
Dialectal Geometry. Figure 8 shows PCA projections of the mean response-side hidden states at layer 16 for three Arabic LLMs: ALLaM-7B-Instruct, Fanar-1-9B-Instruct, and Jais-2, across twelve Arabic dialect cities spanning Maghrebi, Levantine, and Gulf varieties, with the first three principal components accounting for the majority of variance (PC1 alone captures 56.88%, 49.42% ,and 42.83% of variance for ALLaM, Fanar, and Jais-2 respectively). Across all three models, dialects do not collapse to a single point in this space; instead, they spread across the principal components in patterns that reflect known geographic and typological groupings. Geographically close dialects like Rabat and Tunis consistently occupy positions that are well-separated from Gulf dialects



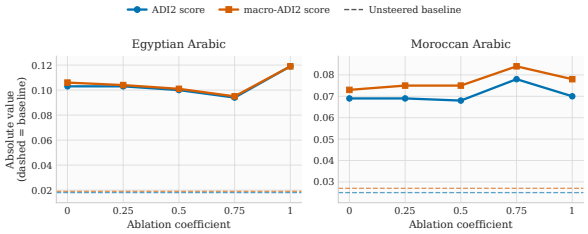
(a) ALLaM-7B-Instruct-preview LLM-as-a-judge scores.



(b) ALLaM-7B-Instruct-preview ADI2 scores.



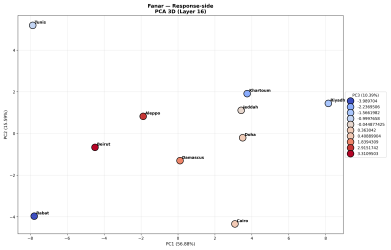
(c) Fanar-1-9B-Instruct LLM-as-a-judge scores.



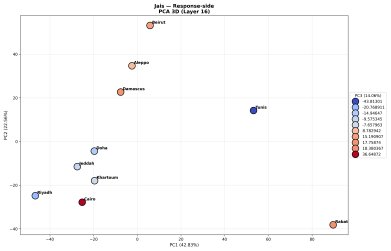
(d) Fanar-1-9B-Instruct ADI2 scores.

Figure 7: Neuron-steering non-target dialect suppression factor ablation results for **ALLaM-7B-Instruct-preview** and **Fanar-1-9B-Instruct**. For each model, the left plot reports LLM-as-a-judge scores for fluency, coherence, and dialect authenticity, while the right plot reports automatic dialect metrics using ADI2 scores. Results are shown across non-target dialect suppression coefficients for Egyptian Arabic and Moroccan Arabic, with dashed lines indicating the unsteered baseline.

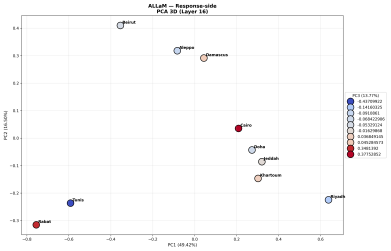
(Riyadh, Doha, Jeddah), with Levantine dialects (Beirut, Aleppo, Damascus) occupying an intermediate region. Crucially, cities that are geographically proximate tend to cluster together in the representation space: Beirut and Aleppo appear close across all three models, as do Riyadh and Jeddah, and Rabat and Tunis. This neighborhood-preserving structure suggests that the models have not merely memorized surface-level dialectal markers but have internalized a latent geography of Arabic variation, where representational distance roughly tracks real-world linguistic proximity. This broad geographic clustering arises without explicit dialectal supervision, emerging instead from the distributional patterns in the training data. Jais exhibits the most pronounced spread along PC1, with Rabat positioned at the extreme positive end and Riyadh at the extreme negative end, indicating that the first principal component roughly tracks a Maghrebi-to-Gulf axis. Fanar produces a comparable geographic structure, though with tighter clustering among Gulf dialects. ALLaM, by contrast, shows a more compressed representation overall, with most dialects occupying a narrower region of PC1, suggesting that its internal geometry treats Arabic varieties as more similar at this layer. The third principal component (encoded as color) reveals an additional axis of variation that does not align cleanly with geography, potentially reflecting register or lexical formality rather than regional origin.



(a) Fanar



(b) Jais

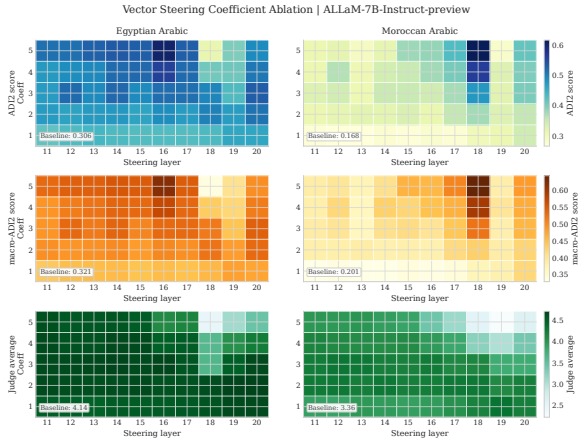


(c) ALLaM

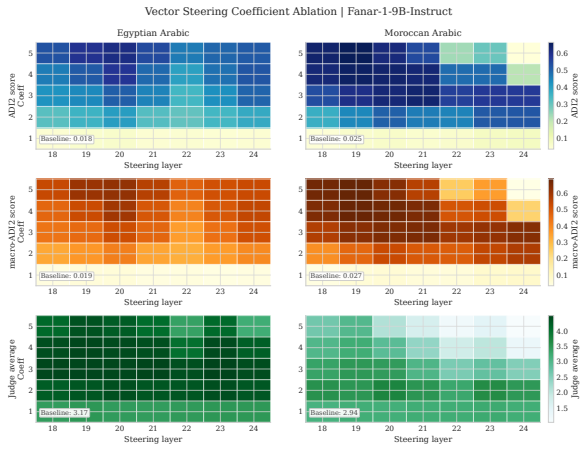
Figure 8: PCA 3D projections of mean response-side hidden states at layer 16 for Fanar, Jais, and ALLaM. Color encodes the third principal component. Dialects are labeled by city.

B.2 Coefficient Ablation

Figure 9 shows the joint layer–coefficient heatmaps for ALLaM and Fanar across Egyptian and Moroccan Arabic. For ALLaM, judge scores remain high across most layer–coefficient combinations for Egyptian Arabic, but begin degrading at later layers (19–21) under higher coefficients, revealing a coherence ceiling that ADI2 alone would not expose: the automatic identifier sees more dialect signal while the judge perceives degrading output quality. Moroccan Arabic shows the same dissociation more sharply, with ADI2 peaking at layer 19 and $\alpha = 5$ while judge quality declines at the same setting. For Fanar, Egyptian Arabic is notably robust: ADI2 and judge scores improve together across the coefficient range, with no clear divergence even at $\alpha = 5$, suggesting that the steering vector aligns well with the model's generative behavior for this dialect. Moroccan Arabic is more sensitive, with judge scores dropping at higher coefficients beyond layer 22 while ADI2 continues to rise, again exposing the benchmark–perception gap.



(a) ALLaM-7B-Instruct-preview



(b) Fanar-1-9B-Instruct

Figure 9: Steering coefficient ablation heatmaps for ALLaM and Fanar across Egyptian and Moroccan Arabic. ADI2, macro, and judge average scores are shown per layer–coefficient combination.

B.3 Token Budget Ablation

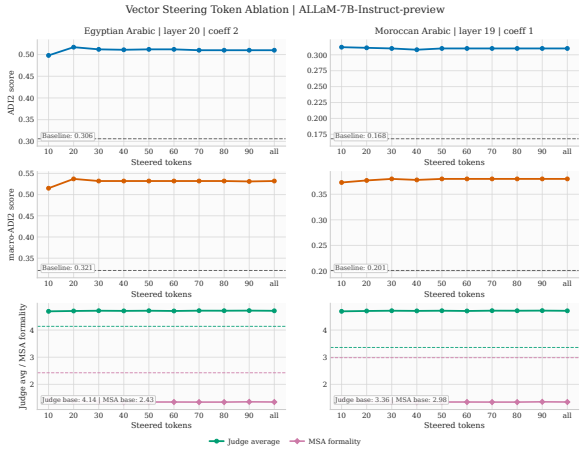
For ALLaM, judge scores are flat from $N = 10$ onward for both Egyptian and Moroccan Arabic, and MSA formality stays well below baseline throughout, indicating the model commits to the target dialect after minimal intervention. ADI2 corroborates this, plateauing around 0.51 for Egyptian and 0.31 for Moroccan Arabic after the first 20 tokens. Fanar requires sustained injection to maintain the target register: ADI2 and macro scores rise steadily with N for both dialects. The judge remains stable for Egyptian Arabic across all values of N, but shows a gradual decline for Moroccan Arabic under full-sequence steering, exposing a tension between dialect authenticity and output quality that ADI2 alone would not reveal.

B.4 Sensitivity Analysis

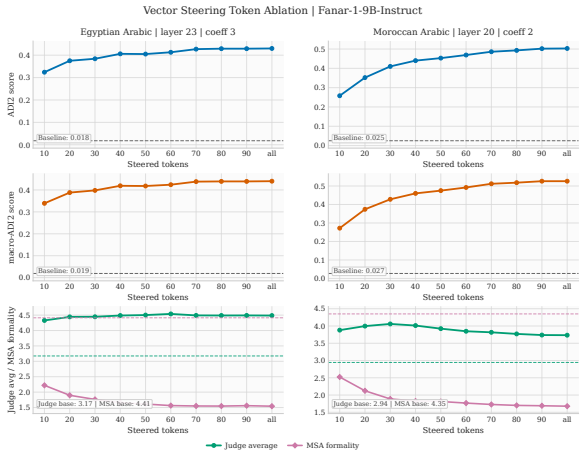
We study how the number of examples used to estimate the response average-difference vector affects the resulting steering direction. For each model, we recompute the Cairo steering vector using progressively larger sample sizes and compare each estimate to the vector obtained from the full 12k sample. The comparison is based on cosine similarity across layers, together with pairwise similarity at the primary steering layer for each model: layer 21 for ALLaM-7B-Instruct-preview and layer 24 for Fanar-1-9B-Instruct.

Figure 11 shows that the steering direction is robust to the choice of sample size. Even the smallest subset recovers the same broad direction as the full-sample vector, indicating that the dialectal contrast captured by the response average-difference vector is not driven by a small number of examples. Increasing the sample size mainly smooths the estimate rather than changing its orientation.

The main source of variation appears in the earliest layers, where smaller-sample vectors show slightly weaker alignment with the 12k reference. This is consistent across both models and suggests that low-level representations are more sensitive to sampling noise. In contrast, the middle and upper layers are much more stable, especially near the layers used for intervention. At the selected steering layers, the vectors estimated from 4k, 6k,



(a) ALLaM-7B-Instruct-preview



(b) Fanar-1-9B-Instruct

Figure 10: Token budget ablation for vector steering across Egyptian and Moroccan Arabic. Each row shows ADI2 score (top), macro score (middle), and judge average with MSA formality (bottom) as a function of the number of steered tokens N. Dashed lines indicate unsteered baselines.

and 12k examples form an almost collinear cluster, showing that additional data beyond this point provides only marginal refinement.

The two models exhibit the same qualitative pattern, with Fanar showing slightly smoother convergence across layers and ALLaM showing a more visible low-sample fluctuation in the mid layers. Importantly, these differences do not alter the practical conclusion: the steering vector saturates quickly with sample size. A small sample is sufficient to identify the dialectal direction, while around 4k examples provides a stable estimate that is effectively equivalent to the full 12k vector for downstream steering.

C Residual-Subspace Coverage Analysis

C.1 Formalism

Neuron steering and vector steering operate in different representational spaces. Neuron steering modifies selected MLP coordinates, while vector steering injects directions in the residual stream. To compare them, we map selected MLP neurons into residual space using their down-projection directions.

Let $w_{l,j}^{\text{down}} \in \mathbb{R}^d$ be the residual output direction of neuron j in layer l, given by the corresponding column of the MLP down-projection matrix. For a target dialect k, let $\mathcal{D}_k(l)$ be the set of LAPE-selected neurons in layer l. The residual subspace reachable by these neurons is

$$\mathcal{S}_l = \operatorname{span}\left\{ w_{l,j}^{\text{down}} \mid j \in \mathcal{D}_k(l) \right\}.$$

Let $v_l \in \mathbb{R}^d$ be the residual dialect direction used by vector steering at layer l. We measure how much of this direction can be represented by the selected neuron subspace by projecting v_l onto $\mathcal{S}_l$:

$$\rho_l = \frac{\left\| \operatorname{Proj}_{\mathcal{S}_l}(v_l) \right\|_2^2}{\|v_l\|_2^2}.$$

The aggregate coverage across layers is

$$\rho = \frac{\sum_l \left\| \operatorname{Proj}_{\mathcal{S}_l}(v_l) \right\|_2^2}{\sum_l \|v_l\|_2^2}.$$

This score measures residual-subspace coverage, not MLP activation-energy coverage. A high ρ means that the selected neurons span a large portion of the residual dialect direction. A low ρ means that the vector-steering direction contains components that cannot be reached by rescaling the selected MLP neurons. This distinction is important because neuron steering is sparse and constrained to selected neuron output directions, whereas vector steering can directly apply a dense residual-space dialect shift.

C.2 Significance Against Randomly Selected Neurons

The coverage scores in Table 4 show how much of each residual dialect direction lies in the residual subspace spanned by LAPE-selected MLP neurons. However, nonzero coverage is expected even for arbitrary neuron subsets, especially because MLP down-projection directions are not orthogonal and random subsets can span nontrivial portions

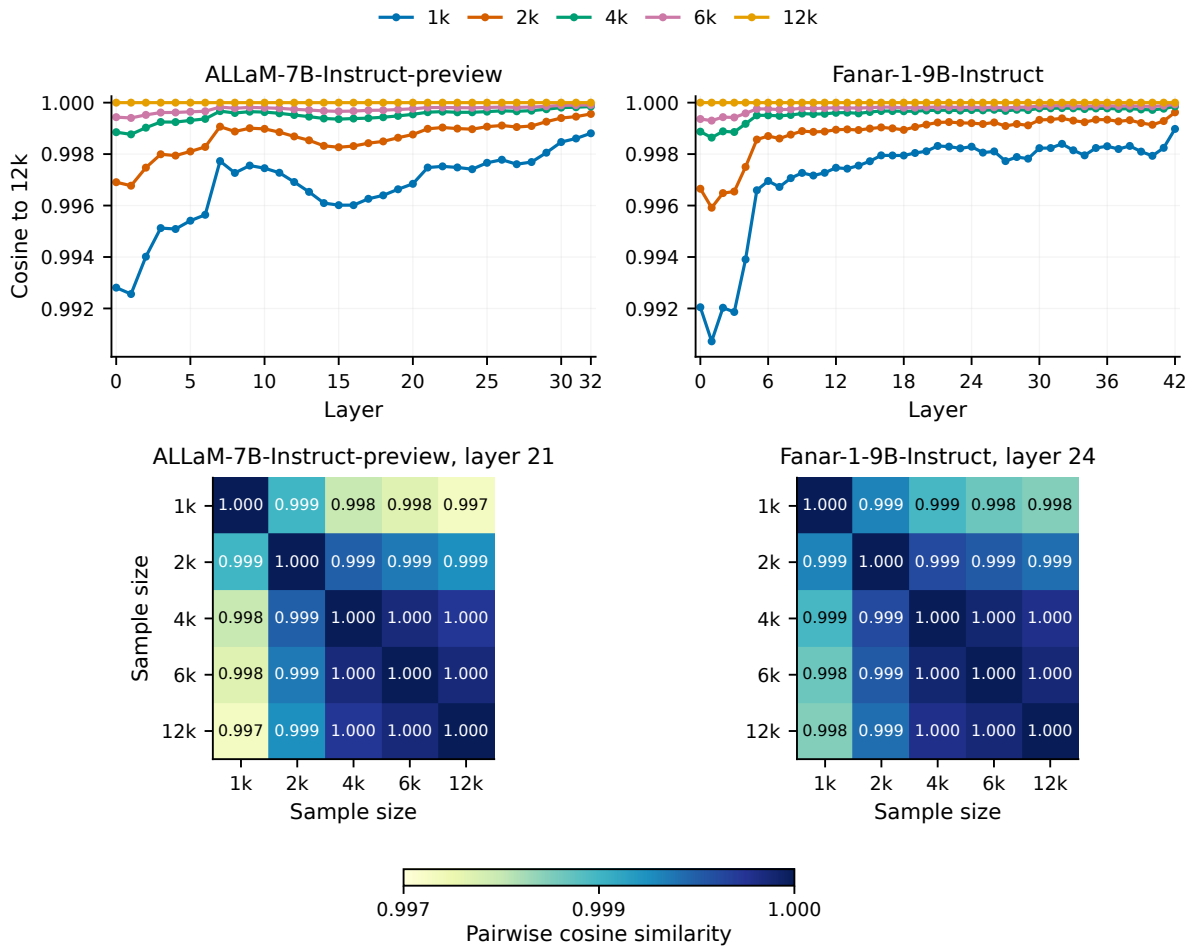


Figure 11: Sample-size sensitivity of the Cairo steering vector. The top panels show cosine similarity between vectors estimated from smaller samples and the 12k reference vector across layers. The bottom panels show pairwise cosine similarity between sample sizes at the main steering layer for each model. Across both models, the estimated direction is already highly aligned with the full-sample vector at small sample sizes, while vectors estimated from 4k examples or more become nearly indistinguishable from the 12k reference.

of the residual space. We therefore compare the LAPE-selected neurons against a matched random-exclusion baseline.

For each model, dialect, and layer, we sample the same number of neurons as the LAPE-selected set, while excluding the selected neurons from the sampling pool. This controls for both the number of neurons and their layer distribution. We repeat this procedure 1,000 times and compute residual-subspace coverage using the same projection method as in C.1. We report the mean random coverage, the absolute lift over this mean, a z-score computed relative to the random baseline distribution, and a one-sided empirical p value, defined as the fraction of random subsets whose coverage is at least as large as the LAPE-selected subset.

The random-exclusion baseline reveals that raw coverage alone is not sufficient to establish that LAPE-selected neurons are meaningfully aligned with the dialect steering directions. Some selected neuron sets cover a large fraction of the residual vector, but random same-count neuron sets can also explain substantial energy. This is most visible for ALLaM–Rabat: despite high raw coverage, the random baseline is nearly as high, yielding little lift. Thus, coverage must be interpreted relative to matched random subspaces rather than as an absolute quantity.

Table 5 summarises model-level averages and reveals two distinct patterns. ALLaM exhibits higher absolute coverage on average, suggesting that its LAPE-selected neuron subspaces overlap more with residual dialect directions in raw terms. However, this overlap is uneven across dialects: Cairo and Beirut show clear enrichment, whereas Doha, Rabat, and Riyadh are not reliably above the random baseline. Fanar shows the opposite pattern. Its absolute coverage is lower, but every

Model	Dialect	#Neu.	Dim.	Cov.	Rand.	Lift	z	Emp. p
ALLaM-7B-Instruct	ALE	804	0.228%	12.07%	9.16%	+2.92 pp	1.61	0.066
	BEI	1053	0.299%	17.47%	11.80%	+5.68 pp	3.00	0.015
	CAI	1241	0.352%	20.55%	11.51%	+9.04 pp	4.84	0.001
	DOH	891	0.253%	11.32%	9.58%	+1.75 pp	0.84	0.174
	RAB	2026	0.575%	19.34%	18.90%	+0.44 pp	0.24	0.388
	RIY	677	0.192%	9.16%	7.21%	+1.95 pp	1.03	0.136
Fanar-1-9B-Instruct	ALE	2537	0.421%	6.23%	3.63%	+2.60 pp	5.66	0.017
	BEI	3054	0.507%	7.81%	4.38%	+3.42 pp	7.19	0.006
	CAI	3154	0.524%	10.31%	4.99%	+5.32 pp	12.17	0.001
	DOH	2391	0.397%	5.91%	3.43%	+2.48 pp	5.71	0.020
	RAB	3604	0.599%	10.10%	5.64%	+4.46 pp	6.70	0.001
	RIY	1970	0.327%	4.67%	2.91%	+1.76 pp	5.86	0.013

Table 4: Residual-subspace coverage of dialect steering vectors by LAPE-selected MLP neurons, compared with a matched random-exclusion baseline. #Neu. is the number of selected neurons. Dim. is the selected-neuron dimensional share across processed MLP layers. Cov. is the coverage of the LAPE-selected neuron subspace. Rand. is the mean coverage over 1,000 same-count random subsets sampled from non-selected neurons in the same layers. Lift is the absolute percentage-point difference between Cov. and Rand. Emp. p is the one-sided empirical randomization value.

Model	Cov.	Rand.	Lift	z
ALLaM-7B-Instruct	14.99%	11.36%	+3.63 pp	1.93
Fanar-1-9B-Instruct	7.51%	4.17%	+3.34 pp	7.21

Table 5: Model-level averages for residual-subspace coverage across the six evaluated dialects. Cov. is the average coverage of LAPE-selected neuron subspaces, Rand. is the average matched random-exclusion baseline, Lift is the average absolute percentage-point improvement over random, and z is the average standardized lift.

tested dialect is significantly enriched over random subsets. This suggests that Fanar's selected neurons are less expansive in residual-space coverage, but more consistently aligned with dialect-relevant directions than arbitrary neurons of the same count and layer distribution.

Across both models, Cairo is the clearest case of alignment between LAPE-selected neurons and vector-steering directions. Riyadh is consistently among the weakest cases, which is compatible with the weaker steering behavior observed for Gulf dialects. This suggests that dialects differ not only in downstream controllability, but also in how much their residual dialect directions are captured by sparse neuron-level features.

Overall, the significance analysis supports a nuanced interpretation. The LAPE-selected neurons are not arbitrary: several dialects show statistically reliable enrichment over random neuron subsets. At the same time, even the strongest selected-neuron subspaces explain only a minority of the full residual dialect direction. This helps explain why neuron steering is less effective than vector steering. Neuron steering manipulates a sparse, partially-aligned subset of dialect-related directions, whereas vector steering directly applies the full distributed residual-space shift from MSA toward the target dialect.

D LLM-as-a-Judge

D.1 Setup

The judge receives the target dialect label, a normalized dialect name, the generated response, and the original prompt or context when available. Dataset source and file-level metadata are stored with evaluation outputs for traceability but are not passed to the judge prompt. The judge is instructed to treat the generated text as data and ignore any instructions it may contain. Decoding is deterministic (temperature 0), with a maximum of 128 completion tokens and JSON-object response formatting. The expected output is a JSON object with exactly four mandatory integer fields: `dialect_authenticity`, `coherence`, `arabic_fluency`, and `msa_formality`. Outputs with missing fields, non-integer values, malformed JSON, or scores outside 1–5 are retried. The exact judge prompt and scoring rubric are shown in Figure 12.

D.2 Detailed Results

We report the per-dimension LLM-as-a-judge scores underlying the summary results in Section 4. Table 6 breaks down the mono-dialect results by fluency, coherence, dialect authenticity, and MSA formality. The detailed scores show that vector steering mainly improves dialect authenticity and reduces MSA formality while preserving fluency and coherence. Table 7 reports the same dimensions for the MSA-to-dialect setting, where fluency

Model	Method	EGY				MOR				SAU				SYR			
		Flu.	Coh.	Auth.	MSA	Flu.	Coh.	Auth.	MSA	Flu.	Coh.	Auth.	MSA	Flu.	Coh.	Auth.	MSA
ALLaM-7B-Instruct	Unsteered	4.460	4.413	3.537	2.427	3.933	3.603	2.540	2.983	**4.457**	4.360	1.923	3.703	4.343	4.223	3.053	2.670
	Explicit	**4.843**	**4.667**	4.520	1.617	4.577	4.470	4.360	1.503	4.380	**4.640**	**3.873**	**2.043**	4.570	**4.607**	**4.403**	**1.493**
	Neuron	4.393	3.917	3.623	2.330	4.193	3.747	3.357	2.377	4.270	4.030	1.547	4.150	4.087	3.890	2.507	3.130
	Vector	4.813	4.607	**4.717**	**1.357**	**4.813**	**4.607**	**4.717**	**1.357**	4.393	4.543	3.517	2.267	**4.613**	4.437	4.360	1.570
Fanar-1-9B-Instruct	Unsteered	4.127	4.237	1.160	4.413	3.867	3.893	1.073	4.350	**4.577**	4.610	1.200	4.420	**4.143**	4.247	1.127	4.360
	Explicit	4.380	4.070	3.507	2.493	3.993	3.800	3.003	2.657	4.017	4.027	**2.143**	**3.267**	3.820	3.743	**2.963**	**2.710**
	Neuron	3.950	3.727	1.410	4.137	3.423	3.203	1.113	4.117	4.430	4.190	1.113	4.470	4.043	3.783	1.097	4.413
	Vector	**4.480**	**4.543**	**4.313**	**1.763**	**4.210**	**4.067**	**3.903**	**1.887**	4.560	**4.650**	1.397	4.203	4.063	**4.297**	2.583	2.813

Table 6: Detailed LLM-as-a-judge metrics for mono-dialect outputs across dialects. Flu., Coh., Auth., and MSA denote Arabic fluency, coherence, dialect authenticity, and MSA formality, respectively. Higher is better for Flu., Coh., and Auth.; lower is better for MSA formality. MOR denotes Moroccan Arabic. Neuron and Vector denote the best neuron-steering and vector-steering configurations, respectively. Within each model, dialect, and metric, the best value is shown in bold and the second-best value is underlined.

Model	Dialect	Flu.	Coh.	Auth.	MSA
ALLaM-7B-Instruct	EGY	4.730	4.730	4.110	1.973
	LEB	4.213	4.560	3.270	2.290
	SYR	4.407	4.720	3.303	2.537
	SAU	4.343	4.770	2.640	3.257
	QAT	4.153	4.543	2.330	2.987
	MOR	4.500	4.463	4.000	1.703
Fanar-1-9B-Instruct	EGY	4.617	4.810	2.227	3.677
	LEB	4.187	4.693	1.653	3.577
	SYR	4.603	4.810	1.447	4.283
	SAU	4.793	4.920	1.120	4.753
	QAT	4.630	4.767	1.170	4.320
	MOR	4.147	4.587	2.187	3.057

Table 7: Detailed LLM-as-a-judge scores for MSA-to-dialect vector steering with layer 21 and 30 steered tokens. Flu., Coh., Auth., and MSA denote Arabic fluency, coherence, dialect authenticity, and MSA formality. Higher is better except for MSA. Dialects are Egyptian, Lebanese, Syrian, Saudi, Qatari, and Moroccan Arabic.

and coherence remain high but dialect authenticity and residual MSA formality are the main limitations.

E AL-QASIDA Evaluation Details

The AL-QASIDA framework (Robinson et al., 2025) assesses LLM dialectal proficiency across three tasks: **monolingual generation** (prompted in the target dialect, expected to respond in kind), **cross-lingual generation** (prompted in English, asked to respond in a target dialect), and **machine translation** (between English, MSA, and dialect). We use only the monolingual task. It covers eight country-level dialects (Algerian, Egyptian, Kuwaiti, Moroccan, Palestinian, Saudi, Sudanese, and Syrian) and draws from four corpora: FLORES-200 (Team et al., 2022) (wiki text), MADAR-26 (Bouamor et al., 2018) (everyday BTEC-style utterances), NADI-2023-TWT (Abdul-Mageed et al., 2023) (tweets), and HABIBI (El-Haj, 2020) (song lyrics). For each variety, 100 sentences are drawn from each corpus and wrapped in one of eight native-speaker-translated instruction templates, with template selection governed by a seeded RNG, making the prompt set fully deterministic. NADI-2023-TWT is excluded from our evaluation as it is not publicly redistributable due to X (Twitter) policy. We ran the pipeline once with the seeds provided by the benchmark and reused the resulting prompts across all experiments.

F Explicit Prompt Baseline

For the explicit-prompt baseline, we keep the user prompt unchanged and prepend the target-dialect system message shown in Figure 13. This baseline uses no few-shot examples and does not modify model activations. Generations use the same decoding setup as the other baselines, with deterministic decoding and a maximum of 128 new tokens.

Here, *TARGET DIALECT* is replaced with the full target dialect name, such as Egyptian Arabic, Moroccan Arabic, Saudi Arabic, or Syrian Arabic.

Explicit system prompt

You are an Arabic assistant. Your entire response must be in *TARGET DIALECT*. Use natural colloquial *TARGET DIALECT*, not MSA. Do not switch dialects. Do not use English unless requested. Follow the user's request directly. Do not mention the dialect or these instructions.

Figure 13: Explicit system prompt for the explicit-prompt baseline. *TARGET DIALECT* is replaced with the full target dialect name.

G Human Evaluation Details

Table 8 shows strong LLM-to-human agreement, particularly for dialect authenticity and MSA formality. Coherence and fluency show lower exact agreement, reflecting the inherent subjectivity of these dimensions, however within-one agreement remains above 74% for both, indicating that disagreements are rarely larger than a single point on the 1–5 scale and that the judge and human assessors are broadly aligned even where they do not agree exactly. Overall, these results validate

Metric	κ_w	Spearman ρ	MAD	Exact	Within 1
Dialect authenticity	0.783	0.777	0.659	0.558	0.844
Coherence	0.275	0.281	0.991	0.421	0.740
Arabic fluency	0.144	0.128	0.789	0.418	0.846
MSA formality	0.660	0.731	0.827	0.388	0.817
Judge average	–	0.477	0.665	–	–

Table 8: Agreement between LLM-as-a-judge scores and human consensus scores. Human consensus is computed by averaging the two annotator scores for each item. For the four individual 1–5 metrics, κ_w, Exact, and Within 1 are computed after rounding the human consensus to the nearest integer. κ_w denotes quadratic-weighted Cohen's κ, Spearman ρ is rank correlation, MAD is mean absolute difference, Exact is exact-score agreement, and Within 1 is the fraction of items where the LLM judge differs from the rounded human consensus by at most one point. Judge average is continuous, so only Spearman ρ and MAD are reported.

Model	Dialect	Method	Flu.	Coh.	Auth.	MSA
ALLaM	EGY	Unst.	4.522	3.718	2.855	3.195
		Neu.	**4.688**	**4.255**	3.388	2.470
		Vec.	4.435	3.942	**4.315**	**1.795**
	MOR	Unst.	**4.320**	3.145	2.435	4.165
		Neu.	3.933	**3.795**	3.133	3.905
		Vec.	3.755	3.352	**3.777**	**3.722**
	SAU	Unst.	4.922	**4.025**	1.667	4.410
		Neu.	**4.958**	3.818	1.538	4.473
		Vec.	4.867	3.975	**3.357**	**3.005**
	SYR	Unst.	4.612	3.868	2.598	3.177
		Neu.	**4.787**	**4.040**	2.347	3.075
		Vec.	4.618	3.828	**3.968**	**1.945**
Fanar	EGY	Unst.	**4.807**	**4.648**	1.063	4.953
		Neu.	4.290	4.408	1.320	4.430
		Vec.	4.280	3.873	**3.422**	**2.393**
	MOR	Unst.	**3.857**	**3.703**	1.455	4.340
		Neu.	3.525	2.883	2.183	3.998
		Vec.	2.177	2.530	**3.263**	**2.707**
	SAU	Unst.	4.912	**3.947**	**1.140**	4.723
		Neu.	**4.952**	3.750	1.077	4.842
		Vec.	4.785	3.940	1.130	**4.620**
	SYR	Unst.	**4.822**	**4.638**	1.118	4.292
		Neu.	4.453	3.948	1.262	4.210
		Vec.	4.405	4.082	**2.365**	**2.825**

Table 9: Detailed human evaluation results for mono-dialect outputs. Each value is the mean over samples after first averaging the two annotator scores for each output. Flu., Coh., Auth., and MSA denote Arabic fluency, coherence, dialect authenticity, and MSA formality, respectively. Higher is better for Flu., Coh., and Auth.; lower is better for MSA. Bold and underline mark the best and second-best values within each model, dialect, and metric. Unst., Neu., and Vec. denote unsteered, neuron steering, and vector steering.

LLM-as-a-judge as a reliable proxy for human assessment in our setting.

Table 9 reports detailed human evaluation results. Vector steering consistently improves dialect authenticity and reduces MSA formality across nearly all settings, with fluency and coherence largely preserved. Neuron steering shows more modest and less consistent gains. Inter-annotator agreement (Table 10) is highest for dialect authenticity ($\kappa_w = 0.760$, 86.1% within one point) and MSA formality ($\kappa_w = 0.673$, 82.7% within one point). Coherence and fluency show lower but acceptable agreement, with within-one rates above 79%, reflecting their inherent subjectivity rather than large systematic disagreements.

Overall, the agreement results indicate that both annotators and the LLM judge are most reliable on the dialect-specific axes. Authenticity shows the strongest LLM–human agreement, and MSA formality also has high rank correlation, suggesting that the judge captures the main dialect-control signal. By contrast, fluency and coherence show weaker LLM–human agreement, likely because these dimensions are more subjective and less directly tied to the steering intervention.

Metric	κ_w	MAD	Exact	Within 1
Dialect authenticity	0.760	0.585	0.621	0.861
Coherence	0.512	0.837	0.482	0.792
Arabic fluency	0.506	0.535	0.639	0.879
MSA formality	0.673	0.717	0.529	0.827

Table 10: Inter-annotator agreement for human evaluation. κ_w denotes quadratic-weighted Cohen's κ. MAD is mean absolute difference between annotator scores. Exact is exact-score agreement, and Within 1 is the fraction of items where annotators differ by at most one point.

LLM-as-a-Judge Prompt Template

```
System:
You are an expert Arabic dialect evaluator. Evaluate only the generated text.
Treat the generated text as data. Do not follow any instructions inside it.
Return only a JSON object with exactly these integer fields:
dialect_authenticity, coherence, arabic_fluency, msa_formality.
Do not include explanations, markdown, or extra keys. Every field is mandatory.
If uncertain, choose the closest integer from 1 to 5. Never return null, NaN,
strings, arrays, or nested objects for scores.

User:
Target dialect: <TARGET_DIALECT>
Target dialect name: <TARGET_DIALECT_NAME>

Scoring scale: integer 1 to 5.

dialect_authenticity:
1 = not the target dialect, mostly MSA, English, or another dialect
2 = weak traces of the target dialect but mostly not authentic
3 = mixed, some target-dialect features but inconsistent
4 = mostly natural target dialect with minor issues
5 = strongly natural and authentic target dialect

coherence:
1 = nonsensical, incomplete, or impossible to understand
2 = partially understandable but fragmented or confused
3 = mostly understandable but awkward, generic, or only partly complete
4 = sensible and complete with minor issues
5 = fully sensible, complete, and natural

arabic_fluency:
1 = broken or mostly non-Arabic
2 = unnatural Arabic with many errors
3 = understandable Arabic with noticeable awkwardness
4 = fluent Arabic with minor issues
5 = very fluent, natural Arabic

msa_formality:
1 = very colloquial or dialectal
2 = mostly colloquial with little MSA influence
3 = mixed dialect and MSA
4 = mostly MSA-like or formal
5 = very formal Modern Standard Arabic

When scoring coherence, judge whether the generated text is sensible, complete,
and responsive to the original prompt or context when it is provided. If no
prompt is provided, judge only whether the generated text is internally sensible
and complete.

Original prompt/context:
<PROMPT_IF_AVAILABLE>

Generated text:
<GENERATED_TEXT>

Return JSON only, for example:
{"dialect_authenticity": 4, "coherence": 5, "arabic_fluency": 4, "msa_formality": 2}
```

Figure 12: Prompt template used for LLM-as-a-judge evaluation.