WHEN ALIGNMENT HURTS: DECOUPLING REPRESENTATIONAL SPACES IN MULTILINGUAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

It is often assumed that aligning low-resource varieties with high-resource standards improves modeling in multilingual Large Language Models (LLMs). We challenge this view with the first causal study showing that excessive representational entanglement with dominant varieties can reduce generative quality. We introduce an online variational probing method that continuously estimates the subspace of a dominant variety during fine-tuning on a generative task and penalizes it to reduce its span. Across six language families we find that reducing alignment consistently boosts low-resource translation performance, including +11.7 ChrF++ for European Portuguese, +5.3 for Indonesian, +4.6 for Kven Finnish, and +2.7 for Low German. In Arabic, several dialects improve by up to +4.3 ChrF++ despite sharp drops for cross-lingual tasks such as translation to MSA, English, or French, suggesting that the effect extends beyond simple cross-lingual alignment. Alongside these causal results, we present qualitative and observational evidence from information-theoretic and geometric probing that further supports our hypothesis. Together, our findings establish that disentangling high-resource subspaces can unlock capacity for related low-resource varieties and provide practical tools for controlling representational allocation in multilingual LLMs. Code will be released.

1 INTRODUCTION

Large Language Models (LLMs) have achieved remarkable progress in multilingual Natural Language Understanding (NLU) and Generation (NLG) tasks (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2022; Aryabumi et al., 2024). Beyond English, these models show strong cross-lingual transfer, enabling low-resource varieties to benefit from related high-resource languages (Hu et al., 2020; Conneau et al., 2020; Xue et al., 2021).

A less understood question, however, is whether closer alignment with a dominant, high-resource variety always benefits related low-resource ones. Dialects provide a natural test case: they are linguistically distinct, socially important, yet often heavily entangled with their standardized counterpart in both data and models. Arabic exemplifies this dynamic, where Modern Standard Arabic (MSA) dominates pretraining resources while dozens of dialects remain underrepresented and underperform on benchmarks (Kantharuban et al., 2023). Similar dynamics arise in other orthographically and lexically close pairs such as Czech–Slovak, Indonesian-Malay, Standard-Low German, Brazilian-European Portuguese, and Kven-Finnish. Understanding representational interactions in such settings is crucial for inclusive generative modeling.

This paper challenges the assumption that alignment with a high-resource standard is always beneficial. By studying six diverse linguistic groups, we show that excessive representational entanglement with the higher-resource variety may hinder generative performance. Since parallel and labeled corpora for other generative tasks across dialects/similar languages are scarce, we focus on machine translation as a controlled proxy for dialect-sensitive generation.

Our study proceeds in two stages. First, we introduce a *novel* **online variational probing** framework that continuously estimates the subspace of the high-resource standard during fine-tuning on a generative task like machine translation, enabling a novel subspace decoupling strategy. This causal intervention promotes orthogonal representations and improves generative capacity for lower-resource varieties. Then, we shift to a more qualitative/observational analysis honing in on Arabic to analyze

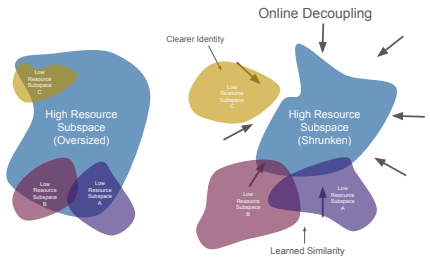

Figure 1: Visualization of the intuition behind our method.

how LLMs internally represent Modern Standard Arabic (MSA) and dialects, revealing that stronger generative performance correlates with greater representational separability from MSA.

Applied to 6 diverse language groups, our approach yields consistent improvements over standard fine-tuning, consistently boosting lower-resource performance, including **+11.7** ChrF++ for European Portuguese, **+5.3** for Indonesian, **+4.6** for Kven Finnish, and **+2.7** for Low German. In Arabic, several dialects improve by up to **+4.3** ChrF++ despite drops in cross-lingual tasks such as translation to MSA, English, or French, indicating the presence of factors that go beyond simple cross-lingual alignment. More broadly, our findings provide the first causal evidence that representational dominance by high-resource standards can limit generative modeling in closely related varieties.

Contributions.

- We introduce and verify a novel online probing-based subspace decoupling method that improves generative performance for underrepresented varieties.
- For the first time, we demonstrate that despite helping with cross-lingual performance alignment has a detrimental effect on dialectal/similar-language performance.
- We empirically demonstrate consistent gains across 6 language groups, highlighting implications for related language families where orthographic and lexical similarity creates similar entanglement.
- We present the first large-scale representational analysis of dialects in generative LLMs, unifying geometric and information-theoretic probing.

2 Related Works

This work investigates how LLMs internally allocate representational capacity across closely related language varieties.

Multilingualism in Large Language Models. Recent studies have analyzed how multilingual LLMs encode language-specific knowledge. For example, Wang et al. (2024) and Kojima et al. (2024) explore neuron sharing and language-specific activations, showing that subtle modifications can alter generation in particular languages. Our perspective differs: rather than focusing on neuron-level behavior, we ask whether dialects remain representationally distinct from their standardized counterpart and how this distinction (or entanglement) affects generative performance. This question is not limited to one particular group or family, but applies broadly to orthographically and lexically similar pairs with a resource imabalance.

At the representational level, Chang et al. (2022) show that languages occupy distinct subspaces in encoder-only models, while Shah et al. (2024) link geometric differences to cross-lingual transfer. We extend these insights to large generative models, showing that the degree of subspace separability between varieties correlates with downstream generation quality. Similarly, Nigatu et al. (2023) find that models struggle to capture dialectal nuances; our results both confirm this for recent LLMs and provide causal evidence that mitigating representational entanglement improves performance.

Information-Theoretic Probing. Information-theoretic probes have been used to study how linguistic signals emerge during pretraining (Voita & Titov, 2020; Müller-Eberstein et al., 2023). Build-

ing on this, we introduce probes not just for analysis but as part of training: our "variety probes" continuously estimate the dominant higher-resource variety subspace during fine-tuning, enabling us to directly intervene by penalizing entanglement. This extends probing from a diagnostic tool to a mechanism for causal representational control.

Dialectal and Low-Resource NLP. Dialectal variation presents a persistent challenge for generative modeling. Prior work has documented large performance gaps as dialects deviate from their standardized counterpart (Kantharuban et al., 2023; Ziems et al., 2023). For Arabic, evaluation resources such as AraBench (Sajjad et al., 2020) and MADAR (Bouamor et al., 2018) have been developed, and recent studies examine MT and NLG across varieties (Kadaoui et al., 2023; Nagoudi et al., 2023). Efforts like the Tatoeba challenge (Tiedemann, 2020) and FRMT (Riley et al., 2023) provide region-aware and tail-end language few-shot machine translation resources. Our work departs from these by focusing not on resource creation or evaluation but on how varieties are internally represented in LLMs and how interventions on representational subspaces can improve generative capacity. While Arabic provides a uniquely rich testbed given its extensive dialectal spectrum, the implications extend to other under-resourced language varieties that share high orthographic and lexical overlap with a dominant variety.

3 Background: Dialects and Similar Language Varieties

Languages vary internally due to cultural, environmental, geographical, and administrative factors (Honkola et al., 2018). These variations often diverge into distinct varieties, with speakers of minority varieties facing socioeconomic disadvantages that are mirrored in multilingual LLMs (Kantharuban et al., 2023). While LLMs leverage scraped data and cross-lingual transfer, such benefits are less evident for lower-resource varieties closely related to higher-resource ones than for more distinct low-resource languages. We address this gap by moving beyond alignment-based solutions and investigating representational dominance in LLMs as a key driver of disparities. The distinction between "dialects" and "languages" is scientifically and politically problematic, often yielding artificial boundaries (Melinger, 2018). We therefore use the neutral term **variety** to refer to any spoken or written linguistic form, and group varieties based on demonstrated lexical and orthographic similarity. An illustration for Arabic varieties is shown in Table 1.

Table 1: Sample of 5-way parallel sentences meaning "How much does the breakfast cost?" in 5 different varieties of Arabic from the MADAR 26 corpus (Bouamor et al., 2018). The yellow highlights the interrogative element (roughly "how much"), the green (when present) highlights the explicit cost word, and the blue highlights the breakfast term.

Dialect	Arabic	Transliteration (Buckwalter)
Modern Standard Arabic	كم تكلفة الإفطار؟	kam taklifaT al-'ifTar?
Egyptian Arabic	بكام الفطار؟	bkam al-fiTar?
Levantine Arabic	أدي حق الترويقة؟	'addi Haq al-tarwiqa?
Gulf Arabic	بكم الريوق؟	bkam al-riyooq?
Maghrebi Arabic	بقداش فطور الصباح؟	bqaddash fuToor al-SabaaH?

4 Methodology

We present a methodology designed to first diagnose and then causally intervene in the representational geometry of multilingual models. Our approach uses a controlled generative task to probe model capabilities, analyze the underlying representations through geometric and information-theoretic lenses, and introduce a novel training technique to mitigate representational entanglement. We clarify our Large Language Model use for this paper in Appendix H.

4.1 Task Formulation: Machine Translation as a Generative Testbed

To study varietal generation in a controlled setting, we formulate the task of **Inter-variety Machine Translation** (VarMT). Given a sentence in a higher-resource variety, the model must generate the semantically equivalent sentence in a lower-resource variety. This setup serves as a proxy for broader conditional generation, enabling precise measurement of a model's ability to manipulate linguistic style while preserving meaning. We adopt MT as our testbed due to the relative availability of parallel data, in contrast to other generative tasks (e.g., summarization, open-ended dialogue). Prompting details are provided in Appendix A. For our causal experiments (Sec. 4.3), we fine-tune models with a bidirectional VarMT objective (higher-resource $\leftrightarrow$ lower-resource). This prevents models from trivially degrading higher-resource representations in favor of lower-resource performance, ensuring a fairer evaluation of subspace dynamics and intervention effects. The only exceptions are Indonesian–Malay and Czech–Slovak. Since these groups are typically considered distinct languages, we instead train with English as a pivot (English $\rightarrow$ language), providing complementary evidence to the VarMT setup.

4.2 Quantifying Performance and Representational Geometry

Evaluation. We evaluate generation quality using chrF++ (Popović, 2015), a character n-gram F-score. Its character-level nature makes it well-suited for morphologically rich languages and robust to the minor lexical variations common across varieties, while remaining sensitive to subtle character-level shifts. Like other automated metrics, chrF++ cannot fully capture the nuances of "varietal distinctness." Human annotation would be the only true alternative, but is largely infeasible given the small native-speaker populations of many varieties. LLM-based metrics are also unsuitable, as they risk reintroducing the very biases that hinder variety-sensitive generation.

4.3 Causal Intervention: Online Subspace Decoupling

To test the hypothesis that representational entanglement with a high-resource varieties harms low-resource generation, we introduce a novel training method: **Online Subspace Decoupling**. This method acts as a causal intervention by actively discouraging lower-resource varietal representations from overlapping with the higher-resource variety subspace during fine-tuning.

The procedure is as follows:

1. **Identify Higher-resource Variety Subspace:** We train a variational linear probe (as in Sec. 4.4) to distinguish the higher-resource variety from all other varieties in a group. We then use Singular Value Decomposition (SVD) on the learned probe weights to extract an orthonormal basis $\mathbf{U}_{\text{HR}}$ for the higher-resource subspace and form its projection matrix: $\mathbf{P}_{\text{HR}} = \mathbf{U}_{\text{HR}}\mathbf{U}_{\text{HR}}^{\top}$.
2. **Define Decoupling Loss:** During fine-tuning on the VarMT task, we add a penalty term to the standard language modeling loss. This *decoupling loss* penalizes the magnitude of the projection of the model's hidden states $\mathbf{H}$ onto the higher-resource subspace:

$$\mathcal{L}_{\text{decouple}} = \mathbb{E}\left[\|\mathbf{H}\mathbf{P}_{\text{HR}}\|_2\right] \tag{1}$$

   The total loss is $\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda\,\mathcal{L}_{\text{decouple}}$, where λ is a hyperparameter (we use λ =1e-4 across all setups after ablation, check Appendix E.2.).

Crucially, the probe is periodically retrained on fresh model checkpoints during fine-tuning. This **online updating** of $\mathbf{P}_{\text{HR}}$ ensures that our intervention targets the evolving higher-resource subspace before the probe becomes too stale, enabling a precise and adaptive causal manipulation of the model's representational geometry. Further details are in Appendix E.

Representational Geometry. To understand *how* models represent varieties, we analyze their internal geometry. For the observational part of our study we hone in on Arabic dialects due to the unique availability of 28-way parallel resources (MADAR 26 (Bouamor et al., 2018)). Furthermore, Arabic provides a plethora of different varieties each with their own unique characteristics which can be compared to the standard. We measure the **Geometric Separability** between sentence representations using L2 and cosine distance, anchoring all comparisons to Modern Standard Arabic

(MSA) representations. This allows us to quantify how distinct dialectal representations are from the high-resource standard. Furthermore, we compute **Subspace Angles (SSA)** (Müller-Eberstein et al., 2023) to measure the alignment between subspaces corresponding to different dialects. Smaller angles indicate greater alignment. This allows us to track how fine-tuning and our proposed interventions reshape the model's internal organization of linguistic information.

4.4 Information-Theoretic Probing

To complement the geometric analysis, we employ an information-theoretic variational linear probe (similar to our online subspace decoupling intervention) (Voita & Titov, 2020; Müller-Eberstein et al., 2023). The probe is a sparsity-regularized classifier trained to identify a variety from token-level representations. The resulting negative cross-entropy provides a tight lower bound on the mutual information $I(\mathbf{h}^{(\ell)}; Y)$ between a model's hidden states and the variety's identity. This allows us to quantify how easily variety-specific information can be linearly decoded from the model's representations, layer by layer, and how this changes during training. Further details are in Appendix D. Again similarly to the geometric analysis, when this tool is used observationally (i.e. not in our causal intervention) we hone in only on Arabic for practical considerations.

4.5 Experimental Setup

Data. We cover six groups of varieties. To be able to cover this range, we utilize data resources from four dataset resources. For Arabic we use the MADAR 26 corpus (Bouamor et al., 2018), which contains 2,000 parallel sentences across 25 city-level Arabic dialects, MSA, English, and French. This fine-grained, multi-dialect parallel resource is unique and enables controlled observational study. For Brazilian-European Portuguese, we use the FRMT resource (Riley et al., 2023). For Indonesian-Malay and Czech-Slovak we use the Flores-200 dataset (Costa-jussà et al., 2022; Goyal et al., 2021). For Standard-Low German and Kven-Finnish we use the Tatoeba challenge (Tiedemann, 2020). We elaborate on the precide processing and splits of each dataset in Appendix B

Models. We analyze a suite of state-of-the-art open-weight multilingual models: Jais-family 30B (Sengupta et al., 2023), Gemma 3 1B (Team, 2025a), Aya expanse 8B (Dang et al., 2024), and Qwen 3 14B (Team, 2025b). For our causal intervention experiments, we deliberately select Gemma 3 1B. For finetuning we start with the base (non-instruction tuned model). Its smaller parameter count implies a more constrained representational space, making it a challenging and informative test case for the benefits of explicit subspace management. Furthermore, its weaker baseline performance provides a clear opportunity to measure improvement from our method.

5 Results and Analysis

We now present our empirical investigation, which first validates our hypothesis with a causal intervention on multiple language groups then explores the representational pathologies hindering dialectal generation in multilingual models by focusing on Arabic. We place the numerical results for all setups in Appendix F.

5.1 Causal Validation: Online Subspace Decoupling Boosts Performance

We test our hypothesis of excessive representational dominance/-conflation negatively affecting lower-resource generative abilities of multilingual LLMs on lower-resource varieties by directly using our proposed **Online Subspace Decoupling** method (Section 4.3). By adding an explicit penalty term that penalizes over-sized higher-resource variety subspaces. The choices for which variety's subspace to penalize per group are outlined in Table 2. We make these choices based on previously reported performance disparities (Kantharuban et al., 2023) and if they are considered the standard (if applicable).

Table 2: Higher-resource Varieties

Language Group	Higher-Resource Variety
Portuguese	Brazilian Portuguese
Czech/Slovak	Czech
Finnish/Kven	Finnish
German	Standard German
Malay/Indonesian	Indonesian
Arabic	Modern Standard Arabic

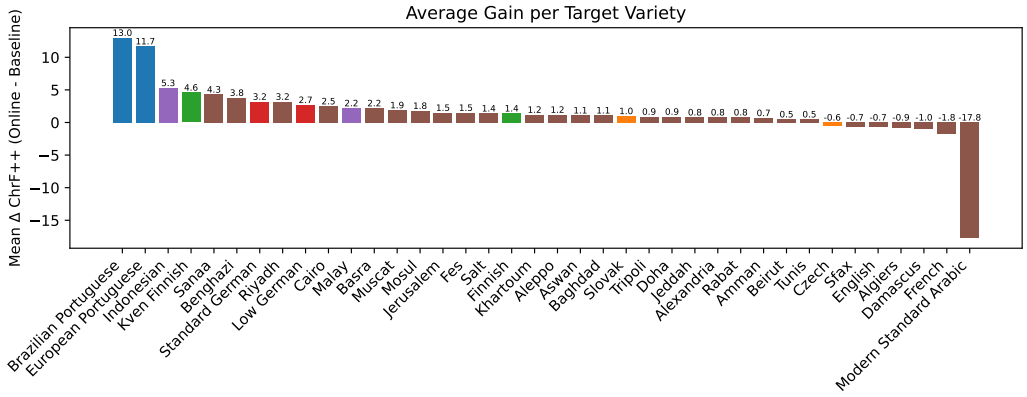

Figure 2: The mean delta in ChrF++ on several target varieties and languages between our Online Decoupling Training and Baseline SFT on VarMT. A positive delta indicates superior performance from our method.

In Figure 2, we compare online subspace decoupling against baseline supervised fine-tuning. Improvements are most consistent for lower-resource target varieties, where inflated subspaces of high-resource counterparts are explicitly penalized and disentangled. European Portuguese is particularly illustrative: despite Brazilian Portuguese dominating corpus size and representational allocation, decoupling yields a striking +11.7 ChrF++ improvement, showing that naive fine-tuning can in fact be hindered by conflation with a related high-resource variety. Smaller but still meaningful gains are observed for Kven (+4.6 ChrF++), Low German (+2.7), and Slovak (+1.1). Importantly, the dominant higher-resource varieties change little under decoupling (with the exception of Czech, which shows a minor decline, and MSA, which worsens). This supports the claim that our method reallocates representational capacity toward underrepresented varieties rather than amplifying already dominant ones. A one-sided Wilcoxon signed-rank test on overall ChrF++ scores confirms that online decoupling significantly outperforms baseline fine-tuning across variety setups (excluding variety→MSA translation), yielding a p-value of 0.00195. Online decoupling achieves higher ChrF++ in 9 of 10 setups, with an average gain of +4.45 points. At the sentence-level, one-sided Wilcoxon tests further show that online decoupling significantly improves over the baseline ($p < 0.05$) for nearly all varieties, with exceptions in Finnish, Czech, and many Arabic dialects (notably Amman, Cairo, Muscat, and Salt do reach significance) (see Appendix G). Together, these results demonstrate that decoupling systematically benefits lower-resource varieties but may not harm their higher-resource counterparts on many of the language groups observed.

Table 3: ChrF++ of Random Subspace Decoupling vs. Baseline SFT.

Target	Baseline	Random	Δ Random-Baseline
Sanaa	18.4	17.9	-0.5
Benghazi	21.0	18.3	-2.7
Riyadh	23.0	19.9	-3.1
Cairo	16.9	16.1	-0.8
Basra	20.1	17.5	-2.6
Muscat	18.6	18.0	-0.6
Mosul	20.9	17.7	-3.2
Fes	22.5	21.2	-1.3
Jerusalem	22.6	18.5	-4.1
Salt	22.0	17.9	-4.1
Aleppo	20.6	17.2	-3.4
Khartoum	22.1	18.2	-3.9
Baghdad	20.4	15.7	-4.7
Aswan	20.2	18.8	-1.4
Tripoli	21.2	18.8	-2.4
Doha	21.8	17.7	-4.1
Rabat	19.9	19.0	-0.9
Alexandria	21.9	19.1	-2.8
Jeddah	21.2	18.0	-3.2
Amman	20.4	16.7	-3.7
Beirut	18.5	16.7	-1.8
Tunis	17.9	16.0	-1.9
Sfax	18.2	18.1	-0.1
Algiers	23.6	20.1	-3.5
Damascus	20.6	17.1	-3.5
French	27.7	22.0	-5.7
English	31.4	22.9	-8.5

Arabic dialects provide further support for our hypothesis. Constraining Modern Standard Arabic (MSA) subspaces yields gains of up to +4.3 ChrF++ (Sanaa) for many dialects, even as MSA itself and cross-lingual transfers (e.g., to English or French) decrease. This asymmetry shows that dominance by the standard variety does not linearly benefit dialect modeling and can suppress dialectal

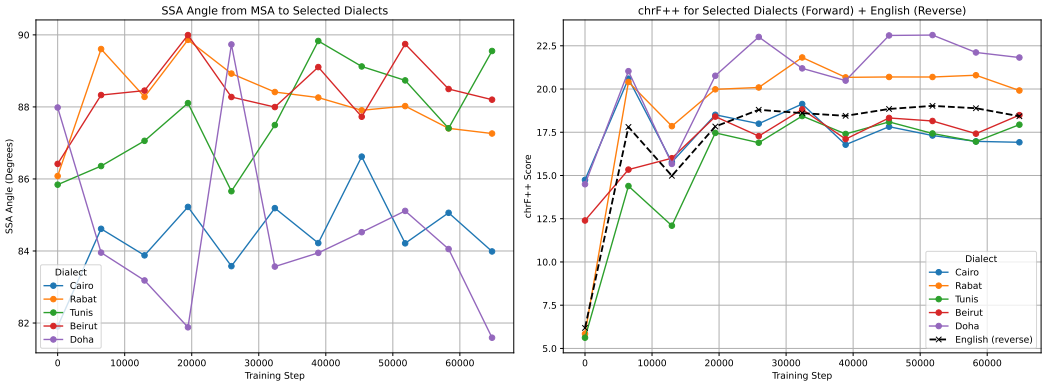

Figure 3: **(Left)** During baseline SFT, the subspace angle (SSA) between MSA and dialects consistently increases, indicating growing representational separation. **(Right)** This increase in separation correlates directly with improved chrF++ scores. This provides strong evidence that disentangling from MSA is a key mechanism for improving dialectal generation.

expressivity. Online Subspace Decoupling effectively reallocates capacity to underrepresented dialects, unlocking performance that would otherwise be constrained by MSA.

Interestingly, some higher-resource varieties also improve under decoupling: Indonesian gains +5.3 ChrF++ and Brazilian Portuguese +13.0, the largest observed increase. This indicates that entangled subspaces can distort both lower- and higher-resource varieties. By disentangling, decoupling could be sharpening the boundaries between varieties, reducing interference and enabling specialization. Penalizing oversized subspaces may also prevent dominant varieties from overfitting shared structures, benefiting both high- and low-resource generation.

To rule out gains from generic hidden space regularization, we tested random subspace shrinking on Arabic dialects. As shown in Table 3, performance consistently dropped below baseline for MSA, the dialects, French, and English, confirming that improvements arise specifically from disentangling oversized higher-resource subspaces rather than from indiscriminate regularization.

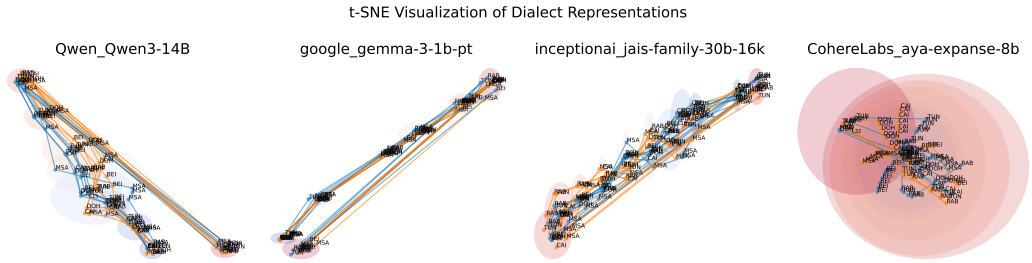

Figure 4: t-SNE of sentence representations. Higher-performing models (e.g., Qwen, Aya) exhibit clearer separation between dialectal clusters in their intermediate layers, unlike weaker models (Gemma, Jais).

5.2 Exploration I: Geometric Analysis Links Performance to Representational Separation

Now we move on from causal experimentation to quantitative and qualitative observation of model geometry across multiple models of varying sizes and families on Arabic. To investigate the underlying representational geometry, we visualize the hidden states of parallel sentences from MADAR 25 using t-SNE (with perplexity=20) (Figure 4). The visualizations reveal a striking pattern: stronger models like Qwen and Aya learn to separate representations by dialect in their intermediate layers, whereas weaker models like Jais and Gemma maintain entangled representations. This qualitative

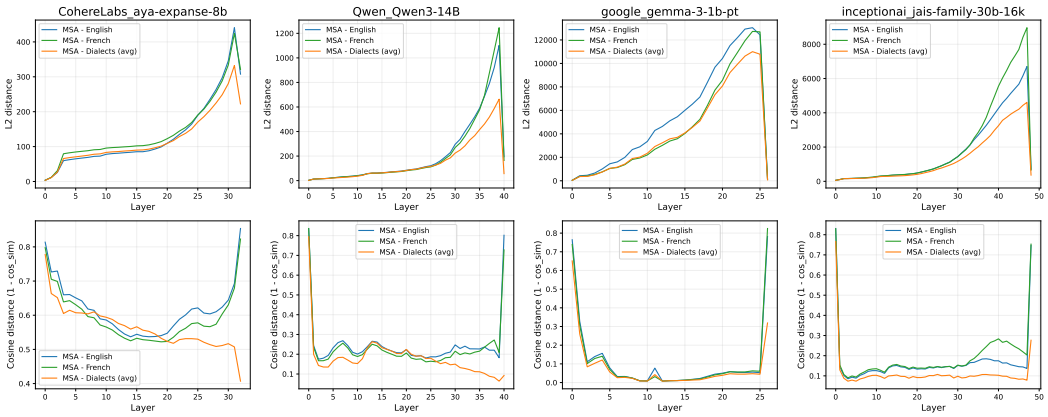

Figure 5: Layer-wise L2 (Top) and Cosine (Bottom) distance between dialectal representations and MSA. High-performing models show distinct geometric patterns, with Aya treating dialects more like separate languages (high L2 distance).

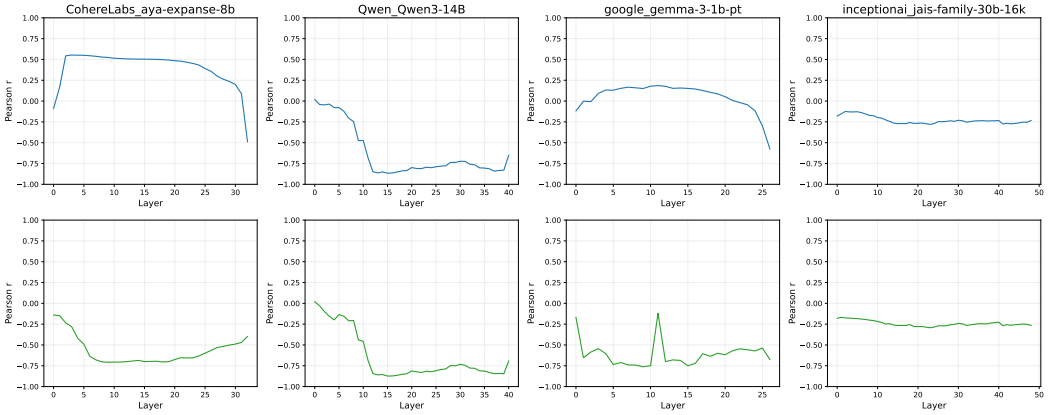

Figure 6: Layer-wise Pearson correlation between representational distance from MSA (L2-Top, Cosine-Bottom) and downstream generation performance. The consistent negative correlation with cosine distance suggests that subspace alignment is beneficial.

observation suggests a link between a model's ability to geometrically isolate dialectal subspaces and its downstream generative performance.

We quantify this by measuring the L2 and cosine distance between MSA and dialectal sentence representations across all layers (Figure 5). We observe that different distance metrics capture different geometric properties: L2 distance reflects the degree of spatial separation, while cosine distance measures the alignment of subspaces. To substantiate the link to performance, we compute the layer-wise correlation between these distances and the chrF++ score (Figure 6). A consistent negative correlation emerges between cosine distance and performance, especially in early-to-mid layers. This suggests that better alignment (lower cosine distance) in these layers is beneficial, likely facilitating the transfer of semantic information from the high-resource MSA. Conversely, the relationship with L2 distance is more complex, with models like Aya benefiting from greater spatial separation in intermediate layers. This indicates a delicate balance: subspaces must be aligned enough for knowledge transfer but separate enough to preserve unique dialectal features.

5.3 Exploration II: Information-Theoretic Evidence of MSA's Representational Dominance

The geometric analysis suggests entanglement with MSA is problematic. We further interrogate this using information-theoretic probing during standard supervised fine-tuning (SFT) on the VarMT task in Arabic. We track the ELBO code length required to identify dialects from the model's hidden states (a proxy for how accessible this information is). As shown in Figure 7, standard fine-tuning causes the code length for all dialects to increase slightly, as the model specializes for generation rather than classification. However, the increase is **disproportionately large for MSA**. This indicates that the model is actively making MSA-specific information less linearly accessible, suggesting its pre-trained MSA representation is oversized and detrimental to the dialectal generation task.

This "pruning" of the MSA subspace has a direct geometric consequence. As we fine-tune, the Subspace Angle (SSA) between MSA and the dialectal subspaces consistently increases (Figure 3, left). That is, the dialectal subspaces systematically drift away from the MSA subspace. Crucially, this growing separation directly correlates with improvements in generation performance (Figure 3, right).

Taken together, these analyses provide compelling correlational evidence for our central hypothesis: the representational dominance of the higher-resource varieties actively hinders a model's ability to generate text in related low-resource varieties. Fine-tuning implicitly alleviates this by pushing dialectal representations away from the MSA subspace.

There are a few limitations to keep in mind, the MADAR dataset, while unique in its breadth of dialects, is composed of relatively short sentences. This setting may not fully capture model behaviors on tasks requiring longer-form generation, thereby defining the scope of our current findings. We hope future work addresses this gap in data availability.

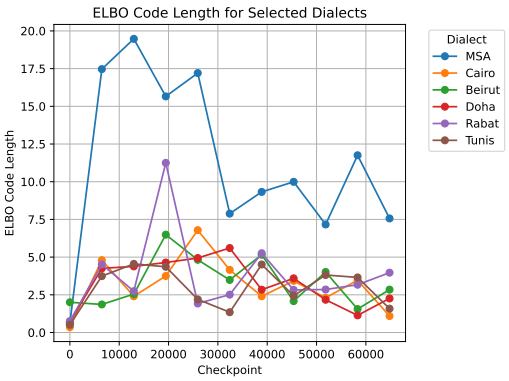

Figure 7: Code Length evolution over baseline training.

6 Discussion & Future Work

This work shows that representational entanglement with high-resource languages is a key bottleneck for generative modeling in closely-related, low-resource varieties. Using *online subspace decoupling*, we dynamically discourage dominance by higher-resource varieties during fine-tuning, establishing causal evidence across six language groups that managing subspace dominance yields substantial gains (up to +13.0 ChrF++). Geometric and information-theoretic analyses on Arabic dialects further suggest that Modern Standard Arabic (MSA) dominance impedes dialectal generation, highlighting the importance of representational allocation in multilingual models. Future directions include:

- **Scalable and Efficient Methods:** Developing computationally cheaper alternatives, such as subspace-aware adapters or pre/post-training objectives that balance representational spaces.
- **Inference-Time Interventions:** Using activation steering, targeted neuron editing, or distributional shifts to mitigate higher-resource interference without gradient updates.
- **Interpretability for Benefit Prediction:** Investigating which representational or linguistic factors (e.g., corpus size, syntactic divergence, shared subspaces) most strongly influence the gains from decoupling, enabling more principled and predictive model design.

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Se-

bastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. Aya 23: Open weight releases to further multilingual progress, 2024. URL https://arxiv.org/abs/2405.15032.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. The MADAR Arabic dialect corpus and lexicon. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1535.

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

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. The geometry of multilingual language model representations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 119–136, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.9. URL https://aclanthology.org/2022.emnlp-main.9.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2020.

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, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and The NLLB Team. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022. URL https://arxiv.org/abs/2207.04672.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) and the 11th International Joint Conference on Natural Language Processing (IJCNLP): System Demonstrations*, pp. 586–598. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-demo.67. URL https://aclanthology.org/2021.acl-demo.67.

T. Honkola, K. Ruokolainen, K. J. J. Syrjänen, L. Savolainen, M. Lehtinen, and K. Vesakoski. Evolution within a language: environmental differences contribute to divergence of dialect groups. *BMC Evolutionary Biology*, 18(132), 2018. doi: 10.1186/s12862-018-1238-6. URL https://doi.org/10.1186/s12862-018-1238-6.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*, 2020.

Karima Kadaoui, Samar Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. TARJAMAT: Evaluation of bard and ChatGPT on machine translation of ten Arabic varieties. In Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham (eds.), *Proceedings of ArabicNLP 2023*, pp. 52–75, Singapore (Hybrid), December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.arabicnlp-1.6. URL https://aclanthology.org/2023.arabicnlp-1.6.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. Quantifying the dialect gap and its correlates across languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7226–7245, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.481. URL https://aclanthology.org/2023.findings-emnlp.481.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6919–6971, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.384. URL https://aclanthology.org/2024.naacl-long.384.

Alissa Melinger. Distinguishing languages from dialects: A litmus test using the picture-word interference task. *Cognition*, 172:73–88, 2018. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2017.12.006. URL https://www.sciencedirect.com/science/article/pii/S0010027717303086.

Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13190–13208, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.879. URL https://aclanthology.org/2023.findings-emnlp.879/.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. Dolphin: A challenging and diverse benchmark for Arabic NLG. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1404–1422, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.98. URL https://aclanthology.org/2023.findings-emnlp.98.

Hellina Nigatu, Atnafu Tonja, and Jugal Kalita. The less the merrier? investigating language representation in multilingual models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12572–12589, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.837. URL https://aclanthology.org/2023.findings-emnlp.837.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (eds.), *Proceedings of the Tenth Workshop on Statistical Machine*

Translation, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. Frmt: A benchmark for few-shot region-aware machine translation. *Transactions of the Association for Computational Linguistics*, 11:671–685, 06 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00568. URL https://doi.org/10.1162/tacl_a_00568.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. AraBench: Benchmarking dialectal Arabic-English machine translation. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5094–5107, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.447. URL https://aclanthology.org/2020.coling-main.447.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models, 2023.

Cheril Shah, Yashashree Chandak, Atharv Mahesh Mane, Benjamin Bergen, and Tyler A. Chang. Correlations between multilingual language model geometry and crosslingual transfer performance. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4059–4066, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.361.

Gemma Team. Gemma 3. 2025a. URL https://goo.gle/Gemma3Report.

Qwen Team. Qwen3 technical report, 2025b. URL https://arxiv.org/abs/2505.09388.

Jörg Tiedemann. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri (eds.), *Proceedings of the Fifth Conference on Machine Translation*, pp. 1174–1182, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.139/.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. Sharing matters: Analysing neurons across languages and tasks in llms, 2024. URL https://arxiv.org/abs/2406.09265.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. Multi-VALUE: A framework for cross-dialectal English NLP. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 744–768, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.44. URL https://aclanthology.org/2023.acl-long.44.

A VarMT Prompts

Language Group	Prompt Template

Non-Arabic (e.g., Portuguese, Finnish, German, Malay, Czech/Slovak)	`Translate the following sentence from {src_lang_name} to {tgt_lang_name}:`
	`{src_lang_name}: {src_sentence}`
	`{tgt_lang_name}: {tgt_sentence}`

Arabic (MSA to dialect)	`Rewrite the following MSA sentence to the dialect of {city}:`
	`MSA: {msa}`
	`{city}:`

Arabic (MSA to English/French)	`Rewrite the following MSA sentence to {city}:`
	`MSA: {msa}`
	`{city}:`

Table 4: Prompting templates used for fine-tuning. For non-Arabic pairs we use a direct *Translate the following sentence* prompt, while for Arabic we adopt a *Rewrite to dialect* formulation that mirrors natural usage of MSA as the standard reference. For instruction-tuned or chat models, these prompts are wrapped inside the model's recommended system/user templates.

B Data Processing and Splits

To ensure consistency across language groups and prevent data leakage, we follow the principles below:

B.1 Arabic (Madar 26)

For Arabic, we exclusively use the Madar 26 split. Both the translation models and dialect identification probes are trained on this split. The official test set is **never used during training or probing** and serves solely for final evaluation, ensuring no data leakage.

B.2 FLORES-200

For FLORES-200, we use the `devtest` split as the training/dev set and the `test` split as the test set. Because FLORES-200 is already small, we do **not perform additional sampling**. Otherwise, we follow the same principles as for other datasets: probes are trained on the same set as the translation model, and the test set is kept fully separate to prevent leakage.

B.3 Consistent Train/Dev Splits for Probing

Across all languages and models, probes are always trained on the **same training/dev set as the translation model**, while the test split is kept entirely separate. This ensures that no information from the test set can influence probe training or model tuning.

B.4 Low-Resource / Few-Shot Sampling

For large datasets (e.g., Tatoeba or other multilingual resources), we adopt a **controlled low-resource setup** to normalize the training regime across languages:

- From any dataset with more than approximately 1000 parallel training samples, we create a **training/dev split of 1000 parallel sentences**.
- The test split also consists of **1000 parallel sentences**.
- This approach allows us to simulate a few-shot scenario and maintain comparability between high- and low-resource language pairs.

For example, in the Tatoeba preprocessing pipeline:

```
de_lo_mt = make_mt_dataset(de_lo, "de", "lo", dev_size=1000
, test_size=1000)
de_lo_id = make_dialect_id_dataset(de_lo_mt["dev"].to_pandas(),
"de", "lo", 1, 0)
```

B.5 Portuguese (FRMT Dataset)

For Brazilian and European Portuguese, we process multiple FRMT buckets to extract aligned sentence pairs. The workflow mirrors the Tatoeba setup:

- Merge BR/PT parallel sentences on the English pivot to create the translation dataset (`tr_dataset`).
- Sample **1000 sentences per variant** for dialect identification probes (`id_dataset`).
- Push both datasets to Hugging Face Hub for standardized access.

```
tr_dataset = build_translation_dataset(files)
id_dataset = build_dialect_id_dataset(files)
```

B.6 Kven–Finnish

Kven–Finnish is an inherently low-resource language pair, with only 797 total parallel sentence samples in Tatoeba. To handle this, we create a training/dev set of 500 samples and use the remaining samples as the test set. This setup ensures the training and probing data remain separate from evaluation data while respecting the limited resource size.

B.7 Rationale

This uniform low-resource setup across all language groups ensures comparability, even though **parallel sentence availability varies greatly** across language pairs. For instance, Kven–Finnish has far fewer resources than German–Low German or Portuguese. Limiting all datasets to a few-shot regime allows systematic study of translation and dialect probing under consistent conditions. We release all the splitting code for reproducibility (with fixed random seeds, however we can not share the data directly as we do not have permission to do so.

C City Names to Dialect Code for Arabic

D More Information About Probing

To complement geometric subspace analysis, we adopt an information-theoretic variational linear probe (Voita & Titov, 2020; Müller-Eberstein et al., 2023) to quantify how much dialect identity information is recoverable from token-level model representations. For a given token, let $\{\mathbf{h}^{(0)}, \ldots, \mathbf{h}^{(\ell)}\} \in \mathbb{R}^d$ denote its hidden states from all ℓ layers, including the non-contextualized layer 0. The probe computes a learned weighted average over layers:

$$\mathbf{h}' = \sum_{i=0}^{\ell} \alpha_i \mathbf{h}^{(i)},$$

where $\boldsymbol{\alpha} \in \mathbb{R}^\ell$ are learned combination weights.

City	Code
Rabat	RAB
Fes	FES
Algiers	ALG
Tunis	TUN
Sfax	SFX
Tripoli	TRI
Benghazi	BEN
Cairo	CAI
Alexandria	ALX
Aswan	ASW
Khartoum	KHA
Jerusalem	JER
Amman	AMM
Salt	SAL
Beirut	BEI
Damascus	DAM
Aleppo	ALE
Mosul	MOS
Baghdad	BAG
Basra	BAS
Doha	DOH
Muscat	MUS
Riyadh	RIY
Jeddah	JED
Sana'a	SAN

Table 5: City Names and Their Codes

This aggregated representation is fed to a linear classifier with weight matrix $\boldsymbol{\theta} \in \mathbb{R}^{d\times c}$ for c dialect classes. Following Voita & Titov (2020), each weight w in $\boldsymbol{\theta}$ is drawn from a normal distribution

$$w \sim \mathcal{N}(z\mu, z^2\sigma^2),$$

where the scaling factor z is also drawn from

$$z \sim \mathcal{N}(\mu_z, \sigma_z^2).$$

The pair (w, z) is given a joint normal–Jeffreys prior

$$\gamma(w, z) \propto |z|^{-1}\, \mathcal{N}(w \mid 0, z^2)$$

which encourages sparsity by pushing weights toward zero with low variance.

The probe parameters $(\boldsymbol{\alpha}, \boldsymbol{\theta})$ are trained to minimize

$$\mathcal{L} = \mathrm{CE}(y, \hat{y}) \;+\; \beta\, D_{\mathrm{KL}}(q(\boldsymbol{\theta}) \,\|\, \gamma(\boldsymbol{\theta}))\,,$$

where CE is the cross-entropy loss for one-vs-rest dialect classification, and the KL term regularizes $\boldsymbol{\theta}$ toward the sparsity-inducing prior. This objective maximizes compression while preserving predictive accuracy, yielding a layer-combined, token-level estimate of recoverable dialect identity information. The one-vs-rest objective hones in on dialect specific information that can help the model discern between similar dialects and offers counter-examples. We construct the training set for each dialect/variety/language by taking all the target's sentences in MADAR 26's training set, we construct an equal number of counter-examples from all the other dialects and languages. We make this data available (anonymized). We include training hyperparameters for the probes in Table 6.

E Online Decoupling Training Details

This appendix outlines the key design decisions underlying our online higher-resource variety subspace decoupling method, as well as the exact hyperparameters used in our experiments.

Hyperparameter	Value
Model name	`google/gemma-3-1b-pt`
KL weight	1.0
Number of epochs	30 (for analysis) 15 (for decoupling training)
Early stopping patience	5

Table 6: Training hyperparameters for variational probe experiments.

E.1 Design Choices

Projection Subspace Estimation. We estimate the higher-resource variety subspace using a *variational linear probe* trained on a higher-resource variety vs. lower-resource variety identification task over the training sets of the corpora used. We recover the subspace basis from the learned probe parameters using Singular Value Decomposition (SVD) of the parameter matrix θ_{HR}. The number of retained singular vectors equals the probe's latent dimension.

Online Updating. Rather than estimating the higher-resource variety subspace once before training, we periodically retrain the probe on the current model checkpoint during fine-tuning. This ensures that the projection matrix $\mathbf{P}_{\text{HR}}$ remains synchronized with the evolving hidden representation geometry. The projection matrix is updated every N_{update} gradient steps. We ablate the choice of N_{update} in Appendix E.3.

Layer Aggregation. Hidden representations from all layers are combined using a learned set of attention weights $\alpha \in \mathbb{R}^{L+1}$ from the variational probe. This allows the method to focus the decoupling penalty on layers most predictive of MSA features.

Penalty Formulation. We penalize the ℓ_2 norm of the projection of the aggregated hidden states onto the higher-resource subspace:

$$\mathcal{L}_{\text{decouple}} = \mathbb{E}\left[\|\mathbf{H}\mathbf{P}_{\text{HR}}\|_2\right], \tag{2}$$

where $\mathbf{H}$ are the contextual hidden states and $\mathbf{P}_{\text{HR}}$ is the projection matrix.

Loss Weighting. The decoupling penalty is scaled by a coefficient λ and added to the standard causal language modeling loss:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda \cdot \mathcal{L}_{\text{decouple}}. \tag{3}$$

Bidirectional Training Data. To encourage symmetric modeling of both higher-resource variety $\rightarrow$ lower-resource variety and dialect $\rightarrow$ MSA directions, we construct bidirectional rewriting prompts for each sentence pair.

E.2 Loss Coefficient λ Ablation

λ	Average ChrF++	std
1e-4	**21.8749**	2.0365
1e-3	18.1038	1.4634
1e-2	20.8508	1.3012
0.1	20.7082	1.7187
1.0	11.4210	1.0222
10.0	7.0045	0.4444

Table 7: Average Chrf++ and standard deviation across Arabic dialects over several values of λ.

N_{update}	Average ChrF++	std
100	21.8	1.8
500	22.7	2.0
1000	21.5	1.8

Table 8: Average Chrf++ and standard deviation across Arabic dialects over several values of N_{update}.

Parameter	**Value / Setting**
Base model	`google/gemma-3-1b-pt`
Tokenizer	Matching HF tokenizer (`pad_token = eos_token`)
Batch size (per device)	1
Gradient accumulation steps	4
Max sequence length	512
Optimizer	AdamW (via HF Trainer default)
Learning rate	5×10^{-5} (default HF schedule)
Loss coefficient λ	1e-4
Probe update steps N_{update}	500
Probe training epochs	15
Probe input type	Sequence-level dialect identification
Number of probe classes	2 (MSA vs. non-MSA)
Projection estimation	SVD on θ_{HR}
Subspace dimensionality	Full rank of θ_{HR}
Layer aggregation	Learned attention weights α
Early stopping patience	3 epochs (validation loss)
Early stopping threshold	0.01
Train/validation split	90% / 10%

Table 9: Hyperparameters used in online decoupling experiments.

E.3 N_{UPDATE} ABALATION

E.4 HYPERPARAMETERS

F DETAILED DECOUPLING RESULTS FOR ALL VARIETIES

G STATISTICAL SIGNIFICANCE TESTING PER VARIETY/SETUP

H LLM USE

We utilize LLM assistants in this paper as follows:

- **Paper Writing:** LLMs are used to polish language and style, as well as for brevity and phrasing throughout this paper. The analysis, however, is originally drafted by the authors.
- **Coding:** LLM assistants were used to help draft and clean the code used for our methodology, experimentation, and visualization. The code was manually reviewed and tested/reviewed for correctness.

src	target	online	baseline	random	delta_online_baseline	delta_online_random
European Portuguese	Brazilian Portuguese	45.900	32.900	_	13.000	_
Brazilian Portuguese	European Portuguese	46.200	34.500	_	11.700	_
English	Indonesian	50.100	44.800	_	5.300	_
Finnish	Kven Finnish	50.400	45.800	_	4.600	_
Modern Standard Arabic	Sanaa	22.700	18.400	17.900	4.300	4.800
Modern Standard Arabic	Benghazi	24.800	21.000	18.300	3.800	6.500
Low German	Standard German	49.500	46.300	_	3.200	_
Modern Standard Arabic	Riyadh	26.200	23.000	19.900	3.200	6.300
Standard German	Low German	52.300	49.600	_	2.700	_
Modern Standard Arabic	Cairo	19.400	16.900	16.100	2.500	3.300
English	Malay	50.000	47.800	_	2.200	_
Modern Standard Arabic	Basra	22.300	20.100	17.500	2.200	4.800
Modern Standard Arabic	Muscat	20.500	18.600	18.000	1.900	2.500
Modern Standard Arabic	Mosul	22.700	20.900	17.700	1.800	5.000
Modern Standard Arabic	Fes	24.000	22.500	21.200	1.500	2.800
Modern Standard Arabic	Jerusalem	24.100	22.600	18.500	1.500	5.600
Kven Finnish	Finnish	40.500	39.100	_	1.400	_
Modern Standard Arabic	Salt	23.400	22.000	17.900	1.400	5.500
Modern Standard Arabic	Aleppo	21.800	20.600	17.200	1.200	4.600
Modern Standard Arabic	Khartoum	23.300	22.100	18.200	1.200	5.100
Modern Standard Arabic	Baghdad	21.500	20.400	15.700	1.100	5.800
Modern Standard Arabic	Aswan	21.300	20.200	18.800	1.100	2.500
English	Slovak	32.300	31.300	_	1.000	_
Modern Standard Arabic	Tripoli	22.100	21.200	18.800	0.900	3.300
Modern Standard Arabic	Doha	22.700	21.800	17.700	0.900	5.000
Modern Standard Arabic	Rabat	20.700	19.900	19.000	0.800	1.700
Modern Standard Arabic	Alexandria	22.700	21.900	19.100	0.800	3.600
Modern Standard Arabic	Jeddah	22.000	21.200	18.000	0.800	4.000
Modern Standard Arabic	Amman	21.100	20.400	16.700	0.700	4.400
Modern Standard Arabic	Beirut	19.000	18.500	16.700	0.500	2.300
Modern Standard Arabic	Tunis	18.400	17.900	16.000	0.500	2.400
English	Czech	31.600	32.200	_	-0.600	_
Modern Standard Arabic	Sfax	17.500	18.200	18.100	-0.700	-0.600
Modern Standard Arabic	English	30.700	31.400	22.900	-0.700	7.800
Modern Standard Arabic	Algiers	22.700	23.600	20.100	-0.900	2.600
Modern Standard Arabic	Damascus	19.600	20.600	17.100	-1.000	2.500
Modern Standard Arabic	French	25.900	27.700	22.000	-1.800	3.900
Muscat	Modern Standard Arabic	15.800	19.200	_	-3.400	_
Khartoum	Modern Standard Arabic	15.600	19.400	_	-3.800	_
Algiers	Modern Standard Arabic	14.400	18.900	_	-4.500	_
Riyadh	Modern Standard Arabic	14.100	19.000	_	-4.900	_
Jeddah	Modern Standard Arabic	13.700	18.900	_	-5.200	_
Aswan	Modern Standard Arabic	13.100	18.300	_	-5.200	_
Fes	Modern Standard Arabic	13.800	19.100	_	-5.300	_
Cairo	Modern Standard Arabic	13.700	19.000	_	-5.300	_
Tripoli	Modern Standard Arabic	12.800	18.200	_	-5.400	_
Salt	Modern Standard Arabic	13.200	18.700	_	-5.500	_
Aleppo	Modern Standard Arabic	12.800	18.300	_	-5.500	_
Baghdad	Modern Standard Arabic	13.700	19.200	_	-5.500	_
Basra	Modern Standard Arabic	13.000	18.600	_	-5.600	_
Jerusalem	Modern Standard Arabic	12.800	18.500	_	-5.700	_
Sanaa	Modern Standard Arabic	13.000	18.700	_	-5.700	_
Alexandria	Modern Standard Arabic	13.200	19.000	_	-5.800	_
Benghazi	Modern Standard Arabic	13.300	19.100	_	-5.800	_
Sfax	Modern Standard Arabic	11.800	17.600	_	-5.800	_
Rabat	Modern Standard Arabic	12.600	18.600	_	-6.000	_
Mosul	Modern Standard Arabic	12.400	18.600	_	-6.200	_
Amman	Modern Standard Arabic	13.400	19.600	_	-6.200	_
Doha	Modern Standard Arabic	12.900	19.200	_	-6.300	_
Damascus	Modern Standard Arabic	12.700	19.200	_	-6.500	_
Beirut	Modern Standard Arabic	12.000	18.500	_	-6.500	_
Tunis	Modern Standard Arabic	11.400	18.300	_	-6.900	_
French	Modern Standard Arabic	0.300	17.700	_	-17.400	_
English	Modern Standard Arabic	0.300	18.400	_	-18.100	_

Table 10: All the results across all of our experimental settings.

Group	Target	N	CHRFF(base)	CHRFF(online)	Δ	$p_{1-sided}$	sig
Arabic	Aleppo	200	21.31	20.45	-0.87	0.501	
Arabic	Alexandria	200	22.97	20.75	-2.22	0.706	
Arabic	Algiers	200	24.50	20.01	-4.49	0.999	
Arabic	Amman	200	21.37	23.80	2.43	0.0128	*
Arabic	Aswan	200	20.23	20.49	0.26	0.178	
Arabic	Baghdad	200	21.07	20.56	-0.51	0.65	
Arabic	Basra	200	20.06	21.04	0.98	0.153	
Arabic	Beirut	200	18.94	19.08	0.13	0.224	
Arabic	Benghazi	200	22.16	23.00	0.84	0.147	
Arabic	Cairo	200	16.50	18.22	1.72	0.00656	*
Arabic	Damascus	200	21.97	22.21	0.24	0.113	
Arabic	Doha	200	23.28	24.09	0.81	0.0677	
Arabic	English	200	33.69	28.99	-4.70	0.999	
Arabic	Fes	200	23.24	19.77	-3.47	0.998	
Arabic	French	200	27.52	23.65	-3.87	1	
Arabic	Jeddah	200	23.23	21.95	-1.29	0.515	
Arabic	Jerusalem	200	24.15	23.59	-0.56	0.502	
Arabic	Khartoum	200	23.30	21.86	-1.44	0.691	
Arabic	Mosul	200	21.75	21.17	-0.58	0.814	
Arabic	Muscat	200	18.43	19.76	1.33	0.0133	*
Arabic	Rabat	200	19.59	19.02	-0.58	0.641	
Arabic	Riyadh	200	24.71	25.27	0.56	0.238	
Arabic	Salt	200	22.61	25.42	2.81	0.00118	*
Arabic	Sanaa	200	18.36	19.98	1.62	0.165	
Arabic	Sfax	200	17.98	17.26	-0.72	0.647	
Arabic	Tripoli	200	21.78	20.42	-1.36	0.661	
Arabic	Tunis	200	18.11	17.25	-0.86	0.777	
Czech-slovak	eng_Latn_to_ces_Latn	1012	29.33	28.63	-0.71	0.964	
Czech-slovak	eng_Latn_to_slk_Latn	1012	28.44	29.38	0.94	0.00086	*
German-low_german	de_to_lo	1000	52.10	54.50	2.40	1.85e-06	*
German-low_german	lo_to_de	1000	48.59	51.49	2.90	5.22e-07	*
Indo-malay	eng_Latn_to_ind_Latn	1012	42.47	47.96	5.49	3.81e-41	*
Indo-malay	eng_Latn_to_zsm_Latn	1012	44.91	47.22	2.30	3.23e-12	*
Kven-finnish	fi_to_fkv	297	47.90	53.10	5.20	0.000155	*
Kven-finnish	fkv_to_fi	297	41.49	43.09	1.61	0.0896	
Portuguese	br_to_pt	985	35.24	45.88	10.64	6.42e-64	*
Portuguese	pt_to_br	985	33.01	45.24	12.23	2.28e-74	*

Table 11: Individual Wilcoxon p-test on each translation direction (online vs. baseline SFT).