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 multilingual modeling in large language models (LLMs). We challenge this view with the first intervention-based study showing that excessive representational entanglement with dominant varieties can degrade generative quality in machine translation, suggesting a causal link between representational dominance and weaker downstream performance on low-resource varieties. We introduce an online variational probing fine-tuning method that continuously estimates the subspace of a dominant variety during generative fine-tuning (mainly translation) and penalizes it to reduce its span. Across six language families, reducing alignment consistently improves low-resource translation quality, with gains of up to +11.7 ChrF++ / +10.1 COMET for European Portuguese, +5.3 / +4.3 for Indonesian, +4.6 / +4.2 for Kven Finnish, and +2.7 / +2.1 for Low German. In Arabic, several dialects improve by up to +4.7 ChrF++ and +1.4 COMET despite sharp drops for cross-lingual tasks (e.g., translation to MSA, English, or French), suggesting that the effect extends beyond simple cross-lingual alignment. Alongside these intervention results, we present qualitative and geometric analyses that further support our hypothesis. Together, our findings show that disentangling high-resource subspaces can unlock representational capacity for related low-resource varieties and provide a practical means of 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 intervention promotes orthogonal representations and improves generative capacity for lower-resource varieties, allowing us to study the effect of representational entanglement on downstream task performance beyond simple correlation. Then, we shift to a more qualitative/observational analysis honing in on Arabic to analyze 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 largely consistent improvements over standard fine-tuning, boosting lower-resource performance, including **+11.7** ChrF++ / **+10.1** COMET for European Portuguese, **+5.3 / +4.3** for Indonesian, **+4.6 / +4.2** for Kven Finnish, and **+2.7 / +2.1** for Low German. In Arabic, several dialects improve by up to **+4.7** ChrF++ and **+1.4** COMET 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 mechanistic 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 finetuning method that improves generative performance on machine translation 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 multilingual LLMs allocate representational capacity across closely related language varieties.

Multilingualism in Large Language Models. Work on multilingual LLMs shows that models distribute linguistic knowledge unevenly, with architectural and activation-level analyses revealing language-specific neuron sharing (Wang et al., 2024; Kojima et al., 2024). Studies of representational dominance find that LLMs, especially English-centric ones, bias toward high-resource languages (Wendler et al., 2024; nostalgebraist, 2020), and that non–English-centric models reduce but do not eliminate this effect (Zhong et al., 2025). Other work suggests that imbalance can sometimes aid transfer (Schäfer et al., 2024) or support cross-lingual abstractions (Brinkmann et al., 2025). Our contribution is to examine dominance explicitly and geometrically by testing whether similar varieties form separable subspaces and showing that entanglement predicts generative failures under resource imbalance. This extends geometric findings from encoder models (Chang et al., 2022; Shah et al., 2024) to large generative LLMs. Consistent with evidence that models struggle with dialectal nuance (Nigatu et al., 2023), we provide mechanistic evidence that reducing representational entanglement improves generation for closely related varieties.

Information-Theoretic Probing. Information-theoretic probes have been used to measure linguistic information in representations (Voita & Titov, 2020; Müller-Eberstein et al., 2023). We extend their use from analysis to training: our "variety probes" continuously estimate dominant subspaces during fine-tuning and penalize entanglement, turning probing into a mechanism for mechanistic representational control. Unlike earlier work focusing on specific linguistic features (e.g., POS tagging), we generalize probes to capture cross-varietal geometry and integrate them directly into the model's learning process.

Dialectal and Low-Resource NLP. Dialectal variation remains a key challenge in multilingual generation, with large performance gaps as dialects diverge from standardized forms (Kantharuban et al., 2023; Ziems et al., 2023). For Arabic, datasets such as AraBench (Sajjad et al., 2020) and MADAR (Bouamor et al., 2018) enable evaluation, while works like Kadaoui et al. (2023) and Nagoudi et al. (2023) examine translation across dialects. Broader multilingual efforts, including the Tatoeba challenge (Tiedemann, 2020) and FRMT (Riley et al., 2023), address low-resource translation. Our contribution differs by examining how varieties are represented inside LLMs and how direct interventions on subspaces can improve generation. While Arabic provides a rich testbed, the findings generalize to other varieties with high lexical and orthographic overlap.

Orthogonal Subspace Methods. Subspace orthogonality has been explored for mitigating interference in continual learning (Saha et al., 2021; Wang et al., 2023; Farajtabar et al., 2020) and for multi-objective alignment in LLMs (Lin et al., 2025). Our work draws from this literature but differs in scope and mechanism: rather than enforcing orthogonality between tasks, we *implicitly encourage* it by penalizing projections onto dominant high-resource subspaces. To our knowledge, this is the first use of orthogonalization-based interventions to study and control representational alignment across languages and dialects.

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 mechanistically 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 I.

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. On this setup, lexical similarity and shared script cannot be exploited directly. This setup controls for surface overlap, introduces a neutral semantic anchor, and extends our analysis from intra-varietal to cross-lingual alignment, offering a more robust test of the generality of our intervention.

4.2 Quantifying Performance and Representational Geometry

Evaluation. We evaluate generation quality primarily using primarily chrF++ (Popović, 2015), a character n-gram F-score, and also COMET (Rei et al., 2022), a neural quality estimation metric trained to predict human judgments and a standard in machine translation evaluation. ChrF++'s character-level design makes it well-suited for morphologically rich languages and robust to the minor lexical variations common across varieties, while remaining sensitive to subtle orthographic shifts making it a suitable primary metric. COMET complements this by leveraging pretrained multilingual representations to model semantic adequacy and fluency, offering a more holistic measure of translation quality. However, like other automatic metrics, neither fully captures the nuances of varietal distinctness, that is, whether the output reflects the intended dialectal features rather than generic correctness. Human evaluation would provide the most reliable judgment, but is difficult at the breadth that we aim for with this study given the limited availability of native speakers for many varieties, while LLM-based evaluators risk reintroducing the same high-resource biases our work aims to mitigate.

4.3 Mechanistic Intervention: Online Subspace Decoupling

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

The procedure is as follows:

1. **Identify Higher-resource Variety Direction.** We train a variational linear probe (as in Sec. 4.4) to distinguish the higher-resource variety (positive class) from all other varieties in its group (negative class). The probe's learned weight vector $\boldsymbol{\theta}_{\text{HR}} \in \mathbb{R}^d$ defines the most discriminative direction separating higher- from lower-resource examples. We treat this normalized vector as a one-dimensional subspace representing the higher-resource variety:

$$\mathbf{u}_{\text{HR}} = \frac{\boldsymbol{\theta}_{\text{HR}}}{\|\boldsymbol{\theta}_{\text{HR}}\|}, \qquad \mathbf{P}_{\text{HR}} = \mathbf{u}_{\text{HR}}\mathbf{u}_{\text{HR}}^{\top}.$$

   This projection matrix $\mathbf{P}_{\text{HR}}$ therefore captures the component of any hidden representation that lies along the high-resource direction.

2. **Define Decoupling Loss.** During fine-tuning on the VarMT task, we add a penalty term to the standard language modeling loss that discourages alignment of model hidden states $\mathbf{H}$ with the high-resource direction:

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

> This loss minimizes the magnitude of the projection of $\mathbf{H}$ onto the high-resource direction, effectively reducing the extent to which all tokens encode features associated with the high-resource variety. We apply this penalty to **all** tokens, including those of the high-resource variety, to ensure that the representational geometry itself, rather than only specific examples, becomes disentangled. The underlying intuition is that the internal representational real estate allocated to the high-resource variety should be globally constrained, preventing it from dominating the shared latent space. While this design choice may not yield the most empirically optimal results, and may indeed be overly harsh for the high-resource variety, it enables a controlled investigation of the effect of a global representational intervention, independent of token-level specifics. Future work may explore the specifics of when and where to penalize the subspaces for more effective training. The total loss is
>
> $$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda\, \mathcal{L}_{\text{decouple}},$$
>
> where λ is a scaling hyperparameter (we use $\lambda = 10^{-4}$ across all setups; see Appendix E.2).

Crucially, the probe is periodically retrained on updated model checkpoints during fine-tuning. This **online updating** of $\mathbf{P}_{\text{HR}}$ ensures that our intervention continuously tracks the evolving high-resource direction, preventing the probe from becoming stale and enabling a precise and adaptive causal manipulation of the model's representational geometry. Additional design details are provided 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. In our case, since the subspace is being estimated by a vector, $\boldsymbol{\theta}_{\text{variety}}$, (hence in one dimension) this corresponds to measure the angle between the vectors corresponding to different varieties. Namely:

$$\textbf{SSA} = \arccos\left(\frac{|\boldsymbol{\theta}_{\text{x}}^{\top}\boldsymbol{\theta}_{\text{y}}|}{\|\boldsymbol{\theta}_{\text{x}}\|_2\,\|\boldsymbol{\theta}_{\text{y}}\|_2}\right),$$

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 in form to the probe used in our online subspace decoupling intervention) (Voita & Titov, 2020; Müller-Eberstein et al., 2023). This probe implements the *minimum description length* (MDL) formulation of information-theoretic probing introduced by Voita & Titov (2020), which quantifies both the predictability and the complexity of the probe. Rather than maximizing a mutual information bound, the probe minimizes the expected *codelength* required to transmit both the data and the probe parameters, given by the variational evidence lower bound:

$$\mathcal{L}_{\text{probe}} = -\mathbb{E}_{\boldsymbol{\theta}\sim q(\boldsymbol{\theta})}\big[\log p_{\boldsymbol{\theta}}(y \mid x)\big] + \beta\, D_{\text{KL}}(q(\boldsymbol{\theta}) \,\|\, \gamma(\boldsymbol{\theta}))\,,$$

where $q(\boldsymbol{\theta})$ is the learned posterior over probe weights and $\gamma(\boldsymbol{\theta})$ is a sparsity-inducing prior. The cross-entropy term measures the probe's fit to the data, while the KL term regularizes its complexity by penalizing deviation from the prior. Minimizing this objective corresponds to compressing (x, y) pairs in as few bits as possible, providing an information-theoretic measure of how compactly variety identity is encoded in the model's hidden states. The probe's linear form facilitates interpretability and integration into our geometric analysis (e.g., for computing subspace angles) and online subspace decoupling training. Exploring more complex probe architectures and objectives is left to future work. Further details are provided in Appendix D. For tractability, when this probe is used observationally and not part of our intervention study it is applied to Arabic varieties only.

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 precise 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 mechanistic 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 an intervention on multiple language groups, establishing evidence consistent with a causal relationship, 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 Intervention-based Validation: Online Subspace Decoupling Boosts Performance

We test the hypothesis that excessive representational dominance and conflation with high-resource varieties impair the generative abilities of multilingual LLMs on low-resource varieties using our proposed **Online Subspace Decoupling** method (Section 4.3). This method introduces an explicit penalty term that discourages oversized higher-resource variety subspaces during fine-tuning. The specific higher-resource varieties penalized in each family are listed in Table 2, selected based on prior evidence of performance disparities (Kantharuban et al., 2023) and their status as the standard within each group, where 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

In Figure 1, 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++ and +10.1 COMET 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++ / +4.2 COMET), Low German (+2.7 / +2.1).

Importantly, dominant high-resource varieties do not necessarily suffer under decoupling (Brazilian Portuguese, Indonesian, and Standard German all remain stable or even improve) supporting the claim that the method reallocates representational capacity toward underrepresented varieties rather than amplifying dominant ones. A one-sided Wilcoxon signed-rank test on overall ChrF++ confirms that online decoupling significantly outperforms baseline fine-tuning across variety setups (excluding variety→MSA translation), yielding $p = 0.00195$. Online decoupling achieves higher ChrF++ in 9 of 10 setups, with an average gain of +4.45 points. Sentence-level Wilcoxon tests further show significant gains ($p < 0.05$) for nearly all varieties, except Finnish, Czech, and several

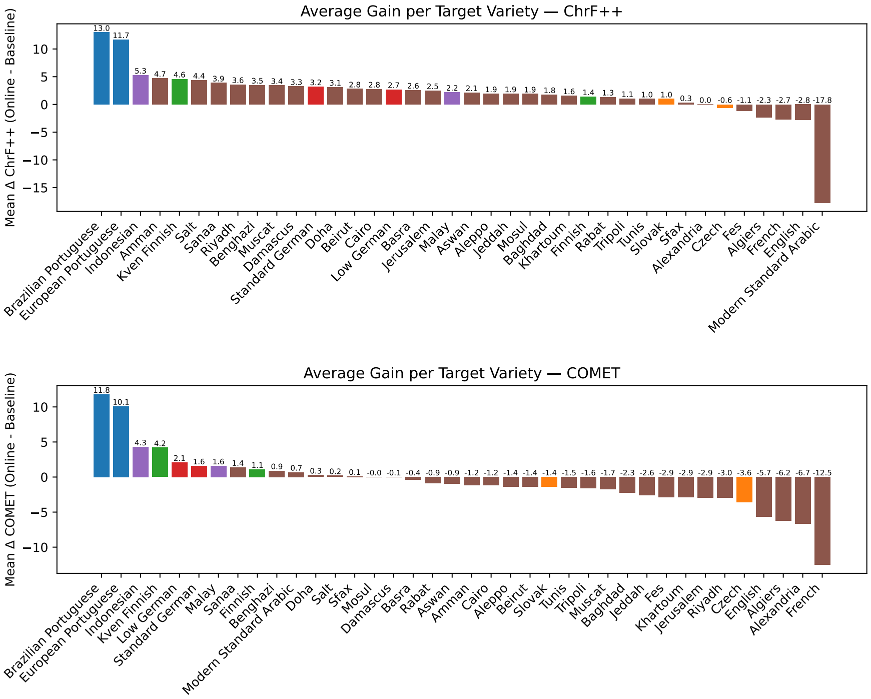

Figure 1: The mean delta in ChrF++ (**Top**) and COMET (**Bottom**) on several target varieties and languages between our Online Decoupling Training and Baseline SFT on VarMT. A positive delta indicates superior performance from our method.

Arabic dialects (notably Amman, Cairo, Muscat, and Salt do reach significance; see Appendix G). Together, these results demonstrate that decoupling systematically benefits low-resource varieties without consistently harming their high-resource counterparts.

Arabic dialects provide further support for this hypothesis. Constraining Modern Standard Arabic (MSA) subspaces yields gains of up to +4.7 ChrF++ (Amman) 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 expressivity. Online Subspace Decoupling effectively reallocates capacity to underrepresented dialects, unlocking performance otherwise constrained by MSA.

Interestingly, some high-resource varieties also benefit from decoupling: Indonesian gains +5.3 ChrF++ / +4.3 COMET, and Brazilian Portuguese +13.0 / +11.8, the largest observed increase. This indicates that entangled subspaces can distort both high- and low-resource varieties. Disentangling may sharpen boundaries between varieties, reduce interference, and enable more stable specialization. Penalizing oversized subspaces can also prevent dominant varieties from overfitting shared structures, benefiting **both** high- and low-resource generation in **some** instances.

While COMET scores largely follow the same trends observed across most language families, showing similar gains in Portuguese, Indo-Malay, German, and Kven-Finnish, they diverge for Arabic and Czech–Slovak. In these cases, more Arabic dialects and both Czech and Slovak (the latter with a smaller drop of -1.4) exhibit COMET decreases not mirrored in ChrF++. This discrepancy may suggest that our intervention enhances surface-level, variety-specific realization while slightly compromising semantic adequacy or fluency. Interestingly, MSA shows the opposite pattern: COMET remains largely stable despite heavy losses in ChrF++. Arabic and Czech–Slovak also start from lower baseline COMET scores (around the 50s, compared to the 60s for other families; see Table 11), reflecting weaker baseline representations in the underlying model. Consequently, decoupling may further expose fragility in these languages. Future work should investigate how representational disentanglement interacts with semantic and fluency-oriented metrics across resource levels.

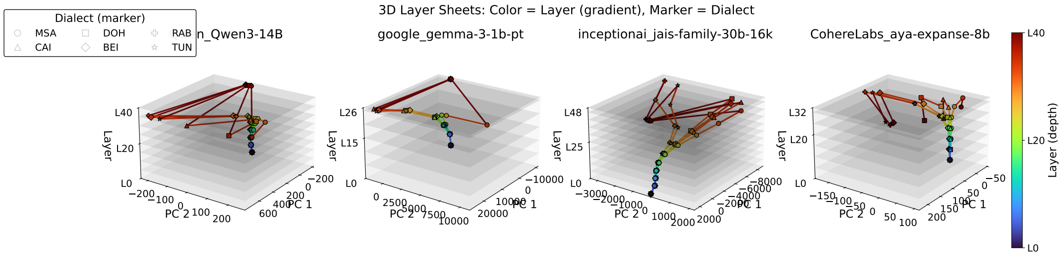

Figure 2: Layer-wise representational trajectories of the same sentence written in six Arabic varieties in four models.

Finally, to rule out the possibility that improvements stem from generic regularization rather than targeted disentanglement, we test random subspace shrinking on Arabic dialects. As shown in Table 3, performance consistently drops below baseline for MSA, the dialects, French, and English, confirming that gains arise specifically from penalizing oversized high-resource subspaces rather than from indiscriminate regularization.

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

5.2 Exploration I: Geometric Analysis Links Performance to Representational Separation

To complement our intervention-based experiments, we now examine how representational geometry evolves across layers and model families. Figure 2 visualizes the trajectories of six Arabic varieties (MSA, Cairo, Doha, Beirut, Rabat, Tunis) for the same sentence across model layers. Each point represents a variety's layer-wise hidden-state centroid projected into a shared PCA space (x–y), with color indicating layer depth (cool colors for early layers, warm colors for higher ones) and marker shape denoting dialect. The z-axis corresponds to layer index, forming a vertical progression through the model. In this view, early layers cluster tightly, reflecting shared low-level linguistic processing, while mid and upper layers begin to diverge, revealing distinct representational directions for each variety. Larger models such as Qwen3-14B, Aya-Expanse-8B, and Jais Family 30B show stronger, more stable separation across higher layers, indicating clearer dialectal structuring and more disentangled latent spaces. In contrast, smaller models like Gemma-3-1B-PT exhibit overlapping and less consistent trajectories, suggesting weaker specialization. Overall, the figure demonstrates that increasing model capacity (and as a result performance) may lead to more organized and semantically stable cross-varietal representations.

We quantify these patterns by measuring the L2 and cosine distances between MSA and dialectal sentence representations across all layers (Figure 3). The two metrics seem to capture complementary geometric aspects: L2 distance may reflect the extent of spatial separation, while cosine distance seems to reflect directional (subspace) alignment. To link these geometric patterns to generation quality, we compute the layer-wise Pearson correlation between representational distance and chrF++ scores (Figure 4). A consistent negative correlation emerges between cosine distance and performance, particularly in early to mid layers, suggesting that stronger directional alignment with MSA facilitates the transfer of semantic information from the high-resource variety.

Conversely, the relationship with L2 distance is more nuanced: models such as Aya benefit from greater spatial separation in intermediate layers, whereas Qwen exhibits the opposite tendency. When we apply our Online Decoupling intervention on Gemma 3 1B PT (Appendix, Figure 7), L2 distance increases relative to both baseline finetuning and the pretrained model, while cosine distance trends remain largely unchanged. This pattern pro-

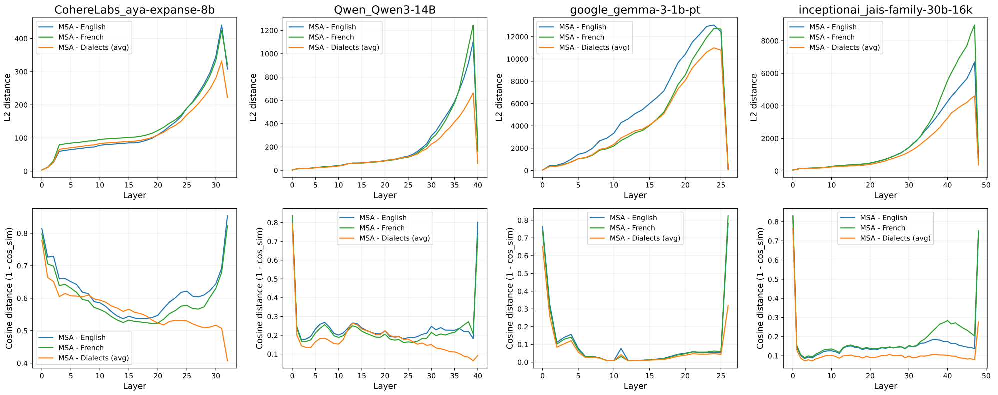

Figure 3: Layer-wise L2 (Top) and Cosine (Bottom) distance between dialectal representations and MSA.

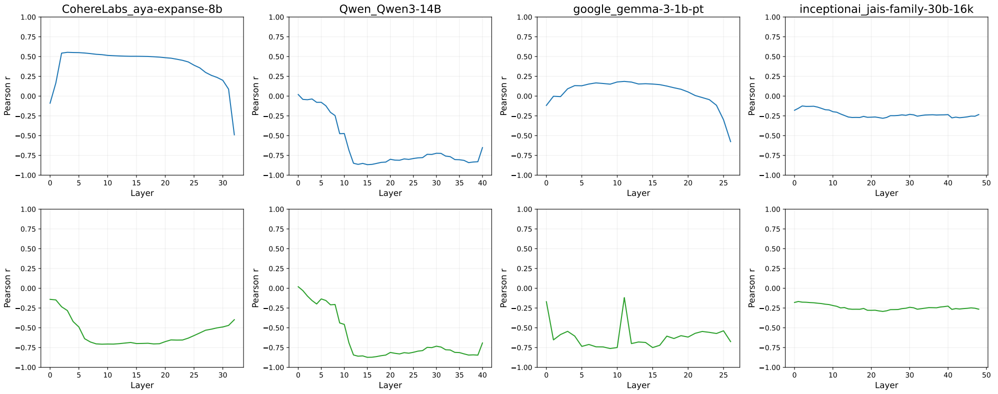

Figure 4: Layer-wise Pearson correlation between representational distance from MSA (L2-Top, Cosine-Bottom) and downstream generation performance. The consistent negative correlation with cosine distance suggests that subspace directional alignment is beneficial.

vides tentative support for the hypothesis that effective transfer requires subspaces to be sufficiently aligned for knowledge sharing (lower cosine) yet distinct enough to preserve variety-specific features (higher L2). Nonetheless, these correlations are observational and should be interpreted cautiously, as they do not generalize uniformly across all models examined.

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 5, standard fine-tuning causes the code length for all dialects to initially increase slightly, as the model specializes for generation rather than classification. However, the increase

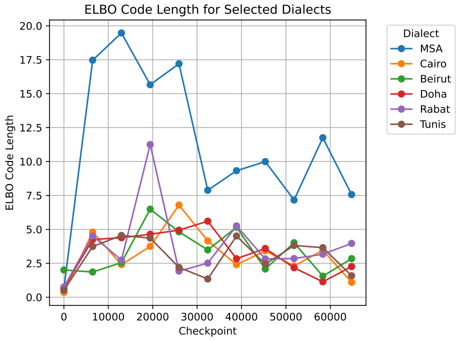

Figure 5: Code Length evolution over baseline training.

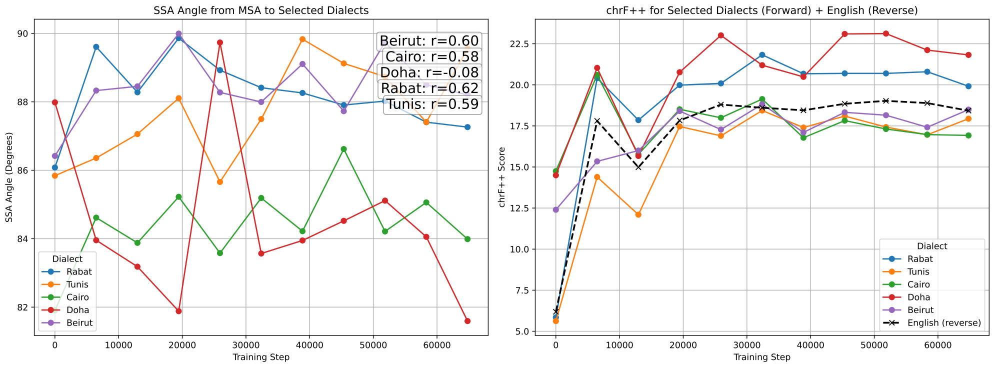

Figure 6: **(Left)** During baseline SFT, the subspace angle (SSA) between MSA and dialects generally shows an increasing trend across all the dialects represented (Except for Doha), indicating growing representational separation. **(Right)** This increase in separation seems to correlate with improved chrF++ scores (pearson r coefficients are shown on the left plot). This provides evidence that disentangling from MSA could be a key mechanism for improving dialectal generation.

is **disproportionately large for MSA**. This indicates that the model is actively making MSA-specific information less linearly accessible, suggesting its initial 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 shows an increasing trend for all dialects shown except for Doha (Figure 6, left). That is, the dialectal subspaces systematically drift away from the MSA subspace. Crucially, this growing separation trend correlates with improvements in generation performance with a pearson correlation coefficient of approximately +0.6 for dialects excluding Doha (Figure 6, 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.

6 Discussion & Future Work

This work identifies *representational entanglement with high-resource languages* as a key barrier to generative modeling in related low-resource varieties. Using *online subspace decoupling*, we dynamically limit high-resource dominance during fine-tuning, showing across six language groups that controlling subspace overlap yields substantial gains (up to **+13.0** ChrF++ / **+10.1** COMET). Geometric and information-theoretic analyses of Arabic dialects further reveal that Modern Standard Arabic (MSA) dominance hinders dialectal generation, underscoring the importance of balanced 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, Sebastian 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`.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. Large language models share representations of latent grammatical concepts across typologically diverse languages. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6131–6150, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.312. URL `https://aclanthology.org/2025.naacl-long.312/`.

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.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773. PMLR, 2020. URL http://proceedings.mlr.press/v108/farajtabar20a.html.

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.

Liang Lin, Zhihao Xu, Junhao Dong, Jian Zhao, Yuchen Yuan, Guibin Zhang, Miao Yu, Yiming Zhang, Zhengtao Yao, Huahui Yi, Dongrui Liu, Xinfeng Li, and Kun Wang. Orthalign: Orthogonal subspace decomposition for non-interfering multi-objective alignment, 2025. URL https://arxiv.org/abs/2509.24610.

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.

nostalgebraist. interpreting gpt: the logit lens, August 31 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. LessWrong / AI Alignment Forum blog post.

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.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52/.

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.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=3AOj0RCNC2.

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.

Anton Schäfer, Shauli Ravfogel, Thomas Hofmann, Tiago Pimentel, and Imanol Schlag. The role of language imbalance in cross-lingual generalisation: Insights from cloned language experiments, 2024. URL `https://arxiv.org/abs/2404.07982`.

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`.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.715. URL `https://aclanthology.org/2023.findings-emnlp.715/`.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL `https://aclanthology.org/2024.acl-long.820/`.

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`.

Chengzhi Zhong, Qianying Liu, Fei Cheng, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. What language do non-English-centric large language models think in? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26333–26346, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1350. URL `https://aclanthology.org/2025.findings-acl.1350/`.

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.

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

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 $\{\mathbf{h}^{(0)}, \dots, \mathbf{h}^{(\ell)}\} \in \mathbb{R}^d$ 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.

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.

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 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.

E.1 Design Choices

Projection Direction Estimation. We estimate the higher-resource variety direction using a *variational linear probe* trained to distinguish the higher-resource variety (positive class) from all other related varieties (negative class). The probe's learned weight vector $\boldsymbol{\theta}_{\text{HR}} \in \mathbb{R}^d$ captures the most discriminative direction separating high-resource from non–high-resource representations. We normalize this vector to obtain the unit direction

$$\mathbf{u}_{\text{HR}} = \frac{\boldsymbol{\theta}_{\text{HR}}}{\|\boldsymbol{\theta}_{\text{HR}}\|},$$

and construct the associated projection matrix

$$\mathbf{P}_{\text{HR}} = \mathbf{u}_{\text{HR}}\mathbf{u}_{\text{HR}}^{\top}.$$

This matrix projects any hidden representation onto the one-dimensional subspace corresponding to the high-resource variety direction.

Online Updating. Rather than estimating this high-resource direction once before training, we periodically retrain the probe on the current model checkpoint during fine-tuning. This **online updating** keeps the projection matrix $\mathbf{P}_{\text{HR}}$ aligned with the model's evolving hidden representation geometry. The projection matrix is refreshed every N_{update} gradient steps; we analyze the effect of this update frequency 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 high-resource direction. This encourages all representations to become more orthogonal to the dominant high-resource feature axis:

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

where $\mathbf{H}$ are the contextual hidden states and $\mathbf{P}_{\text{HR}} = \mathbf{u}_{\text{HR}}\mathbf{u}_{\text{HR}}^{\top}$ is the projection matrix formed from the normalized high-resource direction vector $\mathbf{u}_{\text{HR}}$. This loss minimizes the magnitude of the component of $\mathbf{H}$ aligned with the high-resource direction, thereby discouraging representational overlap with that dominant axis.

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 → lower-resource variety and dialect → 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.

E.3 N_{update} Abalation

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}.

E.4 Hyperparameters

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.

F Detailed Decoupling Results for All Varieties

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 in ChrF++.

src	target	online	baseline	random	delta_online_baseline	delta_online_random
European Portuguese	Brazilian Portuguese	72.600	60.800	_	11.800	_
Brazilian Portuguese	European Portuguese	71.700	61.600	_	10.100	_
English	Indonesian	75.700	71.400	_	4.300	_
Finnish	Kven Finnish	74.800	70.600	_	4.200	_
Standard German	Low German	63.800	61.700	_	2.100	_
Aswan	Modern Standard Arabic	60.000	58.200	_	1.800	_
Low German	Standard German	65.900	64.300	_	1.600	_
Mosul	Modern Standard Arabic	59.300	57.700	_	1.600	_
English	Malay	71.900	70.300	_	1.600	_
Tripoli	Modern Standard Arabic	59.000	57.500	_	1.500	_
Modern Standard Arabic	Sanaa	56.890	55.500	48.900	1.390	7.990
French	Modern Standard Arabic	58.400	57.300	_	1.100	_
Kven Finnish	Finnish	74.800	73.700	_	1.100	_
Baghdad	Modern Standard Arabic	59.400	58.400	_	1.000	_
Modern Standard Arabic	Benghazi	56.505	55.600	47.700	0.905	8.805
Fes	Modern Standard Arabic	58.500	58.000	_	0.500	_
Amman	Modern Standard Arabic	59.500	59.100	_	0.400	_
Benghazi	Modern Standard Arabic	59.400	59.000	_	0.400	_
Muscat	Modern Standard Arabic	60.300	59.900	_	0.400	_
Aleppo	Modern Standard Arabic	58.700	58.400	_	0.300	_
Jerusalem	Modern Standard Arabic	58.600	58.300	_	0.300	_
English	Modern Standard Arabic	59.400	59.100	_	0.300	_
Modern Standard Arabic	Doha	57.679	57.400	49.400	0.279	8.279
Modern Standard Arabic	Salt	58.105	57.900	48.200	0.205	9.905
Basra	Modern Standard Arabic	59.300	59.100	_	0.200	_
Rabat	Modern Standard Arabic	58.700	58.600	_	0.100	_
Modern Standard Arabic	Sfax	50.991	50.900	47.600	0.091	3.391
Algiers	Modern Standard Arabic	58.900	58.900	_	0.000	_
Alexandria	Modern Standard Arabic	59.200	59.200	_	0.000	_
Cairo	Modern Standard Arabic	59.700	59.700	_	0.000	_
Modern Standard Arabic	Mosul	54.978	55.000	47.200	-0.022	7.778
Modern Standard Arabic	Damascus	55.538	55.600	48.000	-0.062	7.538
Jeddah	Modern Standard Arabic	58.600	58.700	_	-0.100	_
Riyadh	Modern Standard Arabic	59.300	59.600	_	-0.300	_
Modern Standard Arabic	Basra	55.817	56.200	47.600	-0.383	8.217
Beirut	Modern Standard Arabic	57.600	58.000	_	-0.400	_
Damascus	Modern Standard Arabic	58.400	58.900	_	-0.500	_
Sanaa	Modern Standard Arabic	58.400	58.900	_	-0.500	_
Doha	Modern Standard Arabic	59.200	59.800	_	-0.600	_
Khartoum	Modern Standard Arabic	59.100	59.700	_	-0.600	_
Modern Standard Arabic	Rabat	53.827	54.700	48.900	-0.873	4.927
Modern Standard Arabic	Aswan	57.264	58.200	53.400	-0.936	3.864
Sfax	Modern Standard Arabic	58.600	59.600	_	-1.000	_
Modern Standard Arabic	Amman	56.608	57.800	47.300	-1.192	9.308
Modern Standard Arabic	Cairo	55.493	56.700	50.100	-1.207	5.393
Modern Standard Arabic	Aleppo	55.014	56.400	49.100	-1.386	5.914
Modern Standard Arabic	Beirut	54.006	55.400	48.400	-1.394	5.606
English	Slovak	46.500	47.900	_	-1.400	_
Salt	Modern Standard Arabic	58.200	59.600	_	-1.400	_
Modern Standard Arabic	Tunis	51.188	52.700	46.500	-1.512	4.688
Modern Standard Arabic	Tripoli	55.589	57.200	49.100	-1.611	6.489
Tunis	Modern Standard Arabic	57.700	59.400	_	-1.700	_
Modern Standard Arabic	Muscat	54.555	56.300	47.600	-1.745	6.955
Modern Standard Arabic	Baghdad	54.334	56.600	45.500	-2.266	8.834
Modern Standard Arabic	Jeddah	56.623	59.200	49.400	-2.577	7.223
Modern Standard Arabic	Fes	53.297	56.200	50.300	-2.903	2.997
Modern Standard Arabic	Khartoum	55.881	58.800	47.700	-2.919	8.181
Modern Standard Arabic	Jerusalem	57.659	60.600	48.500	-2.941	9.159
Modern Standard Arabic	Riyadh	59.325	62.300	49.100	-2.975	10.225
English	Czech	47.800	51.400	_	-3.600	_
Modern Standard Arabic	English	55.203	60.900	46.600	-5.697	8.603
Modern Standard Arabic	Algiers	51.452	57.700	49.700	-6.248	1.752
Modern Standard Arabic	Alexandria	56.422	63.100	52.900	-6.678	3.522
Modern Standard Arabic	French	38.685	51.200	37.600	-12.515	1.085

Table 11: All the results across all of our experimental settings in COMET.

G Statistical Significance Testing per Variety/Setup in ChrF++ Scores

H L2 and Cosine Distances After Decoupling Training

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

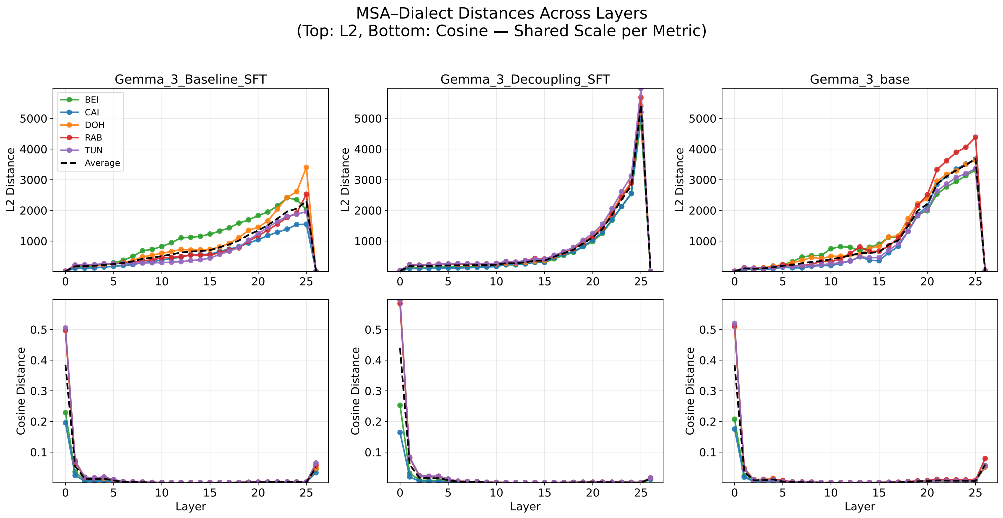

Figure 7: Layer-wise distances between MSA and Arabic dialect representations across Gemma 3 1B before finetuning (**right**) after decoupling (**center**) and baseline (**left**) SFT. Each column corresponds to a model, with the **top** row showing the average L2 distance and the **bottom** row showing the cosine distance between Modern Standard Arabic (MSA) and each dialect's sentence representation at different layers. Colored lines denote individual dialects, and the dashed black line shows the mean distance across dialects.

I 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.